UNIVERSITY OF CALIFORNIA,
IRVINE

Functional Evolution of a Newly Evolved Tandem Multigene Family

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Bryan David Clifton

Dissertation Committee:
Associate Professor José M. Ranz, Chair
Associate Professor J.J. Emerson
Professor Brandon S. Gaut

2022

DEDICATION

To

my parents, Kandra and David Clifton,
and my grandparents Dale and Kathy Rush, Marlene and Paul Krienke, and Bruce and Grace Clifton

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Completing this dissertation would not be possible without the support of many colleagues, friends, and family. Foremost thanks go to my advisor and mentor Prof. José Ranz. Thank you for pushing me farther than I thought I could go while never letting me feel like you would let me fail. Thank you for always being so available and generous with your time when I needed help understanding things. You definitely did not have to put as much effort into me as you did, and I will always be grateful for that.

I would also like to thank my dissertation committee members Prof. J.J. Emerson and Prof. Brandon Gaut. Thank you both for all the stimulating conversations, always having unusually positive attitudes for academics, opening your labs to me for my experiments, and encouraging my scientific career. I also want to thank my past committee members Prof. Adriana Briscoe, Prof. Zeba Wunderlich, and Prof. John Avise for their advice and taking the time to be a part of my committee. I would like to thank Prof. Steve Frank for inspiring my interest in questions in evolutionary biology outside my thesis and improving my ability to review large amounts of literature. I would also like to thank Prof. Nancy Aguilar-Roca for teaching me many pedological methods and being a model educator. You are all very inspiring people.

I want to thank Dr. Shu-Dan Yeh who mentored me when I was an undergraduate. Thank you for teaching me most of my molecular biology and lab organizational skills as an undergraduate and boosting my self-esteem as a scientist, without which I would not have survived graduate school. I want to thank Dr. Mahul Chakraborty for teaching me how to extract high molecular weight DNA and assemble genomes, and always being super friendly and helpful when I can’t figure out how computers work. I also want to thank Dr. Andrei Tatarenkov for always generously taking the time to help with anything he can.

I would like to thank many other supportive members of the UCI community. Thank you Prof. Mike Mulligan and Prof. Jen Martiny for encouraging me to apply for fellowships and awards and for helping me get some slack at a time when my life was overwhelming and I wanted to quit. I want to thank members of the UCI Genomics High Throughput Facility, Dr. Melanie Oakes, Seung-Ah Chung, and Valentina Ciobanu for guidance and sequencing so many of my samples. I would also like to thank EEB department administrators Rodrigo Aguayo, Meranda Aguilar, Carmen Bastos, Kyuri Byun, Warda Bzeih, Thelma Castro, Kate Fuentes, Melanie Nakanishi, Anu Nanjappa, Marie Navas, Marissa Reyes, Irma Rodriguez, and Marcus Umali for countless purchase orders, reservations, TAing-related issues, and key rentals. I also thank Prof. Ali Mortazavi for granting me access to his qPCR machine for my second and third chapters.

I would like to thank the many co-authors of my chapters who have greatly improved the quality of my research. I am grateful to have had many wonderful lab mates and the privilege to mentor many undergraduate researchers while in the Ranz lab. Thank you Michael Caldwell, Kevin Cabrera, Carolus Chan, Stephensen Chea, Kania Gandasetiawan, Imtiyaz Hariyani, Suvini Jayasekera, Karen Ma, Moises Paramo, Daphne Real, Rene Rivera, and Edwin Solares for a fun lab environment and helpful discussions of projects and methods. Thank you Arina Adourian, Sarah Bedford, Liam Bui, Diego Canchola, Zeinab Chahine, Dandan Chen, Kristopher Hernandez, Amir Jaberizadeh, Jamie Jimenez, Eunice Kim, Ashlyn Kimura, Sophia Luo, Brian Luu, Robert Magie, Alvin Nguyen, Kent Nguyen, Edward Oropeza-Rodriguez, Ronni Park, Matthew Raft, Carolina Rojas, Tina Soroudi, Ryan Su, Hayden Tran, Vincent Trieu, Christine Vu, and Amanda Woodbury for your hard work and dedication. Each of you have made my job more enjoyable and rewarding, and it makes me very proud to see all you go on to achieve.

I am very grateful for the friends I have made in graduate school. Your friendship and support have made have made life enjoyable despite countless research setbacks and stresses. Thank you especially to Lucas and Hunna Ustick, Edwin Solares, Alyse Larkin, David Swartout, Newton Hood, Skylar Wyant, Heidi Waite, Andie Nugent, Joseph Suratt, Amanda Barth, Alex Duman, Po, Michelle Herrera, Karina Brocco French, Sam Mahanes, Adrien Arias, Amy Henry, and Martin Minns. I will definitely remember all the nights of barbecues, dinners, trivia, and videogames way more fondly than any of the lab work. Thank you, Edwin, for helping me learn bioinformatics. Thank you, Sam, for always being so welcoming to new people in the department and introducing me to my partner. I also want to thank all the members of the UCI SACNAS chapter. It has been a privilege to get to do outreach with so many welcoming and passionate people.

I am especially grateful to have met my partner Marlo McCarter in graduate school. Thank you, Marlo, for showing me that insects are special, for telling me it’s going to be okay when I’m anxious about work, and for taking me to the hospital when I cut my hand open. I would not have been able to get through preparing for my defense without you. Special thanks to Stan the cat for helping me stay as sane as possible while being home and writing my dissertation.

I want to thank my family for making me part who I am today and helping me get here. I especially want to thank my parents Kandra and David. Thank you both for always encouraging my education and not making me feel bad for still being in school and poor in my 30s. I also thank my grandmother Kathy for allowing me to live close to UCI for practically nothing. If it wasn’t for my cousin Josh watching things like *Cosmos* with me and sharing an interest in learning about science, I am not sure if I would have ever gone to school for biology. I am very grateful for that.

I want to thank my friends Eric, Louis, Ein, Justin, Frank, Anthony, Sean, Robert, Jon, Matt, and AJ for all the fun times during and before graduate school. Thank you for contributing to the life part of my work-life balance, taking part in my hobbies, and texting me jokes.

I would like to acknowledge the funding sources that allowed me to conduct this research. The work in this dissertation was supported by the National Science Foundation (MCB-1157876 and MCB-2129845), the National Institutes of Health (R25-GM056647), the Ministerio de Economía y Competitividad of Spain (CGL2013-45211), and University of California, Irvine intramural funds. I am grateful to have had my stipend momentarily funded by the U.S. Department of Education GAANN fellowship and the National Science Foundation (MCB-2129845). I would like to thank UCI AGS and the EEB department for funding some of my conference fees. Thank you to the University of California, Irvine HPC for facilitating my analyses. Thank you, *Molecular Biology and Evolution*, for permission to include my first two chapters in this dissertation.

Last, but not least, I would like to thank the many, many *Drosophila melanogaster* that were sacrificed for this dissertation.

VITA

Bryan David Clifton

EDUCATION

2014	B.S. in Genetics, University of California, Irvine
2014-2016	Junior Specialist, Department of Ecology and Evolutionary Biology, University of California, Irvine
2016-2021	Teaching Assistant, Department of Ecology and Evolutionary Biology, University of California, Irvine
2019	M.S. in Biological Sciences, University of California, Irvine
2022	Ph.D. in Biological Sciences, University of California, Irvine

PUBLICATIONS

José Ranz, Alwyn Go, Pablo Gonzalez, **Bryan Clifton**, Suzzane Gomes, Amirali Jaberyzadeh, Amanda Woodbury, Carolus Chan, Kania Gandasetiawan, Suvini Jayasekera, Chelsea Gaudreau, Hsiu-Ching Ma, Victor Salceda, Cei Abreu-Goodger, & Alberto Civetta. Gene expression differentiation in the reproductive tissues of *Drosophila willistoni* subspecies and their hybrids. In review - *Molecular Ecology*.

Edward Oropeza-Rodriguez, **Bryan D. Clifton**, and José M. Ranz. 2022. On the genetic basis of the effect of *Spiroplasma* on the male reproductive fitness of *Glossina fuscipes fuscipes*. *PLoS Pathogens*. 18(4):e1010442

Vivek Jayaswal, Cyrille Ndo, Hsiu-Ching Ma, **Bryan Clifton**, Marco Pombi, Kevin Cabrera, Anna Couhet, Karine Mouline, Abdoulaye Diabaté, Roch Dabiré, Diego Ayala, & José M. Ranz. 2021. Intraspecific transcriptome variation and sex-biased expression in *Anopheles arabiensis*. *Genome Biology and Evolution.* 14(2)

José M. Ranz, Pablo M. González, **Bryan D. Clifton**, Nestor O. Nazario, Pablo L. Hernández-Cervantes, Ryan N. Su, Sarah J. Bedford, María J. Palma-Martínez, Dulce I. Valdivia, Andrés Jiménez-Kaufman, Megan M. Lu, Therese Markow, & Cei Abreu Goodger. 2021. A de novo transcriptional atlas in *Danaus plexippus* reveals variability in dosage compensation across tissues. *Communications Biology*. 4, Article number: 791

Bryan D. Clifton, Jamie Jimenez, Ashlyn Kimura, Zeinab Chahine, Pablo Librado, Alejandro Sanchez-Gracia, Mashya Abbassi, Francisco Carranza, Carolus Chan, Marcella Marchetti, Wanting Zhang, Mijuan Shi, Christine Vu, Shu-Dan Yeh, Laura Fanti, Xiao-Qin Xia, Julio Rozas, & José M. Ranz. 2020. Understanding the early evolutionary stages of a tandem *D. melanogaster*-specific gene family: a structural and functional population study. *Molecular Biology and Evolution*. 37(9):2584–2600

José Ranz & **Bryan Clifton**. 2019. Characterization and evolutionary dynamics of complex regions in eukaryotic genomes. *Science China Life Sciences*. 62(4):467-488

Vivek Jayaswal*, Jamie Jimenez*, Robert Magie*, Kien Nguyen, **Bryan Clifton**, Shu-Dan Yeh, & José M. Ranz. 2018. A species-specific multigene family mediates differential sperm displacement in *Drosophila melanogaster*. *Evolution*. 72(2):399-403.

Bryan D. Clifton, Pablo Librado, Shu-Dan Yeh, Edwin A. Solares, Daphne A. Real, Suvini U. Jayasekera, Wanting Zhang, Mijuan Shi, Ronni V. Park, Robert D. Magie, Hsiu-Ching Ma, Xiao-Qin Xia, Antonio Marco, Julio Rozas, & José M. Ranz. 2017. Rapid functional and sequence differentiation of a tandemly repeated species-specific multigene family in *Drosophila*. *Molecular Biology and Evolution*. 34(1):51–65.

Xian B. Mardiros, Ronni Park, **Bryan Clifton**, Gurman Grewal, Amina K. Khizar, Therese A. Markow, José M. Ranz, & Alberto Civetta. 2016. Postmating reproductive isolation between strains of *Drosophila willistoni*. *Fly*. 10(4):162-171.

TEACHING ASSISTANT APPOINTMENTS

Summer 2021	E109: Human Physiology (remote)
Spring 2021	E168: Evolution (remote)
Spring 2021	E112L: Physiology Laboratory (administrative TA; remote)
Winter 2021	E106: Ecology and Evolutionary Biology (remote)
Fall 2020	E153: Genome Evolution (remote)
Fall 2020	E112L: Physiology Laboratory (remote)
Summer 2020	E106: Ecology and Evolution (remote)
Summer 2020	E109: Human Physiology (remote)
Spring 2020	E168: Evolution (remote)
Spring 2020	E112L: Physiology Laboratory (administrative TA; remote)
Winter 2019	E106: Ecology and Evolutionary Biology
Fall 2019	E153: Genome Evolution
Fall 2019	E112L: Physiology Laboratory
Summer 2019	E109: Human Physiology
Winter 2019	E106: Ecology and Evolutionary Biology
Summer 2018	E109: Human Physiology
Spring 2018	E106: Ecology and Evolutionary Biology
Winter 2018	E106: Ecology and Evolutionary Biology

Fall 2017	E112L: Physiology Laboratory
Spring 2017	E112L: Physiology Laboratory
Winter 2017	Bio Sci 94: Organisms to Ecosystems
Fall 2016	E112L: Physiology Laboratory
Fall 2016	E124: Infectious Disease Dynamics

SELECTED AWARDS

2022	Brian G. Atwood '74 and Lynne H. Edminster Graduate Studies Endowment Award
2020	Graduate Fellowship Award
2018	Graduate Assistance in Areas of National Need (GAANN) Fellowship
2017	NSF GRFP Honorable Mention

SELECTED SERVICE

2017	Co-founder, SACNAS at UC Irvine chapter
2017 – 2021	Outreach Chair, SACNAS at UC Irvine chapter

ABSTRACT OF THE DISSERATION

Functional Evolution of a Newly Evolved Tandem Multigene Family

By

Bryan David Clifton

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2022

Associate Professor José M. Ranz, Chair

Species-specific expansions of gene duplicates foster adaptation, genetic innovation, and phenotypic diversification. While it is recognized that events during their early evolutionary history are important for determining if a gene duplicate will be retained, lost, or become nonfunctional, models of gene family evolution have mostly been built from studies of relatively ancient gene families. Therefore, how genes overcome the immediate consequences of duplication, *i.e.,* dosage increase, and accumulate the molecular diversity required for novel functions, while also being impacted by molecular mechanisms such as gene conversion, and evolutionary forces such as genetic drift and selection along the path to fixation, remains largely uncharacterized.

My goal was to characterize the functional evolution of a young tandem gene expansion found only in *Drosophila melanogaster*: *Sperm-specific dynein intermediate chain* (*Sdic*). I aimed to accurately reconstruct the *Sdic* region at the structural and sequence level while obtaining accurate information about sequence diversity among the *Sdic* paralogs in different strains from different geographical origins (Chapters 1 & 2); investigate the extent of *Sdic* copy number variation (CNV) (Chapter 2) while examining the relationship between *Sdic* copy number and total *Sdic* expression (Chapters 2 & 3); and probe the divergence of different expression attributes

among *Sdic* paralogs within and between strains while gauging the impact of *cis* and *trans* regulatory variation (Chapters 1 & 3).

Through my research, I established the correct structure of the *Sdic* region in the *D. melanogaster* reference genome using raw long read sequences and showed the *Sdic* paralogs exhibit variable expression in both abundance and breadth using qRT-PCR and RNA-seq. I generated a precise portrait of *Sdic* copy number variation using reference-quality genome annotations, qPCR, and read-depth methods. Only one *Sdic* paralog is fixed across populations and there is no evidence of pseudogenization among paralogs. While artificially doubling copy number within the same genomic background increased male expression over two-fold, I observed no correlation between copy number and total *Sdic* expression across natural populations, suggesting differential regulatory modifiers likely play key roles in shaping *Sdic* expression. Further, I used RNA-seq to quantify *Sdic* expression in testes from populations with *Sdic* CNV, as well as testis, heads, and accessory glands from males with identical genomes except for different *Y* chromosomes. In testis, I found clear evidence of variable expression among *Sdic* paralogs and a positive correlation between *Sdic* CNV and expression. The *Y* chromosome seems to impact total expression of *Sdic* in accessory glands but not testes or heads.

My dissertation represents a rare interpopulation characterization of a species-specific multigene family at the sequence, structural, and functional levels. *Sdic* epitomizes how quickly a tandem multigene family can functionally diversify at both the coding and regulatory levels, even in the face of gene conversion. Beyond maintaining a minimally optimal expression level, the presence of *Sdic* duplicates appears to act as a catalyst for generating protein and regulatory diversity, showcasing a possible evolutionary path that novel gene functions can follow toward long-term consolidation within eukaryotic genomes.

INTRODUCTION

The evolution of novel gene functions underlies phenotypic innovation and adaptation, contributing ultimately to the diversification of life on Earth. New gene functions most commonly originate through the duplication of existing genes followed by functional divergence among the retained duplicates at the protein coding and/or expression levels. While most duplicates quickly decay into pseudogenes (Force et al 1999), various scenarios or models have been proposed to explain the retention of gene duplicates (see Innan & Kondrashov 2010 for a comprehensive list of evolutionary scenarios; Kuzmin et al 2022). Due to the difficulty of studying young nearly identical duplicates (Ranz & Clifton 2019), the models posed are largely based on studies of evolutionarily ancient duplicates that have acquired multiple secondary mutations, limiting our understanding of the early consequences of gene duplication events on the phenotype and fitness. While these models can propose mechanisms by which new functions arise, comprehensive functional characterizations of young gene duplicates that can test specific hypotheses that evaluate the applicability of these models to the early evolutionary stages of functional diversification among paralogs comprising particular gene families are lacking. Therefore, how gene duplicates overcome the immediate consequences of gene duplication, *i.e.*, dosage increase, and accumulate the molecular diversity required for novel functions, while also being impacted by gene conversion, genetic drift, and natural selection along the path to fixation, remains largely uncharacterized.

The youngest gene duplicates are found in only a single species. Species-specific genes have been shown to impact organismal fitness and contribute to phenotypic change (Chen et al. 2010; Yeh et al 2012; Jugulam et al. 2014; Mayer et al. 2015; Florio et al 2015; Fiddes et al 2018; Chakraborty et al. 2019), with recent expansions of tandemly duplicated genes thought to play key roles in adaptation, phenotypic diversification, and genetic innovation (Brown et al. 1998; Newcomb et al. 2005; Perry et al. 2007; Jugulam et al. 2014). These tandem gene families are thought to primarily originate through DNA-based duplication events mediated by non-allelic homologous recombination (NAHR) events, *i.e.*, unequal crossing over, which occur during meiosis (Hastings et al 2009). Uncovering the mechanisms that shape the functional attributes of individual paralogs within tandem gene families during their early evolutionary stages has been precluded by three major difficulties. First, repetitive regions composed of multiple highly similar tandem repeats, *i.e.*, *structurally complex genomic regions*, remain refractory to accurate sequence reconstruction in even 'reference quality' genome assemblies (Clifton et al 2017, 2020). Second, these genomic regions often display copy number variation (CNV), involving duplicates with high sequence identity that result from NAHR and gene conversion events (Clifton et al 2017, 2020; Loehlin et al 2021). Third, the rules that govern the expression of gene duplicates as they age are still not well understood (Kondrashov 2010; Rody et al 2017; Teufel et al 2018; Loehlin et al 2021; Kuzmin et al 2022). Overall, these difficulties have resulted in a scarcity of intraspecific studies that can properly evaluate paralog diversity and the functional dynamics of tandem multigene families at the early stages of their formation and consolidation in eukaryotic genomes. The goal of this dissertation is to contribute to filling this gap in the literature by characterizing the *Drosophila melanogaster*-specific structurally complex genomic region, *Sperm-specific dynein intermediate chain* (*Sdic*).

Sdic is a tandem multigene family present only in *D. melanogaster*, so therefore originated after the split of the *melanogaster* lineage from the *simulans* clade ~1.4 Ma (Nurminsky et al. 1998; Obbard et al. 2012). *Sdic* is one of the few genetic factors known to influence sperm competition (Civetta & Ranz 2019), *i.e.*, a form of sexual selection that biases fertilization at the postcopulatory level, through an impact on sperm competitive ability (Yeh et al 2012; Jayaswal et al 2018). Due to its young age, tandemly repeated structure, and role in adaptive evolution, the *Sdic* multigene family provides the opportunity to investigate different levels of change and their consequences during the early stages of tandem multigene family evolution, which has been typically neglected despite its importance for understanding the fate of gene duplicates and the origin of new gene functions (Kondrashov 2010; Katju & Bergthorsson 2013; Long et al. 2013; Cardoso-Moreira et al. 2016; Naseeb et al. 2017; Rogers et al. 2017). The original *Sdic* gene originated from a segmental duplication on the *X* chromosome involving two adjacent genes, *short wing* (*sw*) and *Annexin B10* (*AnxB10*), in which the central genes fused into a chimeric entity that essentially encodes a defective form of the sw protein, a cytoplasmic dynein intermediate chain, *i.e.*, a regulatory subunit of the cytoplasmic dynein motor protein complex (Nurminsky et al. 1998; Kardon & Vale 2009). Subsequently, *Sdic* became repeatedly tandemly duplicated, representing one of the most noticeable gene family expansions in *D. melanogaster* (Nurminsky et al. 1998; Hahn et al. 2007; Clifton et al 2017).

The repetitive nature and high sequence similarity among *Sdic* paralogs and the flanking parental genes likely facilitated recurrent NAHR events, which is expected to have caused repeated contractions and expansions of the tandem array, contributing to CNV (Clifton et al 2020; Hastings et al. 2009), as well as rampant gene conversion events, which contributes to high sequence identity levels among the repeats (Clifton et al 2017). Nevertheless, simple questions such as the magnitude of CNV, the most common copy number in populations, whether gene conversion impacts the whole length of the *Sdic* repeat or only particular intervals, or whether different paralogs are expressed at different levels or have entered into the path of nonfunctionalization have not been addressed. A key aspect that explains this lack of knowledge for *Sdic* and other species-specific tandem expansions is that their sequence properties and repetitive nature make them particularly challenging to reconstruct in genome assemblies and characterize molecularly. My dissertation is comprised of three chapters contributing to the characterization of the *Sdic* region of the *D. melanogaster* genome.

The aims of this dissertation are to: i) accurately reconstruct the Sdic *region at the structural and sequence level while obtaining accurate information about sequence diversity among the* Sdic *paralogs in different strains from different geographical origins (Chapters 1 and 2); ii) reveal the extent of* Sdic *CNV present in* D. melanogaster *(Chapter 2) while examining the relationship between* Sdic *copy number and total* Sdic *expression (Chapters 2 and 3); and iii) probe the divergence of different expression attributes (level and tissue presence) among* Sdic *paralogs within and between strains while gauging the impact of* cis *and* trans *regulatory variation (Chapters 1 and 3).*

Specifically, Chapter 1 focuses on properly characterizing the structure of *Sdic* region and expression of the individual paralogs in the *D. melanogaster* reference genome ISO-1 and another common laboratory strain, Oregon-R (Clifton et al 2017). I analyzed individual raw PacBio long sequencing reads to demonstrate that the *Sdic* region of the reference genome assembly (Release 6; dos Santos et al 2015) was incorrectly assembled in both copy number and order and provided the correct position of the copies. I also demonstrated that *Sdic* expression is not limited to sperm as previously believed, *e.g.*, in ovaries (female germline tissue) and heads (unisex somatic tissue), and that individual *Sdic* paralogs show diverged expression in both abundance and breadth in ISO-1. Further, I demonstrated that female expression of sperm enhancing *Sdic* does not induce a sexually antagonistic effect on female fecundity. This chapter highlights the difficulty of obtaining accurate reconstructions of structurally complex regions like *Sdic* and demonstrates how quickly a tandemly arranged multigene family can functionally diversify at both the coding and regulatory levels, even in the face of gene conversion.

Chapter 2 focuses on characterizing *Sdic* structural and functional variation across a range of geographically diverse strains (Clifton et al 2020). I annotated the *Sdic* region in reference-quality genome assemblies from 15 populations of *D. melanogaster,* comparing their copy number with estimates from both qPCR (a quantitative molecular technique) and CNVnator (a computational sequencing read depth-based technique) –which allowed further analysis of 83 individuals– as a metric of proper region assembly. While qPCR and CNVnator showed complete agreement, *Sdic* was only reliably assembled in ~50% of the assemblies. I confirmed the existence of *Sdic* CNV, with ~97% of individuals harboring four to eight copies. Across the eight reliably assembled genomes, I found no evidence of pseudogenization, with only one isoform being fixed across all the strains and the remaining isoforms existing as a floating pool of arguably nonessential duplicates. While synthetic strains carrying a duplication of the *Sdic* region exhibit increased male expression >2-fold, I detected no correlation between copy number and expression variation across different genomic backgrounds from natural populations, suggesting that differential regulatory genome modifiers likely play a central role in shaping *Sdic* expression levels. Further, one *Sdic* duplicate strain showed ~3-fold increased expression and had similar sperm competitive ability to its derived wildtype strain, however another *Sdic* duplicate strain with ~4-fold expression had decreased sperm competitive ability, suggesting *Sdic* dosage might be constrained to an optimal level. Beyond maintaining a minimally optimal expression level, duplication of the *Sdic* copies appears to act as a catalyst of protein and regulatory diversity, detailing a possible evolutionary path that novel gene families can follow toward long-term consolidation within eukaryotic genomes. Importantly, this chapter showcases how refractory structurally complex genomic regions can be to proper assembly, even in supposed reference-quality genomes assemblies.

Chapter 3 focuses on how *cis* variation among *Sdic* paralogs and *trans* variation across the genome, specifically the *Y* chromosome, impact the expression levels of these paralogs within and across populations. I performed RNA-sequencing on testes from four of the populations with reliable genome assemblies used in Chapter 2. I quantified expression of individual paralogs by tracking paralog-specific sequences. I found a positive correlation between *Sdic* copy number and total *Sdic* expression in testis contrary to the pattern seen in whole bodies (Chapter 2), which is compatible with *Sdic* dosage being under positive selection in testis. I also detected a negative correlation between *Sdic*'s parental gene *sw* and total *Sdic* expression in testis, suggesting the possible existence of a regulatory mechanism that maintains total dynein intermediate chain dosage within a limit that does not significantly impede cytoplasmic dynein protein complex assembly and functionality. Within all four strains, I detected significantly different expression among the *Sdic* paralogs, with promoter type, copy position within the tandem array, and coding sequence all having no consistent impact on expression of the *Sdic* paralogs. To study the impact of the *Y* chromosome on expression, I created a set of seven strains differing only by their *Y* chromosome and measured total *Sdic* expression with qRT-PCR, finding that the *Y* chromosome impacts expression of *Sdic* in some strains but not others. From four of these strains, the total RNA from testes, accessory glands, and heads was sequenced. I found that the *Y* chromosome impacts total expression of *Sdic* in accessory glands, a somatic tissue with a role in reproduction, but not in testes or heads. This chapter highlights the importance of integrating precise paralog-specific sequence information with tissue-level expression data to obtain accurate and nuanced portraits of how the functional attributes of recently originated multigene families evolve along their path to fixation and consolidation in the genome.

Overall, this dissertation represents a sophisticated paralog and tissue level characterization of a species-specific multigene family, adding to a few others such as those reported in *Homo sapiens* (Dougherty et al 2018; Fiddes et al 2018) and *Cannabis sativa* (Vergara et al 2019). This work pioneers the proper reconstruction of structurally complex genomic regions while characterizing them at the functional level. As most prior studies of gene duplicates have been performed in the context of evolutionarily older gene families, this dissertation provides a rare portrait of the early evolution of multigene families along the path to consolidation in eukaryotic genomes.

REFERENCES

Brown CJ, Todd KM, Rosenzweig RF. 1998. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Mol Biol Evol*. 15(8):931–942.

Cardoso-Moreira M, Arguello JR, Gottipati S, Harshman LG, Grenier JK, Clark AG. 2016. Evidence for the fixation of gene duplications by positive selection in *Drosophila*. *Genome Res*. 26(6):787–798.

Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. 2019. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat Commun*. 10(1):4872.

Chen S, Zhang YE, Long M. 2010. New genes in *Drosophila* quickly become essential. Science 330:1682–1685.

Civetta A, Ranz JM. 2019. Genetic Factors Influencing Sperm Competition. *Front Genet*. 13;10:820.

Clifton BD, Librado P, Yeh SD, Solares ES, Real DA, Jayasekera SU, Zhang W, Shi M, Park RV, Magie RD, Ma H, Xia X, Marco A, Rozas J, Ranz JM. 2017. Rapid functional and sequence differentiation of a tandemly repeated species-specific multigene family in *Drosophila*. *Mol Biol Evol*. 34(1):51–65

Clifton BD, Jimenez J, Kimura A, Chahine Z, Librado P, Sanchez-Gracia A, Abbassi M, Carranza F, Chan C, Marchetti M, Zhang W, Shi M, Vu C, Yeh S, Fanti L, Xia X, Rozas J, Ranz JM. 2020. Understanding the early evolutionary stages of a tandem *D. melanogaster*-specific gene family: a structural and functional population study. *Mol Biol Evol*. 37(9):2584–2600

dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, Emmert DB, Gelbart WM, FlyBase C, the FlyBase Consortium. 2015. FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res*. 43(D1):D690–D697.

Dougherty ML, Underwood JG, Nelson BJ, Tseng E, Munson KM, Penn O, Nowakowski TJ, Pollen AA, Eichler EE. 2018. Transcriptional fates of human-specific segmental duplications in brain. *Genome Res*. 28(10):1566-1576.

Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, van den Bout A, Bishara A, Rosenkrantz JI, Lorig-Roach R, Field AR, Maeussler M, Russo L, Bhaduri A, Nowakowski TJ, Pollen AA, Dougherty ML, Nuttle X, Addor MC, Zwolinski S, Katzman S, Kriegstein A, Eichler EE, Salama SR, Jacobs FMJ, Haussler D. 2018. Human-specific *NOTCH2NL* genes affect Notch signaling and cortical neurogenesis. *Cell*. 173(6): 1356–1369.e22.

Florio M, Albert M, Taverna E, Namba T, Brandl H, Lewitus E, Haffner C, Sykes A, Wong FK, Peters J, Guhr E, Klemroth S, Prüfer K, Kelso J, Naumaan R, Nüsslein I, Dahl A, Lachmann R, Pääbo S, Huttner WB. 2015. Human-specific gene *ARHGAP11B* promotes basal progenitor amplification and neocortex expansion. *Science*. 347(6229):1465-70

Force A, Lynch M, Pickett FB, Amores A, Yan Y, Postlehwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*. 151:1531–154

Hahn MW, Han MV, Han SG. 2007. Gene family evolution across 12 *Drosophila* genomes. PLoS Genet. 3:e197.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet*. 10(8):551–564.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*. 11:97-108

Jayaswal V, Jimenez J, Magie R, Nguyen K, Clifton B, Yeh S, Ranz JM. 2018. A species-specific multigene family mediates differential sperm displacement in *Drosophila melanogaster*. *Evolution*. 72(2):399–403.

Jugulam M, Niehues K, Godar AS, Koo DH, Danilova T, Friebe B, Sehgal S, Varanasi VK, Wiersma A, Westra P, et al. 2014. Tandem amplification of a chromosomal segment harboring 5-enolpyruvylshikimate-3-phosphate synthase locus confers glyphosate resistance in *Kochia scoparia*. *Plant Physiol*. 166(3):1200–1207.

Kardon JR, Vale RD. 2009. Regulators of the cytoplasmic dynein intermediate chain. *Nat Rev Mol Cell Biol*. 10:854–865

Katju V, Bergthorsson U. 2013. Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Front Genet*. 4:273.

Kondrashov FA. 2010. Gene Dosage and Duplication. Evolution After Gene Duplication. *Wiley-Blackwell*

Kuzmin E, Taylor JS, Boone C. 2022. Retention of duplication genes in evolution. *Trends in Genetics*. 38(1):59-72

Loehlin DW, Kim JY, Paster CO. 2021. A tandem duplication in *Drosophila melanogaster* shows enhanced expression beyond the gene copy number. *Genetics*.

Long M, VanKuren NW, Chen S, Vibranovski MD. 2013. New gene evolution: little did we know. *Annu Rev Genet*. 47:307–333.

Mayer MG, Rodelsperger C, Witte H, Riebesell M, Sommer RJ. 2015. The orphan gene *dauerless* regulates Dauer development and intraspecific competition in nematodes by copy number variation. *PLoS Genet*. 11:e1005146.

Naseeb S, Ames RM, Delneri D, Lovell SC. 2017. Rapid functional and evolutionary changes follow gene duplication in yeast. *Proc R Soc B*. 284(1861):20171393.

Newcomb RD, Gleeson DM, Yong CG, Russell RJ, Oakeshott JG. 2005. Multiple mutations and gene duplications conferring organophosphorus insecticide resistance have been selected at the Rop-1 locus of the sheep blowfly, *Lucilia cuprina*. *J Mol Evol*. 60:207–220.

Obbard DJ, Maclennan J, Kim KW, Rambaut A, O'Grady PM, Jiggins FM. 2012. Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. *Mol Biol Evol*. 29(11):3459–3473.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet*. 39:1256–1260.

Ranz J, Clifton B. 2019. Characterization and evolutionary dynamics of complex regions in eukaryotic genomes. *Sci China Life Sci*. 62(4):467-488.

Rogers RL, Shao L, Thornton KR. 2017. Tandem duplications lead to novel expression patterns through exon shuffling in *Drosophila yakuba*. *PLoS Genet*. 13(5):e1006795.

Rody HVS, Baute GJ, Rieseberg LH, Oliveira LO. 2017. Both mechanism and age of duplications contribute to biased gene retention patterns in plants. *BMC Genomics*. 18(46)

Teufel AI, Johnson MM, Laurent JM, Kachroo AH, Marcottee EM, Claus OW. 2018. The Many Nuanced Evolutionary Consequences of Duplicated Genes. *Mol. Biol. Evol.* 36(3):304-314

Vergara D, Huscher EL, Keepers KG, Givens RM, Cizek CG, Torres A, Gaudino R, Kane NC. 2019. Gene copy number is associated with phytochemistry in *Cannabis sativa*. *AoB Plants*. 11(6)

Yeh SD, Do T, Chan C, Cordova A, Carranza F, Yamamoto EA, Abbassi M, Gandasetiawan KA, Librado P, Damia E, et al. 2012. Functional evidence that a recently evolved *Drosophila* sperm-specific gene boosts sperm competition. *Proc Natl Acad Sci USA*. 109(6):2043–2048.

CHAPTER 1

Rapid functional and sequence differentiation of a tandemly repeated species-specific multigene family in *Drosophila*

ABSTRACT

Gene clusters of recently duplicated genes are hotbeds for evolutionary change. However, our understanding of how mutational mechanisms and evolutionary forces shape the structural and functional evolution of these clusters is hindered by the high sequence identity among the copies, which typically results in their inaccurate representation in genome assemblies. The presumed testis-specific, chimeric gene *Sdic* originated, and tandemly expanded in *Drosophila melanogaster*, contributing to increased male-male competition. Using various types of massively parallel sequencing data, we studied the organization, sequence evolution, and functional attributes of the different *Sdic* copies. By leveraging long-read sequencing data, we uncovered both copy number and order differences from the currently accepted annotation for the *Sdic* region. Despite evidence for pervasive gene conversion affecting the *Sdic* copies, we also detected signatures of two episodes of diversifying selection, which have contributed to the evolution of a variety of C-termini and miRNA binding site compositions. Expression analyses involving RNA-seq datasets from 59 different biological conditions revealed distinctive expression breadths among the copies, with three copies being transcribed in females, opening the possibility to a sexually antagonistic effect. Phenotypic assays using *Sdic* knock-out strains indicated that should this antagonistic effect exist, it does not compromise female fertility. Our results strongly suggest that the genome consolidation of the *Sdic* gene cluster is more the result of a quick exploration of different paths of molecular tinkering by different copies than a mere dosage increase, which could be a recurrent evolutionary outcome in the presence of persistent sexual selection.

INTRODUCTION

Genes restricted to one or a few closely related species are ubiquitous across phyla (Tautz and Domazet-Loso 2011; Long et al. 2013). Despite their young age, these genes can exert noteworthy effects on organismal viability and fertility (Chen et al. 2010; Mayer et al. 2015), therefore their study is instrumental for determining how early mutational mechanisms and evolutionary forces refine the functional attributes of a gene and its organismal impact shortly after its formation (Hahn 2009; Chen et al. 2013). This is especially important in the case of recent expansions of tandemly duplicated genes, which are thought to play a primary role during species adaptation and differentiation (Brown et al. 1998; Newcomb et al. 2005; Perry et al. 2007; Jugulam et al. 2014).

Genome consolidation of recent duplicates can be achieved throughout different evolutionary paths in which natural selection and genetic drift contribute with different intensities (Innan and Kondrashov 2010; Katju and Bergthorsson 2013). In particular, the expansion dynamics of gene clusters is commonly thought to be associated with a beneficial effect via increased gene dosage (Ohno 1970; Kondrashov 2012). However, this process can be subsequently accompanied by some degree of functional diversification among the duplicates through a secondary functional attribute of the gene product (Bergthorsson et al. 2007). A relevant constraint on functional paralog divergence to consider is the homogenizing effect exerted by interlocus gene conversion, i.e., the non-reciprocal recombination process that results in the transfer of DNA stretches between similar non-allelic sequences, which is particularly relevant in the case of young tandemly arranged duplicates (Chen et al. 2007; Osada and Innan 2008; Casola et al. 2010). Importantly, this homogenizing effect also impacts the retention probability of the duplicates and therefore their ability to contribute to species adaptation (Walsh 1987; Innan 2003; Katju 2012). Critically, the analysis of the functional and evolutionary dynamics of recent tandem expansions of species-specific genes is hindered precisely by the repetitive nature and high sequence identity of the constituent copies. These features limit the resolution of microarray and quantitative PCR technologies as well as the information derived from short-read based sequencing technologies, which typically results in an inaccurate representation of these gene clusters in current genome assemblies in the form of sequence errors or copies being collapsed (Hemingway et al. 2004; Bariami et al. 2012; Krsticevic et al. 2015).

The *Sperm-specific dynein intermediate chain* (*Sdic*) multigene family originated in the *D. melanogaster* lineage less than 4.9 mya (Obbard et al. 2012). The *Sdic* ancestral copy started its formation with a local segmental duplication of two adjacent genes on the *X* chromosome, *AnxB10* and *sw*. This was followed by point mutations and indels of varying size that obliterated sections along the parental genes, resulting in a fusion event between their inner copies, with *AnxB10* not contributing to the transcribed region of *Sdic*, and a de novo exon acquisition from a previously non-coding sequence of *sw* (fig. 1.1A) (Nurminsky et al. 1998b). Subsequently, *Sdic* became repeatedly tandemly duplicated, representing one of the most noticeable gene family expansions in *D. melanogaster* (Hahn et al. 2007). One *Sdic* copy has been shown to be expressed only in males, with its encoded product present in the tail of mature spermatocytes, collectively pointing toward a role in male fertility. Based on functional features and comparative sequence analysis, the Sdic protein was classified as an axonemal, rather than cytoplasmic, dynein intermediate chain (Nurminsky et al. 1998b). Genome engineering experiments coupled with phenotypic tests ultimately uncovered that the *Sdic* region boosts sperm competitive ability (Yeh et al. 2012), in line with its presumed adaptive nature (Kulathinal et al. 2004), making *Sdic* one of the few examples of a recently formed gene cluster that is unambiguously linked to sexual selection.

Due to its short age, highly tandemly repeated nature, and role in adaptive evolution, the *Sdic* multigene family has the potential to reveal key insights about the mode and tempo of the functional evolution that accompanies the formation and consolidation of similar gene clusters in the genome. However, the most recent release of the *D. melanogaster* genome sequence (Release 6) includes the presence of additional copies compared with the previous release (Release 5) (dos Santos et al. 2015), whereas functionally validated information only exists for one of the *Sdic* copies (Nurminsky et al. 1998b). Therefore, the actual structure of the *Sdic* cluster, and the extent to which the different copies exhibit identical functional attributes at the protein and expression levels, remain uncertain. Thus, resolving these questions is essential to evaluating whether the gene cluster is evolving in a concerted manner or has started a diversification process in which some of the copies have entered into a pseudogenization process. Additionally, a genome-wide analysis of the architecture of sexual antagonism in *D. melanogaster* indicated that the variable expression of one of the *Sdic* copies was associated with opposed effects on male and female fitness (Innocenti and Morrow 2010). In summary, the key structural and functional aspects of the *Sdic* gene cluster continue to remain elusive, impeding a correct analysis of the region's patterns of change and a precise view of its contribution to fitness.

Here we have investigated the evolutionary history of the constituent members of the *Sdic* gene cluster. This study first seeks to precisely reconstruct and annotate one the most challenging regions of the euchromatic fraction of the *D. melanogaster* genome by leveraging the increased resolution associated with long-read sequencing technologies, which have been shown to be instrumental in comprehensive studies of complex genomic regions including tandemly arranged duplicates (Huddleston et al. 2014; Krsticevic et al. 2015); second, to evaluate how different molecular mechanisms and evolutionary forces have shaped the current levels and patterns of DNA variability among the copies, ultimately recreating the most plausible scenario underlying the expansion of the cluster; and third, to determine the degree of functional diversification among different *Sdic* copies by performing a copy-specific monitoring of their expression, paying special attention to sex differences and a potential impact on female fitness.

We present a much more complex organizational and functional portrait of the evolution of the *Sdic* multigene family than previously thought (Nurminsky et al. 1998b; Ponce and Hartl 2006). For this, we devised analytical approaches tailored to accommodate the sequence similarity among the copies in order to leverage multiple available assemblies and preassemblies generated by long-read sequencing technologies (Kim et al. 2014; McCoy et al. 2014; Berlin et al. 2015) and RNA-seq datasets from different developmental stages and body parts (Graveley et al. 2011; Brown et al. 2014). We uncover differences with the current annotation of the *Sdic* region, both in number of copies and internal positioning (dos Santos et al. 2015). Our proposed evolutionary scenario for the formation of the *Sdic* multigene family involves a minimum of four unequal-crossing over events, pervasive gene conversion, and two episodes of positive selection. Despite the young age of this multigene family, we find clear signs of expression diversification across biological conditions with a varying expression breadth among its members, including expression in females although without resulting in decreased fertility according to phenotypic tests. Additionally, our results suggest that the *Sdic* protein may not function only as a sperm-specific axonemal dynein intermediate chain. Collectively, the *Sdic* multigene family epitomizes how quickly a tandemly arranged multigene family can functionally diversify at both the coding and regulatory levels, even in the face of gene conversion, through the acquisition of uneven sexually dimorphic expression.

RESULTS

Assessing the Assembly of the *Sdic* Region

The *Sdic* region is located at 19C1 on the *X* chromosome and is composed of tandem repeats absent in other *Drosophila* species (supplementary fig. S1.1, Supplementary Material online). Each repeat consists of three parts of which the transcriptional unit that encodes the *Sdic* protein is the most relevant (fig. 1.1B). Releases 5 and 6 of the genome assembly of the ISO1 strain differ considerably at the *Sdic* region (Hoskins et al. 2007; dos Santos et al. 2015). Release 5 included four copies of the *Sdic* repeat whereas Release 6 added three new copies (CG46275, CG46276, and CG46277; hereafter *SdicA*, *SdicB*, and *SdicC*, respectively), in addition to substantial sequence changes for copies *Sdic3* and *Sdic4* (fig. 1.1B; supplementary table S1.1, Supplementary Material online). This copy number increase is in good agreement with previous estimates at the molecular and computational levels (Benevolenskaya et al. 1995; Yeh et al. 2012). The fewer number of repeats in Release 5 could be the result of collapsed Sanger sequencing reads of high sequence identity.

To verify the organization of the *Sdic* region in Release 6, we examined other assemblies for the strain ISO1 based on long sequencing reads (table 1.1 and supplementary text, Supplementary Material online). Long reads are more likely to harbor sequence stretches distinctive of particular individual or adjacent repeats, informing about their internal positioning. We examined four assemblies: three assembled from the same set of single-molecule real-time (SMRT) sequencing reads, differing only in their assembly methods (Kim et al. 2014; Berlin et al. 2015; S. Koren and C.S. Chin, unpublished data; see Material and Methods), and one obtained with Illumina TruSeq Synthetic Long-Reads (SLRs) (McCoy et al. 2014). Two of the SMRT-based assemblies, Berlin and PBcR hereafter (table 1.1), produced an unfragmented *Sdic* region (Kim et al. 2014; Berlin et al. 2015). Using a set of diagnostic sequence motifs for each *Sdic* copy (supplementary table S1.2, Supplementary Material online), we located all *Sdic* repeats in the assemblies and proceeded with their precise annotation. For the two unfragmented reconstructions, we found the same number of copies, arranged in the same fashion, although displaying some sequence differences. Critically, both reconstructions differ from Release 6 in having one less copy of the two that are identical in sequence (*Sdic3* and *SdicA*), as well as in the relative order of the copies, with *Sdic2* and *Sdic4* switching places (fig. 1.1B). Collectively, these results strongly support that the Berlin and PBcR assemblies should be considered as an alternative to Release 6 for the *Sdic* region, especially the former given the improvements associated with locality-sensitive hashing-based assemblies (Berlin et al. 2015).

Despite providing a fragmented assembly, the extremely low error rate associated with Illumina TruSeq sequencing (McCoy et al. 2014) makes SLRs especially appropriate to validate the reconstruction of the *Sdic* region in the Release 6 and Berlin assemblies (Berlin et al. 2015). The rationale is that the absence of differences between a particular SLR and one of the assemblies likely reflects the actual sequence in the ISO1 strain. Using BLASTn, we retrieved 319 SLRs encompassing exonic sequences from the *Sdic* copies. Next, we filtered out reads that were so long that they contained the same region from two copies as assessed by Blast2seq (Johnson et al. 2008), which could lead to misassembly (Krsticevic et al. 2015), or so short that they did not retain motifs distinctive of individual copies. The combination of these criteria led us to consider 122 4–7.6 kb long SLRs, which were mapped against the two assemblies using BLASR (Chaisson and Tesler 2012) (supplementary fig. S1.2, Supplementary Material online). Most SLRs showed higher sequence identity in their alignment with one of the two assemblies, with 43 SLRs differing in which *Sdic* copy they were mapping against, which followed different patterns (supplementary table S1.3 and fig. S1.3, Supplementary Material online). Importantly, thorough scrutiny of the alignments revealed that the selected SLRs aligned more optimally with the Berlin assembly than with the Release 6 (supplementary fig. S1.4 and text, Supplementary Material online).

To determine the support level for each *Sdic* copy in the two assemblies, we focused on 107 SLRs showing high quality alignments and found a more even coverage across *Sdic* copies in the Berlin assembly (supplementary fig. S1.5 and text, Supplementary Material online). We also screened some diagnostic sequence stretches indicative of a more accurate reconstruction of the region. Specifically, we determined whether any SLR supported distinctive junctions (*Sdic1-Sdic2*, *Sdic2-Sdic3*, and *SdicC-Sdic4* in Release 6; *Sdic1-Sdic4, Sdic4-Sdic3*, and *SdicC-Sdic2* in the Berlin assembly) and same copy differences in the two assemblies (supplementary table S1.4, Supplementary Material online). For both features, we found SLRs solely supporting the Berlin assembly. On balance, our results indicate that the Berlin assembly most accurately recapitulates the *Sdic* region in the ISO1 strain.

Sequence Diversity

The six annotated copies of *Sdic* in the Berlin assembly (Berlin et al. 2015) range in nucleotide sequence identity percentage from 93.9% to 99.1%, with a median value of 97.6% from the start to stop codons (supplementary table S1.5, Supplementary Material online). This identity level decreases only moderately when the whole gene fraction is considered (93.4–98.9%, median = 97.45%). From the transcriptional start to stop site, most nucleotide differences and indels accumulate in exons 4 and 5, the intron residing between them, and the 3’UTR. Only considering differences that result in amino acid replacements, excluding those due to frameshift mutations and deletions (see below), all nine non-synonymous changes found reside in exons 4 and 5, none of them being present across all *Sdic* copies. For the same alignable regions, only two synonymous changes are detected.

At the amino acid level, the sequence identity among the different Sdic protein variants ranges from 86.1% to 100%, with Sdic3 and SdicB being identical (supplementary fig. S1.6, Supplementary Material online). In terms of domain composition, the Sdic protein variants harbor either six or four WD40 motifs as confirmed by protein domain search in INTERPRO (supplementary fig. S1.6, Supplementary Material online); all sw proteins possess six WD40 motifs (supplementary fig. S1.6, Supplementary Material online). Based on the number of carboxyl end WD40 motifs, we grouped the putative Sdic proteins in two sets. The four WD40 motif-containing set includes Sdic1-PC and Sdic4-PE and is characterized by the shortest protein variants as a result of shifts in splice sites. Sdic1-PA also belongs within this first set of variants, although it exhibits a conspicuous structure as a result of three deletions in exon 5 (supplementary fig. S1.7, Supplementary Material online). Further, the six WD40 motif-containing set is characterized by a carboxyl end either identical to that of sw (all Sdic2 isoforms) or affected by several amino acid deletions and replacements (SdicB-PA, SdicC-PA, and Sdic3-PE, Sdic3-PF, Sdic3-PG). Importantly, the nucleotide differences that alter the donor splice site at the 3' end of exon 4 in Sdic4 and SdicC also mediate the automatic conversion of ancestrally intronic sequence from sw into the Sdic coding sequence. In fact, for SdicC, the whole intronic sequence is read through such that it connects exons 4 and 5 (supplementary fig. S1.7, Supplementary Material online).

In addition to the WD40 motifs, all the Sdic and sw protein variants harbor a cytoplasmic dynein 1 intermediate chain 1/2 domain (supplementary fig. S1.6, Supplementary Material online). Further, sequence comparison of the newly evolved N-terminus of the Sdic protein variants against other known axonemal dynein intermediate chain proteins revealed a negligible level of sequence similarity, which was in good agreement with the lack of significant matches in sequence similarity searches with BLASTp (Altschul et al. 1997). Collectively, these results are suggestive of a cytoplasmic role for the Sdic protein variants, without ruling out their function in the axoneme, which would take place through a non-canonical axonemal domain.

Molecular Evolution of the *Sdic* Multigene Family

The evolution of tandemly arranged gene duplicates often involves an initial phase driven by gene conversion, followed by a second phase where genetic drift and/or selection limit further sequence homogenization, enabling functional divergence (Fawcett and Innan 2011). Taking advantage of the validated Berlin assembly, we evaluated the relative contributions of gene conversion and adaptive diversification to the evolution of the six *Sdic* copies.

The analysis of the 5'–3' distribution of the between-copy variation supported the distinction of two broad sections within *Sdic*. The 5' section begins at the transcription start site and ends at the 12 nt long gap present in the stretch that codes for the fourth WD40 domain. The 3' section proceeds from this gap to the transcription stop site (supplementary fig. S1.8 and S1.13, Supplementary Material online). GeneConv (Sawyer 1989) revealed 23 statistically significant gene conversion tracts $P_{\mathrm{adj}}<0.05$), suggesting a scenario where the inner copies (*Sdic2*, *Sdic3*, *Sdic4*, *SdicC*, and *SdicB*) exchange DNA segments with each other, as well as the 50 regions with *Sdic1*, and the 3' regions with *sw* (supplementary table S1.6, Supplementary Material online). This is in line with the physical positions of *Sdic1* and *sw* as the most outermost genes in the region that are involved in these putative gene conversion events. Five out of the 23 gene conversion tracts show lengths larger than the maximum documented genome-wide in *D. melanogaster* (Casola et al. 2010). This unusual length may be due to the high *Sdic* sequence identity, which precludes the accurate delineation of converted tracts, resulting in the artifactual joining of adjacent stretches of exchanged DNA. Further, the boundaries of these converted tracts show a clear co-localization with the five likely recombination breakpoints inferred by ACG (O'Fallon 2013), which split *Sdic* into six partitions with independent evolutionary histories (P1-P6; fig. 1.2A). P1-P4 would correspond to the 5' section of the *Sdic* sequence whereas the 30 section would span P5–P6.

Overall, our results suggest that gene conversion is a major contributor to the shaping of the *Sdic* multigene family's pattern of variability. Nevertheless, the inspection of the local gene genealogies (fig. 1.2A) revealed that the statistical significance supporting the putatively converted DNA segments is partly driven by the accumulation of singletons (i.e., mutations in a single *Sdic* copy; long branches in the local genealogies of P1, P3, P5, and P6; fig. 1.2A). Given that all mutations are confined to one copy, GeneConv systematically infers that the remaining copies must be homogenizing their DNA sequences by exchanging DNA, a pattern also compatible with other evolutionary scenarios, including a relaxation of purifying selection and the action of positive selection. Using models especially devoted to quantifying the impact of natural selection on coding and non-coding regions (see "Material and Methods"), we found that all *Sdic* copies are evolving under purifying selection, with ~90–95% of their nucleotide positions being invariable or having substitutions rates lower than the synonymous substitution rate. However, the intensity of purifying selection does vary across copies and particularly across partitions. For example, the exonization of the intronic region of *sw* in *Sdic* likely resulted in a stochastic accumulation of mutations in the *sw* intron but not the homologous *Sdic* exon, from which they were purged. This is reflected as a long branch in the local genealogy of partition P1, a pattern that could mimic the signal of positive selection (*sw-AnxB10* branch in the P1 genealogy, fig. 1.2A).

The test conducted is also especially robust at detecting positive selection in the face of potentially confounding factors, such as relaxed purifying selection or GC-biased gene conversion (see "Materials and Methods"). We identified two lineages showing statistical evidence for positive selection (supplementary table S1.7, Supplementary Material online). The first corresponds to the basal lineage leading to the ancestor of all *Sdic* copies in P1 and P3, and the second to the external lineage leading to *Sdic1* in P5. The first episode of positive selection occurred after the formation of the ancestral *Sdic* gene, probably driving mutations responsible for its expression to fixation, such as the acquisition of a translation start site. The second subsequent episode exclusively affected *Sdic1* in partition 5, which has accommodated multiple indels and other nucleotide differences that have led to multiple amino acid replacements (supplementary fig. S1.8,

Supplementary Material online). Interestingly, partition P5 encompasses the constitutive fraction of the 3'UTR, which has undergone a profound remodeling of its miRNA binding site composition across copies, especially in the case of *Sdic1* (see below).

We tentatively reconstructed a scenario of duplications that leads to the contemporary organization of the *Sdic* region in the reference strain ISO1 (fig. 1.3). For that, we took into consideration the phylogenetic relationship among the *Sdic* copies inferred from the gene tree topology exhibited by partition P4, as well as key shared diagnostic changes (e.g., in the promoter region –see below–). Unlike a gene topology based on the whole *Sdic* sequence, P4's topology has experienced limited gene conversion and does not exhibit singleton enrichment, and hence more faithfully recapitulates the evolutionary history of the duplication events and the correct gene tree topology of the family (Slightom et al. 1985; McGrath et al. 2009) (fig. 1.2B–C). The proposed scenario puts forward that upon formation of the ancestral *Sdic*, a duplication event took place giving rise to two copies. One of the two copies, the one adjacent to sw, would have evolved to what is known as *Sdic2*. In parallel, the other copy would have become duplicated again giving rise to two copies, the most downstream from *sw* being the ancestor of *Sdic1*, *Sdic3*, and *SdicB*

(*Sdic1/3/B*), and the middle copy being the ancestor to *SdicC* and *Sdic4* (*SdicC/4*). Protocopies *Sdic1/3/B* and *SdicC/4* would have then duplicated jointly, increasing the number of copies from three to five, originating the precursors of *Sdic1* and *Sdic4* on the downstream side, and the ancestors of both *SdicC* and *Sdic3* and *SdicB* (*Sdic3/B*) near the middle of the cluster. An additional duplication of the protocopy *Sdic3/B* would have then occurred, giving rise to the precursors of *Sdic3* and *SdicB*. Only the temporal sequence of origination of *Sdic1*, *Sdic3*, and *SdicB* conflicts with their phylogenetic relationship, which suggests a different sequence of events: *Sdic1/3/B* → *Sdic3* and *Sdic1/B*, then *Sdic1/B* → *Sdic1* and *SdicB*. Nevertheless, the ancestral node joining *Sdic1*, *Sdic3*, and *SdicB* exhibits a low bootstrap value being this parsimonious scenario also supported by the occurrence of 0 amino acid replacements and 13 silent changes between *Sdic3* and *SdicB*. In the proposed scenario, the tandem duplication of the *Sdic* region would have come about via four unequal crossing-over events.

Expression Diversification among *Sdic* Copies

Previous characterization of *Sdic* expression was limited to *Sdic1* (Nurminsky et al. 1998b; Mikhaylova and Nurminsky 2011). To evaluate potential expression differences among *Sdic* copies, we focused on two amplicons for which the design of specific primers was more feasible. One amplicon is associated exclusively with *Sdic1* whereas the other is shared between *Sdic4* and *SdicC* (hereafter *Sdic**). RT-PCR experiments with the OR-R strain uncovered that both *Sdic1* and *Sdic** are expressed in not just testes, but also ovaries, demonstrating that expression of these copies is not male specific (supplementary fig. S1.9, Supplementary Material online). *Sdic* female expression was also reproduced in the African strain ZW-109 (supplementary fig. S1.10, Supplementary Material online). Furthermore, we detected expression of both amplicons in both male and female heads (supplementary fig. S1.9, Supplementary Material online). In order to better quantify expression differences across tissues, sexes, and strains, we performed qRT-PCR experiments. The results confirmed high expression levels of *Sdic1* and *Sdic4* in testes from the two strains, as well as lower expression levels in ovaries and heads from both sexes (supplementary table S1.8 and fig. S1.11, Supplementary Material online). Interestingly, in ZW-109, *Sdic4*, but not *Sdic1*, was overexpressed in male relative to female heads, a pattern not observed for OR-R. These results support a much more complex spatial expression profile for *Sdic* than previously reported (Nurminsky et al. 1998b).

Even if no disruptive amino acid replacement or premature stop codon has altered the functionality of the different *Sdic* protein variants, the pseudogenization of some of the copies can arise from mutations within the promoter region. We observe two nucleotide differences in the promoter region of *Sdic3* and *SdicB* in relation to the remaining *Sdic* copies (supplementary fig. S1.12, Supplementary Material online). These two nucleotide differences were confirmed in *Sdic3* and *SdicB* by 3 and 4 SLRs, respectively. Importantly, one of these differences falls within a sequence stretch that is similar to a motif in the bTub85D gene promoter responsible for testis-expression specificity (Michiels et al. 1989). In order to both determine the potential impact of the nucleotide differences within the promoter region and generate a more comprehensive expression profile of the *Sdic* copies, we searched for copy-specific motifs and scrutinized their presence—no mismatch allowed—across ~3.15 billion RNA-seq reads representing 59 biological samples from different anatomical parts and developmental timepoints (Graveley et al. 2011; Brown et al. 2014). This measure was necessary as many reads have the potential to map against several *Sdic* copies or *sw*. After corroborating their absence in sw, five motifs were delineated within the most 3’ third of *Sdic*1, *Sdic2*, *Sdic3*, *Sdic4*, and *SdicC* (supplementary table S1.10 and fig. S1.13, Supplementary Material online); no informative motif was found for *SdicB*.

Given the conservative nature of our approach, we pooled all reads from the libraries associated with the same biological condition. In this way, we maximized our capability to detect reads containing the diagnostic motifs, which was used as evidence of expression. The number of reads for which we detected perfect alignments, corrected by the sequencing depth of the biological condition in question, was adopted as proxy for expression level (supplementary table S1.9, Supplementary Material online). In spite of limitations derived from, for example, the fact that some motifs have the potential to survey more than one transcript for a particular copy whereas others are specific to a single mRNA transcript variant, it was possible to uncover distinctive characteristics for the expression profile of the different *Sdic* copies (fig. 1.4A–B, supplementary fig. S1.14, Supplementary Material online).

We found evidence of expression for all five copies surveyed, which, combined with the absence of premature stop codons and evidence of purifying selection, reinforces the notion that none of the *Sdic* copies has entered into a pseudogenization process in the ISO1 strain. From the developmental perspective, all copies showed sustained expression from third instar larvae throughout adulthood, although episodic expression of *Sdic3* was detected in earlier developmental stages. The expression level of the *Sdic* copies increases during the pupal stage, reaching maximum values in 5-day-old males, which correlates well with the testes expression evidence obtained via RT- and qRT-PCR experiments for particular *Sdic* copies. In fact, it is in samples unambiguously linked to males only (eight out of 59) that all *Sdic* copies show their highest expression levels. Considering the six samples (three developmental and three anatomical, roughly 10% of the total) in which each copy shows the highest expression levels, we find *Sdic1* and *Sdic4* displaying the most marked trend, with five out of the six samples being linked to males. Among the anatomical samples linked to males, *Sdic1* stands out by showing its highest expression levels in testes and accessory glands of 4-day-old males, whereas *Sdic3* showed its highest expression levels in head samples from males of different ages. Further, although the developmental samples do not show evidence of systematic expression of the *Sdic* copies in females, the anatomical samples clearly show evidence for the expression of *Sdic3* in eight out of 11 samples unambiguously linked to females. Interestingly, we detect profound variation among *Sdic* copies in their contribution to the expression profile of particular biological conditions not previously shown for this multigene family. For example, *Sdic*3 contributes disproportionately more to the global expression of *Sdic* in the central nervous system of third instar larvae and 2-day-old white prepupae than any other copy. Likewise, we find marked differences in expression specificity values (s) among copies (fig. 1.4C). In fact, Monte Carlo simulations showed that *Sdic3* possesses a significantly wider expression breadth (i.e., lower s value) than the rest of the assayed copies ($P < 0.001$).

Variation in expression attributes among the *Sdic* copies can arise through both the pre- and post-transcriptional regulation. The currently annotated promoter sequences are virtually identical barring two nucleotide substitutions. These sequence changes differentiate *Sdic3* and *SdicB* from the rest of the copies, which could result in differential competing ability to recruit transcriptional machinery in the particular biological conditions in which the constituents of this machinery are in limited concentrations. In fact, *Sdic3* exhibits a clearly different expression breadth compared with the rest of the surveyed copies. Alternatively, differences in expression attributes could result from the recruitment of a slightly different set of downstream regulators. This might have happened through the severe 3’UTR remodeling across *Sdic* copies, resulting in differential post-transcriptional regulation via microRNAs. To explore this, we scanned the 3’ UTRs of all *Sdic* and *sw* transcripts for canonical miRNA target sites. We identified target sites for up to 54 distinct mature microRNAs (supplementary table S1.11, Supplementary Material online). By considering the gain/loss profile of orthologous miRNA target sites, we observed that only four target sites were conserved across all *Sdic* and *sw* transcripts. In fact, *sw* and *Sdic2* had a very similar targeting profile (supplementary fig. S1.15A, Supplementary Material online), suggesting a profound remodeling process of the 3'UTRs occurred after the divergence between *Sdic2* and the rest of *Sdic* copies (supplementary fig. S1.15B, Supplementary Material online). *Sdic1*, the copy characterized by the most male-biased profile, also exhibits the most markedly different miRNA binding site profile. *Sdic1* has the largest number of specific, novel target sites (14), harboring sites in exclusive for 10 miRNAs. Overall, we observed regulated *Sdic* expression throughout development and across body parts, the absence of expression silencing, and incipient differences among copies. How the interplay between promoter differences and remodeled 3'UTR miRNA binding site compositions contribute to the observed expression differences is not apparent at this time.

The *Sdic* Region and Female Fertility

All *Sdic* copies are expressed in males whereas 3–4 copies (*Sdic*1, *Sdic3*, and either *Sdic4*, *SdicC*, or both) show expression in females. Further, microarray experiments coupled with hemiclonal analysis pointed to *Sdic3*, now several copies based on our improved annotation, as a locus that displays sexual antagonism with regard to variable gene expression (Innocenti and Morrow 2010); *sw* did not show this pattern. As the *Sdic* region enhances sperm competitive ability (Yeh et al. 2012), this opens the possibility that the *Sdic* region as a whole can have an opposed effect on the fitness of the sexes. We examined the effects of deleting the *Sdic* region in females under the hypothesis that there would be a fitness boost if *Sdic* expression impairs female fertility.

We generated synthetic genotypes for the *Sdic* region using previously engineered deletions of the entire *Sdic* region via non-homologous recombination (Yeh et al. 2012) (supplementary fig. S1.16A, Supplementary Material online). This was done upon reassuring that the changes introduced to the annotation of the *Sdic* region were compatible with no *Sdic* copy remaining in X(19C1), which could compromise the interpretation of any phenotypic test (supplementary fig. S1.17, Supplementary Material online). We assayed three relevant parameters for female fertility: female productivity, i.e., the progeny number; number of eggs laid; and egg hatching rate. Homozygous females for the deletion of the *Sdic* region (A^{-d} and E^{-d}) were compared against wild-type females for the region (B^{+} and I^{+}) by monitoring differences in female productivity over a 33-day-period (Methods and supplementary fig. S1.16B, Supplementary Material online). The knock-out strains did not exhibit increased productivity relative to their wild-type counterparts and w^{1118}, another control strain (supplementary table S1.12, Supplementary Material online). We found statistically significant differences in each timepoint examined, but they mostly resulted from a consistently low productivity of the wild-type control I^{+} (supplementary table S1.12, Supplementary Material online). In relation to the other two wild-type strains B^{+} and w^{1118}, the knock-out strains E^{-d} and A^{-d} did not show any consistent pattern, with at least one of them displaying no significant differences in productivity for most of the timepoints assayed.

No difference in productivity among females with and without the *Sdic* region could result from counteracting factors, e.g., a higher number of eggs laid being offset by a lower hatching rate. We tested for differences in these two parameters over a 6-day period and found no evidence that the absence of the *Sdic* region correlates with a higher number of eggs laid or a higher hatching rate (supplementary table S1.13-S14 and fig. S1.16C, Supplementary Material online). Failure to find statistically significant differences could result from a lack of power due to limited sample size, particularly in the case of hatching rate. However, the global trend seems to be robust, with two of the wild-type strains (B$^+$ and *w*^{1118}) showing very similar values to those of the knockout strains. Overall, these results indicate that *Sdic* expression in females does not impair the fertility of this sex, which does not exclude that it can impact negatively other fitness traits.

DISCUSSION

Our analysis of the *Sdic* region in *D. melanogaster* represents a step forward in the generation of accurate portraits of the organizational, sequence, and functional evolution of recently originated, tandemly arranged multigene families. This is needed as our current knowledge is primarily based on tandemly arrange families of ancient origin such as the globins or rRNA genes (Brown et al. 1972; Zimmer et al. 1980), cases involving young tandem duplicates with a limited number of members (Osada and Innan 2008), or cases in which the functional data is limited or lacking (Moore and Purugganan 2003). Genomic regions harboring recently expanded gene clusters are hotspots for structural and functional change, having the potential to foster adaptive evolution (Brown et al. 1998; Newcomb et al. 2005; Perry et al. 2007; Jugulam et al. 2014). By coupling long-read sequencing technologies (Eid et al. 2009) with RNA-seq data from multiple biological conditions, and tailored analytical approaches that accommodate the particularities of members of these type of multigene families, we can now perform unparalleled multilevel characterizations of these complex genomic regions.

At the organization level, the combined use of different long-sequencing read technologies has prompted us to propose a different organization for the *Sdic* multigene family in the ISO1 strain from the one currently accepted (dos Santos et al. 2015). This alternative organization differs in both number and internal arrangement of the copies. To account for the six copies in this alternative organization, we propose a duplication scenario involving a minimum of four unequal crossing-over events. Further, the inter-copy variability patterns are compatible with a scenario of rampant inter-locus gene conversion, especially involving the outermost members of the cluster. Despite the homogenizing effects of gene conversion, we found a preferential accumulation of mutations towards the 3' end of the *Sdic* copies, affecting both coding and non-coding sequence, which would have been driven partially by positive selection. Examples of positive selection overcoming the effects of gene conversion have also been documented for other recently originated tandem duplicates (Innan 2003; Osada and Innan 2008). Importantly, the role of positive selection in shaping the patterns of nucleotide polymorphism and divergence in the *Sdic* region has been controversial (Brookfield 2001; Kulathinal et al. 2004). We found evidence that copy differentiation at the sequence level is compatible with at least two episodes of positive selection, one shortly after the origin of the ancestral copy, and a more recent episode exclusively affecting the 3' end of one copy (*Sdic1*). These signatures of positive selection and the lack of evidence for pseudogenization of the *Sdic* copies scrutinized provide strong support to the adaptive role of *Sdic*.

The six copies documented encode a variety of Sdic proteins which differ primarily at their C-terminus, where the protein sw presumably interacts with the dynein heavy chain, as inferred from its ortholog in Dictyostelium (dicA; Ma et al. 1999). Importantly, all Sdic and sw variants possess a common cytoplasmic dynein 1 intermediate chain 1/2 domain, suggesting Sdic could function similarly to sw. However, the lack of coiled-coil and serine-rich domains at the N-terminus of Sdic would presumably prevent the Sdic variants from interacting with the dynactin protein complex, which mediates the interaction of the dynein protein complex with a variety of subcellular structures (Nurminsky et al. 1998a; Maet al. 1999). Overall, Sdic and sw might share a limited set of common interactions with other protein complex subunits and subcellular structures. In fact, these structural differences, and the expression profile exhibited by some Sdic copies, are suggestive of a Sdic protein that interacts with non-axonemal dynein complexes present in tissues possessing both ciliated (e.g., sperm) and non-ciliated cells (e.g., salivary glands and imaginal discs). Whether or not Sdic interacts with axonemal dynein complexes cannot be inferred from our results, but the fact that the silencing of the whole multigene family results in a significant reduction in sperm competitive ability does not allow us to discard this possibility (Yeh et al. 2012).

The *Sdic* multigene family shows a pattern of expression consistent with quick regulatory diversification among copies. As is the case for other recently originated genes, *Sdic* was likely expressed in testes at a very early stage (Kaessmann 2010; Zhao et al. 2014). This is the only expression attribute in adults shared across all copies, whereas expression in females was displayed by 3–4 copies, varying across adult samples, including some (*Sdic1* and *Sdic3*) that were inferred to be among the most recently generated in the gene family. *Sdic*'s testis expression could have resulted from a rather simple promoter motif with incipient testis-biased expression (Nurminsky et al. 1998b; FitzGerald et al. 2006), a benign molecular environment (Schmidt and Schibler 1995; Sassone-Corsi 2002), or both. Subsequently, selective pressures such as post-mating male–male competition (Kleene 2005; Singh and Kulathinal 2005) would have mediated the retention and expansion of *Sdic*, as supported by phenotypic assays (Yeh et al. 2012). Exactly when the broadening of expression took place relative to the origination of some the copies is unclear at this time, as is how the differences in promoter sequence and 3'UTR miRNA binding site composition led to the observed expression differences. Nevertheless, these unclarified aspects point to some interesting directions. First, whereas functional broadening over evolutionary time is a hallmark of many old duplicates (Assis and Bachtrog 2013; Kaessmann 2010), including expression in both sexes, *Sdic3* highlights how quickly this broadening trend can occur. Second, functional diversification of tandemly arranged duplicates might proceed through posttranscriptional regulatory changes driven by the evolution of a unique composition of miRNA binding sites (Wang and Adams 2015), as could be the case for *Sdic1*, revealing an important path for the diversification of DNA-mediated duplicates.

The functional complexity of the *Sdic* copies, revealed here through their protein domain compositions and expression profiles, questions whether the phenotypic impact of the *Sdic* region is confined to post-mating male–male competition. It is possible that *Sdic* expression in females can result in a sexually antagonistic effect as circumstantial evidence suggests (Innocenti and Morrow 2010), fitting into the notion that the *X* chromosome, where *Sdic* resides, is a key genomic reservoir of sexually antagonistic genetic variation (Rice 1984; Gibson et al. 2002). Our results for three parameters of female fertility suggest that should this antagonistic effect exist, it impacts either a more subtle fertility component or a completely different type of trait from those tested here.

Regardless of the organismic impact of the *Sdic* region, our results show that the amplification of *Sdic* has not consisted merely in a gene dosage increase. Nevertheless, it remains a challenge to fully understand the evolutionary implications of the *Sdic* amplification. We hypothesize that the Sdic protein could have facilitated the emergence of a secondary, unrefined function of sw (Hughes 1994) or novel interactions between the dynein complex and other protein complexes or cellular components via the novel N-terminus. Additionally, sw has been shown to interact with the p150-Glued subunit of dynactin in a dosage-dependent manner, suggesting that *Sdic*, which is essentially identical to sw but cannot bind the p150-Glued subunit, could act as a competitive inhibitor of the interaction between the dynein and dynactin complexes (Boylan et al. 2000). Whether it is because of an enhanced secondary or an entirely novel function, the benefit of *Sdic* could have become more apparent upon its overexpression via copy number increase

(Bergthorsson et al. 2007), with some of the copies subsequently undertaking different paths of evolutionary tinkering. This pattern is compatible with the variation in domain composition and expression profiles seen for the *Sdic* copies in the ISO1 strain. Equivalent multilevel characterization of the *Sdic* gene cluster in other *D. melanogaster* strains as performed here will help gauge some key aspects. The first is whether *Sdic*'s functional refinement is still ongoing, with some of the copies possibly undergoing pseudogenization, or alternatively whether the existing copies are part of a diversification process associated with balancing selection, both scenarios driven by the permanent action of sexual selection. The second aspect is whether there is an optimal range of copies refractory to the extreme outcomes of unequal crossing-over, i.e., the complete loss of *Sdic* or an unbearably high copy number which would both be detrimental.

MATERIALS AND METHODS

Assembly and Annotation Analysis

All assemblies used are associated with sequencing experiments that made use of the ISO1 isogenic strain *y; cn bw sp* Adams et al. 2000). These include: the complete sequence of BAC10C18 (GenBank accession number AC011705.11); Release 6 plus ISO1MT (GCA_000001215.4; dos Santos et al. 2015); assembly ASM77845v1, which is based on SMRT sequencing reads ASM77845v1 (GCA_000778455.1; Berlin et al. 2015); and an assembly based on Illumina TruSeq SLRs (GCA_000705575.1; McCoy et al. 2014). The assembly ASM77845v1 was generated using the Celera assembler (v8.2) and MHAP as overlapper. Using the same reads as assembly ASM77845v1, two additional preassemblies just differing in computational pipeline aspects, were included. The preassembly reported in Kim et al. (2014) uses the overlapper implemented in the HGAP (hierarchical genome assembly process) pipeline and can be retrieved from http://cbcb.umd.edu/software/pbcr/dmel_cons_asm.tar.gz (last accessed December 1, 2015).

The other SMRT based preassembly was generated using the FALCON v0.1 assembler, which can be retrieved from https://s3.amazonaws.com/datasets.pacb.com/2014/Drosophila/reads/dmel_FALCON_diploid_assembly.tgz (last accessed December 1, 2014). Contigs containing *Sdic* copies that are part of different assemblies were identified using Bowtie2 v2.2.3 (Langmead and Salzberg 2012) under parameter settings –fast-local and –no-unal, whereas using the sequences of the annotated exons of the *Sdic* copies in Release 6 as a query. The annotation of the *Sdic* region in the assembly GCA_000778455.1 was done taking the gene structure of each *Sdic* copy in Release 6 as a reference.

In the case of the scrutiny of SLRs to test the validity of particular assemblies, FASTQ files (Dm4-1 to Dm4-3, and Dm5-1 to Dm5-3) were downloaded from the Illumina BaseSpace site and tested for significant similarity with *Sdic* exonic sequences using BLASTn v2.2.30 (Altschul et al. 1990). The mapping of SLRs against particular assemblies was done using BLASR v1.3.1 (Chaisson and Tesler 2012) under the default minimum percent identity and setting -bestn 1 in order to prevent multiple alignments. Prior to this, the *Sdic* region in each assembly under comparison was indexed using the program sawriter, which is part of the SMRT Analysis toolkit available at the Pacific Biosciences Developer's Community Network Website (DevNet: http://www.smrtcommunity.com/DevNet; last accessed December 1, 2015). TABLET v1.14.10.20 (Milne et al. 2013) was used for alignment visualization and confirmation of key motifs.

Molecular Evolution Mode

A multiple sequence alignment (MSA) composed of the six *Sdic* copies, from the start of the promoter to the end of the 3'UTR, was assembled including as well an artificial composite sequence comprised of the homologous *sw* and *AnxB10* regions (*sw-AnxB10*) as an outgroup. Using MEGA v6.06 (Tamura et al. 2013), sequence alignments were performed with MUSCLE and refined by visual inspection. Levels of divergence along the sequence alignment, plus the number of synonymous and non-synonymous substitutions, were calculated with DnaSP v5 (Librado and Rozas 2009). The maximum likelihood (ML) phylogenetic tree was reconstructed using RAxML v8.12 (Stamatakis 2014) with 1,000 bootstrap replicates.

Gene conversion tracts were inferred using the GeneConv program (Sawyer 1989) under the assumption that no nucleotide mismatches occurred among the tracts, reflecting the negligible probability of these events happening during the very early evolutionary stages of a multigene family like *Sdic*. We applied the Bonferroni correction to obtain the adjusted probability with which a particular tract experienced gene conversion. As GeneConv tracts might modify the local gene genealogy, we further examined whether *Sdic* exhibits incongruent gene genealogies along its sequence by estimating the recombination breakpoints with the ACG program (O'Fallon 2013), which implements explicit models that fully capture the coalescent process with recombination. The ACG Markov chain was run for 20,000,000 iterations, with a burn-in period of 5,000,000.

The HyPhy batch script, written by Oliver Fredigo (https://github.com/ofedrigo/TestForPositiveSelection/blob/master/nonCodingSelection.bf; last accessed January 15, 2016), was used to test for positive selection acting on specific *Sdic* copies (Haygood et al. 2007). This script evaluates whether the substitution rate in a focal class of sites, which can be comprised of any kind of functional category, is higher than in a neutral class of sites (here represented by the synonymous sites). The statistical significance of this test is assessed by comparing two nested models by means of a Likelihood Ratio Test (LRT). The null model assumes three classes of sites, including positions that are (i) selectively neutral, (ii) evolving under purifying selection, or (iii) purged in background lineages, but neutrally evolving in the foreground branch. The alternative model replaces class (iii) with two extra classes that assume a fraction of the sites are evolving under positive selection in the foreground lineages, but under either (iv) neutral or (v) purifying selection in the background lineages. Thus, this test enables distinguish between positive and relaxed purifying selection, as the latter is already accounted for in the null model. To accommodate for the different gene tree topologies found for each partition along the MSA, this test was separately conducted for each of the *Sdic* sequence partitions identified by the ACG recombination breakpoints. Exclusively for this analysis, we included a second artificial composite sequence comprised of the orthologous stretches to *sw* and *AnxB10* in *D. simulans*, which was used as a more external outgroup. This enabled to clearly distinguish, within each partition, whether basal episodes of positive selection occurred in the lineage leading to the ancestor to all *Sdic* copies or in that leading to the *D. melanogaster* composite *sw-AnxB10*.

Strains and Fly Husbandry

D. melanogaster strains used are listed in supplementary table S1.15, Supplementary Material online. Flies were reared on dextrose-cornmeal-yeast medium in a 25C chamber under constant lighting conditions. Adult virgins were collected within 6 h of eclosion, sorted by sex, and then cultured separately in groups of ≤10 individuals. At 4–6 days post-eclosion, entire adult whole bodies and other dissected biological samples (male and female heads, testes, and ovaries) were homogenized and stored in TRIzol (Life Technologies) at –80^{o}C. Dissections were done separately for each type of biological sample in ice-cold 1x PBS solution. All sorting, scoring, collecting, counting, and manipulation of flies was performed under CO2 anesthesia.

Total RNA Extraction and cDNA Synthesis

For the strains Oregon-R and Zimbabwe-109, total RNA was extracted from three biological replicates corresponding to each strain by sex by tissue combination. Following manufacturer's instructions, total RNA was extracted from tissues previously homogenized in

TRIzol. DNA traces were removed by treating 10 mg of each sample with Turbo DNA-free DNase (Ambion). RNA integrity and purity were confirmed using gel electrophoresis and a NanoDrop spectrophotometer respectively. cDNAs for each sample were generated using 1 mg of DNase-treated total RNA, oligo(dT) primers, and SuperScript III reverse transcriptase (Invitrogen) in the presence of RNaseOUT recombinant RNase inhibitor (Invitrogen). All female samples were tested for male contamination by RT-PCR of the Y-linked gene CG41561. cDNA quality was confirmed by RT-PCR of Gapdh2.

PCR-Based Expression Profiling

RT-PCRs were performed using TaKaRa Ex Taq polymerase (Clontech), 2 mL cDNA template, and appropriate primers. The correct identity of each amplicon was confirmed by gel electrophoresis, Sanger sequencing, and subsequent BLASTn analysis. qRT-PCR experiments were performed essentially as described (Yeh et al. 2014). Possible reference genes were selected based on their expression stability as shown by modENCODE RNA-seq data in FlyBase (dos Santos et al. 2015), as well as the expression profile between the sexes as reported in the Sex Bias Gene Expression Database (Gnad and Parsch 2006). Subsequent verification of expression stability, as indicated by the GeNorm program (Statminer, TIBCO Spotfire suite v6.5.3 -Perkin Elmer-), led us to use two reference genes: *clot* and *CG14903*. Estimates for expression differences were obtained using the $-2^{\Delta\Delta Cq}$ method (Livak and Schmittgen 2001). P-values were calculated using the Limma moderate t-test (Smyth 2004) within the Statminer package and the Benjamini-Hochberg multiple test correction (Benjamini and Hochberg 1995). Each normalized Ct value, x_i, was transformed according to:

$$(-1 \text{ x } \log_b y_i) + 1$$

where $y_i = (x_i + |a| + 1)$, a is the minimum value in the range of initial normalized Ct values ($x_1, . . ., x_n$), and b is the maximum of the initially adjusted values ($x_i + |a| + 1, . . ., x_n + |a| + 1$). Accordingly, the highest normalized Ct value is scaled to 0 and the lowest to 1. Primers used are listed in supplementary table S1.16, Supplementary Material online.

RNA-seq Analysis

Ninety-six SRA files corresponding to 59 types of biological samples were retrieved from NCBI using the SRA Toolkit (Graveley et al. 2011; Brown et al. 2014). Reads with remaining adapters, with a percentage of N sites >10%, or with $\geq$ 50% nucleotides with a quality value Q $\leq$ 5 were discarded. One diagnostic motif, a sequence unique to a specific *Sdic* copy, for each of the *Sdic* copies (excluding *SdicB*, for which none could be found) was extended both upstream and downstream up to a total length of 130 nt. All reads from all libraries were then examined for a perfect alignment involving ~76 nt with each of the extended diagnostic motifs using TopHat 2.0.12 (Kim et al. 2013), making sure that the core diagnostic motif was always included. Raw counts per library were obtained using a custom shell script. The level of expression was estimated as the number of reads per kilobase per million reads (RPKM; Mortazavi et al. 2008), although in this case the variable length has no effect since all the motifs are 30 nt long. Within-biological-sample normalized expression values were subsequently log10 transformed. Heatmaps were generated by hierarchical clustering on principal components using FactoMineR (Lê et al. 2008; Diaz-Castillo et al. 2012). Expression specificity, s, was quantified as described (Yanai et al. 2005). For the Monte Carlo simulation analysis, log10 transformed normalized expression values were shuffled 10,000 times and s was recalculated each time for each copy. The resulting dataset allowed for calculating the probability of obtaining by chance alone a s larger or equal to the one observed.

MicroRNA Binding Site Composition

3'UTR sequences were extracted for all *Sdic* transcripts according to our annotation, and for all *sw* transcripts according to FlyBase (dos Santos et al. 2015). The presence of canonical microRNA sites (7mer-A1, 7mer-m8, and 8mer) as previously described (Bartel 2009), was examined using an in-home Perl script and the current microRNA annotation of *D. melanogaster* in miRBase v.21 (Kozomara and Griffiths-Jones 2014). Gains/losses of microRNA target sites were mapped to the *Sdic* phylogeny using the Dollo v3.695 parsimony method implemented in PHYLIP (Felsenstein 2005).

Phenotypic Assays

For the productivity assay, virgin females either possessing (A^+, I^+) or devoid (B^{-d}, E^{-d}) of the *Sdic* region of the *X* chromosome were crossed with naive wild-type males of the Oregon-R strain. Females from the strain w^{1118} were also used as a control for productivity levels of the source genetic background used to create the engineered strains used here (Yeh et al. 2012). Three naïve Oregon-R males were aged to 5 days old then mated to three 1-day-old virgin females from each of the experimental and control strains. Twenty-five replicates of each mating pair were assembled and the adult individuals were transferred to a fresh vial every other day. To compensate for decreasing male fecundity with age, males were removed on day 15 and replaced with another four males, which were in turn removed on day 29. The total progeny emerged from each vial associated with days 1, 3, 11, 13, 21, 31, and 33 was recorded. The progeny number produced was normalized by the number of females still alive at the moment of transferring from the vial associated with that particular day.

In the case of the egg-laying and egg-hatching assays, 10 five-day-old Oregon-R naïve males were mated separately to 10 virgin females of the same age from each of the five strains under comparison for 24 h. Three replicates of each of these crosses were set up. Petri dishes with grape-juice agar were used for easy egg detection against a dark background. To induce egg-laying, yeast was added to the agar (Waskar et al. 2005). Additionally, several scratches were made on the surface of the agar to increase surface area (Atkinson 1983). The adults of each replicate were transferred to a new plate every 24 h for 5 consecutive days and discarded on day 6. The egg number on each plate was recorded immediately after the adults were removed. After incubating for an additional 24 h, the plates were reexamined for unhatched eggs, the number of which was also recorded. These data was used to calculate the hatching rate and the number of eggs laid per female. JMP 12.1 (SAS Institute) was used for statistical analyses.

In Situ Hybridization

A ~4.23 kb *Sdic* genomic fragment present in all *Sdic* copies was generated by PCR and Sanger sequenced for verification. Probe labeling and hybridization on polytene chromosome squashes was performed as described (Ranz et al. 1997). Cytological analysis of the hybridizations was done using the photomap of *D. melanogaster* (Lefevre 1976) with a Nikon Eclipse 90i-automated microscope under phase contrast.

ACKNOWLEDGEMENTS

We thank Esther Betran, Brandon Gaut, Manyuan Long, John Parsch, and Therese Markow for comments; Carolus Chan, Kania Gandasetiawan, Felix Mesak, Andrei Tatarenkov, Hayden Tran, and EunBi Yang for technical help; and Rahul Warrior for providing us one stock. We are also in debt to the HPC Biocluster at UCI for grating us access to their resources. This work was supported by the National Science Foundation (MCB-1157876 to J.M.R.) and by the Ministerio de Economía y Competitividad of Spain (CGL2013-45211 to J.R.). E.S. and D.R. are grateful to the Bridges to the Baccalaureate Program supported by a National Institutes of Health grant (R25-

GM056647). With exception of one author, E.S., the rest declare to have no competing interests.

E.S. has stock ownership in Pacific Biosciences.

REFERENCES

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287:2185–2195.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25:3389–3402.

Assis R, Bachtrog D. 2013. Neofunctionalization of young duplicate genes in *Drosophila*. *Proc Natl Acad Sci U S A*. 110:17409–17414.

Atkinson WD. 1983. Gregarious oviposition in *Drosophila melanogaster* is explained by surface texture. *Aust J Zool*. 31:925–929.

Bariami V, Jones CM, Poupardin R, Vontas J, Ranson H. 2012. Gene amplification, ABC transporters and cytochrome P450s: unraveling the molecular basis of pyrethroid resistance in the dengue vector, *Aedes aegypti*. *PLoS Negl Trop Dis*. 6:e1692.

Bartel DP. 2009.MicroRNAs: target recognition and regulatory functions. Cell 136:215–233.

Bauters M, Van Esch H, Friez MJ, Boespflug-Tanguy O, Zenker M, Vianna-Morgante AM, Rosenberg C, Ignatius J, Raynaud M, Hollanders K, et al. 2008. Nonrecurrent *MECP2* duplications mediated by genomic architecture-driven DNA breaks and break induced replication repair. *Genome Res*. 18:847–858.

Benevolenskaya EV, Nurminsky DI, Gvozdev VA. 1995. Structure of the *Drosophila melanogaster annexin X* gene. *DNA Cell Biol*. 14:349–357.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol*. 57:289–300.

Bergthorsson U, Andersson DI, Roth JR. 2007. Ohno’s dilemma: evolution of new genes under continuous selection. *Proc Natl Acad Sci USA*. 104:17004–17009.

Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol*. 33:623–630.

Boylan K, Serr M, Hays T. 2000. A molecular genetic analysis of the interaction between the cytoplasmic dynein intermediate chain and the glued (dynactin) complex. *Mol Biol Cell*. 11:3791–3803.

Brookfield JF. 2001. Population genetics: the signature of selection. *Curr Biol*. 11:R388–3R390.

Brown CJ, Todd KM, Rosenzweig RF. 1998.Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Mol Biol Evol*. 15:931–942.

Brown DD, Wensink PC, Jordan E. 1972. A comparison of the ribosomal DNA’s of *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. *J Mol Biol* 63:57–73.

Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, et al. 2014. Diversity and dynamics of the *Drosophila* transcriptome. *Nature* 512:393–399.

Casola C, Ganote CL, Hahn MW. 2010. Nonallelic gene conversion in the genus *Drosophila*. *Genetics* 185:95–103.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* 13:238.

Chen JM, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet*. 8:762–775.

Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. *Nat Rev Genet*. 14:645–660.

Chen S, Zhang YE, Long M. 2010. New genes in *Drosophila* quickly become essential. *Science* 330:1682–1685.

Diaz-Castillo C, Xia XQ, Ranz JM. 2012. Evaluation of the role of functional constraints on the integrity of an ultraconserved region in the genus *Drosophila*. *PLoS Genet*. 8:e1002475.

dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, Emmert DB, Gelbart WM, FlyBase C. 2015. FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res*. 43:D690–D697.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–138.

Fawcett JA, Innan H. 2011. Neutral and non-neutral evolution of duplicated genes with gene conversion. *Genes* (Basel) 2:191–209.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

FitzGerald PC, Sturgill D, Shyakhtenko A, Oliver B, Vinson C. 2006. Comparative genomics of *Drosophila* and human core promoters. *Genome Biol*. 7:R53.

Gibson JR, Chippindale AK, Rice WR. 2002. The *X* chromosome is a hot spot for sexually antagonistic fitness variation. *Proc Biol Sci*. 269:499–505.

Gnad F, Parsch J. 2006. Sebida: a database for the functional and evolutionary analysis of genes with sex-biased expression. *Bioinformatics* 22:2577–2579.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van BarenMJ, Boley N, Booth BW, et al. 2011. The developmental transcriptome of *Drosophila melanogaster*. *Nature* 471:473–479.

Hahn MW. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered*. 100:605–617.

Hahn MW, Han MV, Han SG. 2007. Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet*. 3:e197.

Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet*. 39:1140–1144.

Hemingway J, Hawkes NJ, McCarroll L, Ranson H. 2004. The molecular basis of insecticide resistance in mosquitoes. *Insect Biochem Mol Biol*. 34:653–665.

Hoskins RA, Carlson JW, Kennedy C, Acevedo D, Evans-Holm M, Frise E, Wan KH, Park S, Mendez-Lago M, Rossi F, et al. 2007. Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. *Science* 316:1625–1628.

Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, Sudmant PH, Graves TA, Alkan C, Dennis MY, et al. 2014. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res*. 24:688–696.

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* 256:119–124.

Innan H. 2003. A two-locus gene conversion model with selection and its application to the human *RHCE* and *RHD* genes. *Proc Natl Acad Sci USA*. 100:8793–8798.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet*. 11:97–108.

Innocenti P, Morrow EH. 2010. The sexually antagonistic genes of *Drosophila melanogaster*. *PLoS Biol*. 8:e1000335.

Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. 2008. NCBI BLAST: a better web interface. *Nucleic Acids Res*. 36:W5–W9.

Jugulam M, Niehues K, Godar AS, Koo DH, Danilova T, Friebe B, Sehgal S, Varanasi VK, Wiersma A, Westra P, et al. 2014. Tandem amplification of a chromosomal segment harboring *5-enolpyruvylshikimate-3-phosphate synthase* locus confers glyphosate resistance in *Kochia scoparia*. *Plant Physiol*. 166:1200–1207.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res*. 20:1313–1326.

Katju V. 2012. In with the old, in with the new: the promiscuity of the duplication process engenders diverse pathways for novel gene creation. *Int J Evol Biol*. 2012:341932.

Katju V, Bergthorsson U. 2013. Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Front Genet*. 4:273.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 14:R36.

Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, Chin CS, Rapicavoli NA, Rank DR, Li J, et al. 2014. Long-read, whole-genome shotgun sequence data for five model organisms. *Sci Data*. 1:140045.

Kleene KC. 2005. Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells. *Dev Biol*. 277:16–26.

Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc Biol Sci*. 279:5048–5057.

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*. 42:D68–D73.

Krsticevic FJ, Schrago CG, Carvalho AB. 2015. Long-read single molecule sequencing to resolve tandem gene copies: the *Mst77Y* region on the *Drosophila melanogaster Y* chromosome. *G3* (Bethesda) 5:1145–1150.

Kulathinal RJ, Sawyer SA, Bustamante CD, Nurminsky D, Ponce R, Ranz JM, Hartl DL. 2004. Selective sweep in the evolution of a new sperm-specific gene in *Drosophila*. In: Nurminsky D, editor. Selective Sweep. Austin, Texas: Kluwer Academic/Plenum Publishers. p. 1–12.

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

Lê S, Josse J, Husson F. 2008. FactoMineR: an R package for multivariate analysis. *J Stat Softw*. 25:1–18.

Lefevre G. 1976. A photographic representation and interpretation of the polytene chromosomes of *Drosophila melanogaster* salivary glands. In: Ashburner MA, Novitski E, editors. The Genetics and Biology of *Drosophila*. London: Academic Press. p. 31–66.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451–1452.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 25:402–408.

Long M, VanKuren NW, Chen S, Vibranovski MD. 2013. New gene evolution: little did we know. *Annu Rev Genet*. 47:307–333.

Ma S, Trivinos-Lagos L, Graf R, Chisholm RL. 1999. Dynein intermediate chain mediated dynein-dynactin interaction is required for interphase microtubule organization and centrosome replication and separation in *Dictyostelium*. *J Cell Biol*. 147:1261–1274.

Mayer MG, Rodelsperger C, Witte H, Riebesell M, Sommer RJ. 2015. The orphan gene *dauerless* regulates Dauer development and intraspecific competition in nematodes by copy number variation. *PLoS Genet*. 11:e1005146.

McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, Petrov DA, Fiston-Lavier AS. 2014. Illumina TruSeq synthetic longreads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PLoS One* 9:e106689.

McGrath CL, Casola C, Hahn MW. 2009. Minimal effect of ectopic gene conversion among recent duplicates in four mammalian genomes. *Genetics* 182:615–622.

Michiels F, Gasch A, Kaltschmidt B, Renkawitz-Pohl R. 1989. A 14 bp promoter element directs the testis specificity of the *Drosophila beta 2 tubulin* gene. *EMBO J*. 8:1559–1565.

Mikhaylova LM, Nurminsky DI. 2011. Lack of global meiotic sex chromosome inactivation, and paucity of tissue-specific gene expression on the *Drosophila X* chromosome. *BMC Biol*. 9:29.

Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D. 2013. Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform*. 14:193–202.

Moore RC, Purugganan MD. 2003. The early stages of duplicate gene evolution. *Proc Natl Acad Sci USA*. 100:15682–15687.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 5:621–628.

Newcomb RD, Gleeson DM, Yong CG, Russell RJ, Oakeshott JG. 2005. Multiple mutations and gene duplications conferring organophosphorus insecticide resistance have been selected at the *Rop-1* locus of the sheep blowfly, *Lucilia cuprina*. *J Mol Evol*. 60:207–220.

Nurminsky DI, Nurminskaya MV, Benevolenskaya EV, Shevelyov YY, Hartl DL, Gvozdev VA. 1998a. Cytoplasmic dynein intermediate chain isoforms with different targeting properties created by tissue-specific alternative splicing. *Mol Cell Biol*. 18:6816–6825.

Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL. 1998b. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature* 396:572–575.

O’Fallon BD 2013. ACG: rapid inference of population history from recombining nucleotide sequences. *BMC Bioinformatics* 14:40.

Obbard DJ, Maclennan J, Kim KW, Rambaut A, O’Grady PM, Jiggins FM. 2012. Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. *Mol Biol Evol*. 29:3459–3473.

Ohno S. 1970. Evolution by Gene Duplication. New York: Springer-Verlag.

Osada N, Innan H. 2008. Duplication and gene conversion in the *Drosophila melanogaster* genome. *PLoS Genet*. 4:e1000305.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet*. 39:1256–1260.

Ponce R, Hartl DL. 2006. The evolution of the novel *Sdic* gene cluster in *Drosophila melanogaster*. Gene 376:174–183.

Ranz JM, Segarra C, Ruiz A. 1997. Chromosomal homology and molecular organization of Muller's elements D and E in the *Drosophila repleta* species group. *Genetics* 145:281–295.

Rice WR. 1984. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* 38:735–742.

Sassone-Corsi P. 2002. Unique chromatin remodeling and transcriptional regulation in spermatogenesis. *Science* 296:2176–2178.

Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol Biol Evol*. 6:526–538.

Schmidt EE, Schibler U. 1995. High accumulation of components of the RNA polymerase II transcription machinery in rodent spermatids. *Development* 121:2373–2383.

Singh RS, Kulathinal RJ. 2005. Male sex drive and the masculinization of the genome. *Bioessays* 27:518–525.

Slightom JL, Chang LY, Koop BF, Goodman M. 1985. Chimpanzee fetal *G gamma* and *A gamma globin* gene nucleotide sequences provide further evidence of gene conversions in hominine evolution. *Mol Biol Evol*. 2:370–389.

Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 3:Article3.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol*. 30:2725–2729.

Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. *Nat Rev Genet*. 12:692–702.

Walsh JB. 1987. Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? *Genetics* 117:543–557.

Wang S, Adams KL. 2015. Duplicate gene divergence by changes in microRNA binding sites in *Arabidopsis* and *Brassica*. *Genome Biol Evol*. 7:646–655.

Waskar M, Li Y, Tower J. 2005. Stem cell aging in the *Drosophila* ovary. *Age (Dordr)* 27:201–212.

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, et al. 2005. Genomewide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21:650–659.

Yeh SD, Do T, Chan C, Cordova A, Carranza F, Yamamoto EA, Abbassi M, Gandasetiawan KA, Librado P, Damia E, et al. 2012. Functional evidence that a recently evolved *Drosophila* sperm-specific gene boosts sperm competition. *Proc Natl Acad Sci USA*. 109:2043–2048.

Yeh SD, von Grotthuss M, Gandasetiawan KA, Jayasekera S, Xia XQ, Chan C, Jayaswal V, Ranz JM. 2014. Functional divergence of the miRNA transcriptome at the onset of *Drosophila* metamorphosis. *Mol Biol Evol*. 31:2557–2572.

Zhao L, Saelao P, Jones CD, Begun DJ. 2014.Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science* 343:769–772.

Zimmer EA, Martin SL, Beverley SM, Kan YW, Wilson AC. 1980. Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proc Natl Acad Sci USA*. 77:2158–2162.

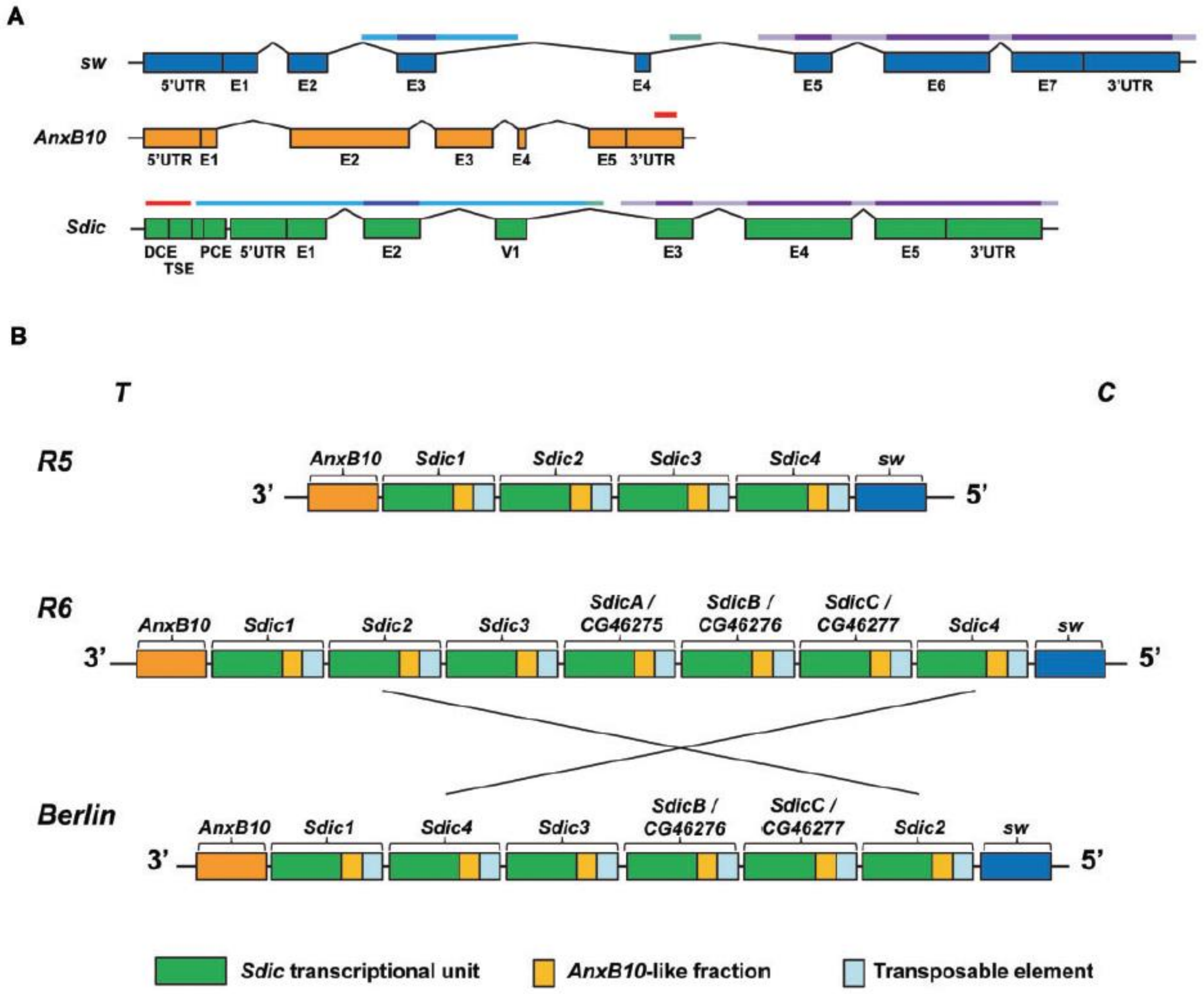


Figure 1.1. Organizational features of the *Sdic* region of *D. melanogaster*. (A) Sequence stretches of the parental genes *sw* and *AnxB10* that contribute to the structure of the chimeric protein-coding gene *Sdic*. Top colored bars denote sequence stretches from parental genes that correspond to sequence stretches in *Sdic*. Dark and light tones, exonic and intronic sequence in *sw* respectively. (B) Different organization of the *Sdic* region in three assemblies of the *D. melanogaster* genome in the ISO1 strain. The *Sdic* cluster is composed of tandem repeats, each consisting of three parts: *Sdic*, originated primarily from stretches of *sw*; another putative transcriptional unit originated from *AnxB10* named *AnxB10*-like; and a ~785 nt stretch from the transposable element Rt1c (Nurminsky et al. 1998b; Ponce and Hartl 2006). The relative location (black lines) and number of repeats vary between assemblies, which determine the size of the region: ~31 kb in Release 5 (R5); ~46 kb in the assembly GCA_000778455.1 (Berlin); and ~54 kb in Release 6 (R6). T, telomere; C, centromere. Distances and lengths of different features are not to a scale.

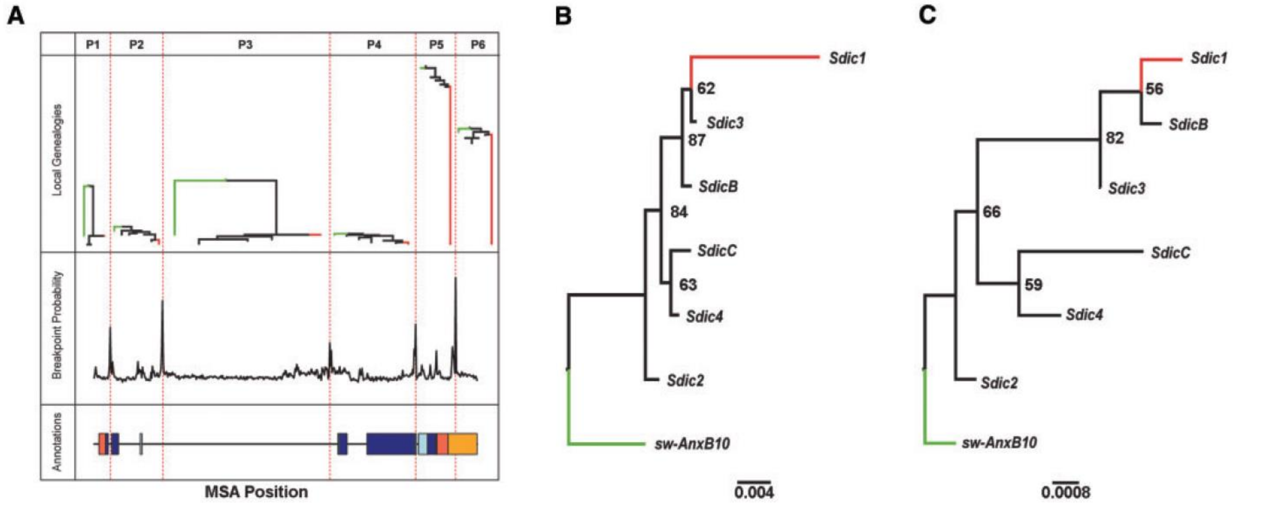


Figure 1.2. Molecular evolution of the *Sdic* multigene family. (A) Top, local gene genealogies for each of the six DNA partitions (labeled by P1–P6) inferred with ACG. The DNA stretches from the different partitions are separated by recombination breakpoints depicted by a red dashed line. Using the exon–intron annotations of all copies except *Sdic4* as a reference, and after omitting stretches of sequence associated with deletions, partitions P5 harbors 11 non-synonymous and eight synonymous substitutions; partitions P1–P4 harbor 5 and 3, respectively. P6 does not include *Sdic4*, as this copy only contains missing data in this region. Middle panel, breakpoint posterior probability as estimated by ACG. Bottom panel, summarization of the exon–intron boundaries of *Sdic* following the color code in supplementary fig. S1.8, Supplementary Material online. MSA, multiple sequence alignment. (B) Maximum Likelihood phylogeny of the *Sdic* multigene family members, using a composite sequenced comprised of the homologous *sw* and *AnxB10* (*sw-AnxB10*) as an outgroup. The numbers in the internal nodes indicate the bootstrap support after 1,000 replicates. (C) Up-close view of the gene genealogy for the P4 partition. This partition has likely not exchanged information by gene conversion or been affected by other evolutionary forces that could potentially obscure the true duplication history of the *Sdic* gene copies. Local gene genealogies are represented with FigTree (http://tree.bio.ed.ac.uk/software/figtree/; last accessed December 1, 2015). Branches colored in red and green highlight *Sdic1* and *sw-AnxB10*, respectively. Scale bars indicate the number of nucleotide substitutions per site.

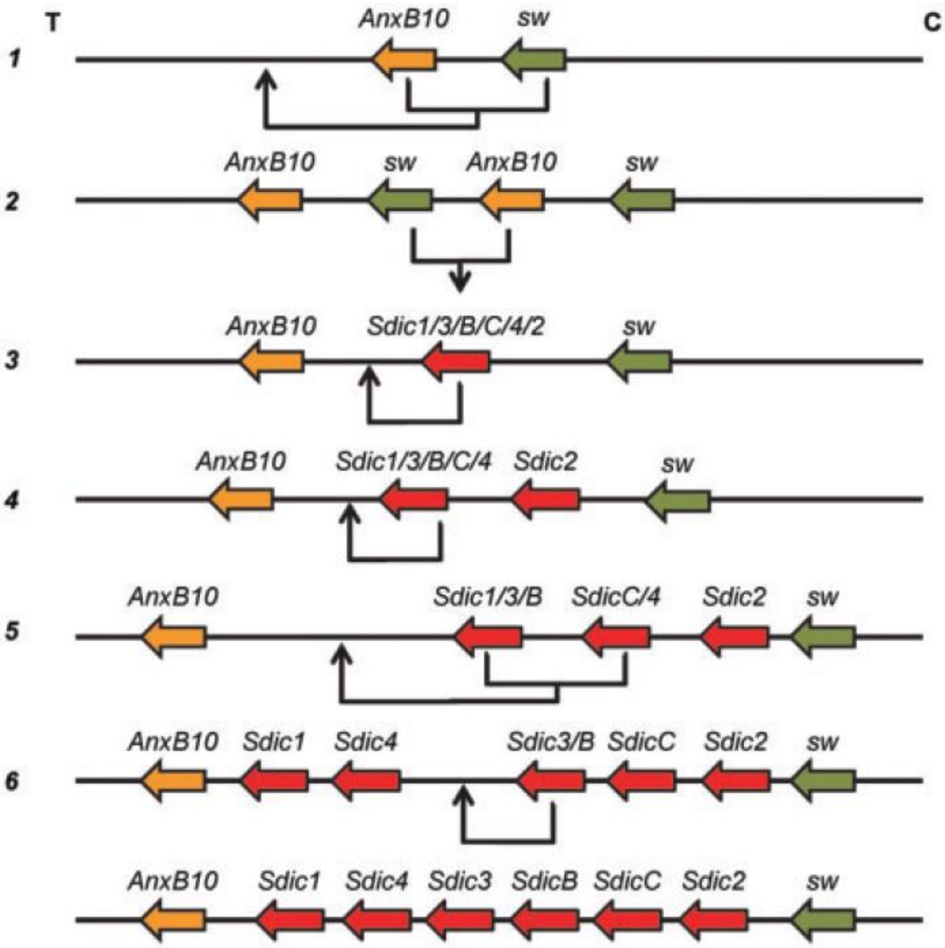


Figure 1.3. Most parsimonious reconstruction of the formation of the *Sdic* region. An unequal crossing-over event between regions upstream of *sw* and downstream of *AnxB10* resulted in a segmental duplication of *sw* and *AnxB10*, although other more complex rearrangement scenarios cannot be ruled out (Bauters et al. 2008) (1). This was followed by the creation of the ancestral *Sdic* copy (*Sdic1/3/B/C/4/2*) through a series of mutations, which notably involved a large deletion event involving the middle copies of *sw* and *AnxB10* (2); a TE also became inserted upstream of the ancestral *Sdic* copy (data not shown). An unequal crossing-over event involving sequence stretches upstream and downstream of the ancestral *Sdic*, but in different homologous chromosomes, would have then resulted in a tandem duplication of the ancestral *Sdic* copy (3). Next, a similar unequal crossing-over event resulted in the tandem duplication of the *Sdic* copy closest to *AnxB10* (4). Subsequently, a third unequal crossing-over event occurred amid the region between *AnxB10* and its closest copy and the region between the two copies closest to *sw* resulting in a tandem duplication of the two copies closest to *AnxB10* (5). Finally, a fourth unequal crossing-over event resulted in a single-copy tandem duplication leading to the formation of the sixth *Sdic* copy (6). Several gene conversion events have likely occurred between *Sdic* copies. After step 3, it is uncertain where the unequal crossing-over events occurred due to the high similarity of the copies. This proposed scenario is in overall good agreement with the phylogenetic tree in fig. 1.2C, with the exception of the sequential generation of *Sdic1*, *Sdic3*, and *SdicB*. Nevertheless, this tree exhibits low bootstrap values. Black arrows, duplication events. T, telomere; C, centromere.

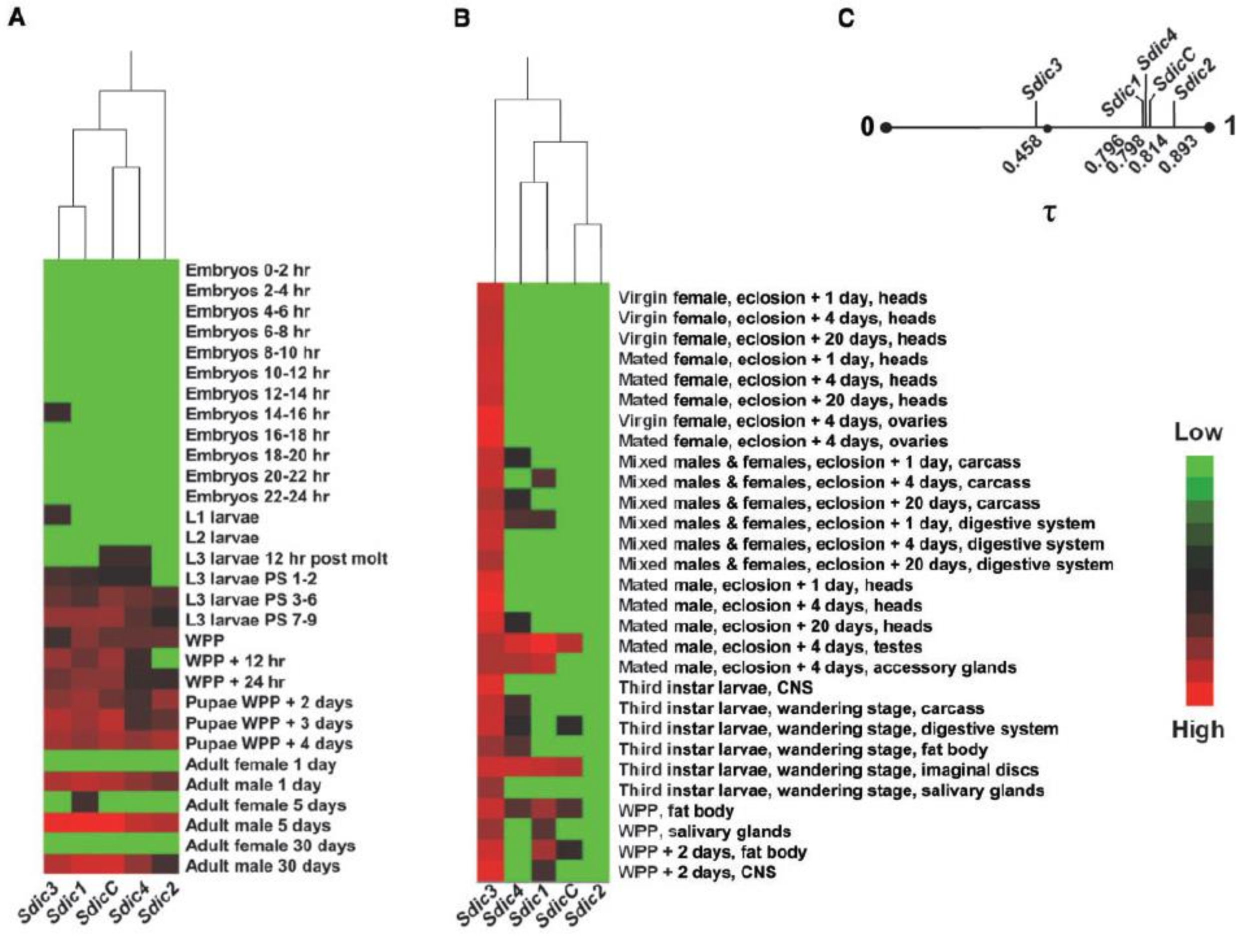
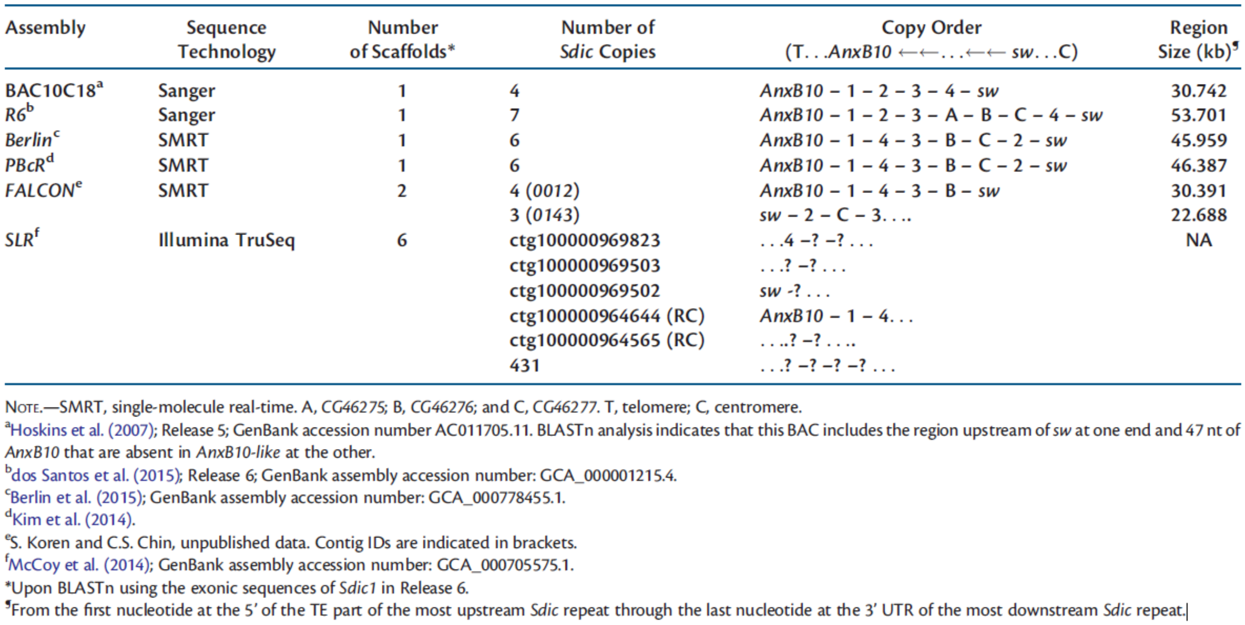


Figure 1.4. Expression profile of five *Sdic* copies. Heatmap for developmental stages (A) and anatomical samples (B) showing evidence of expression diversification among the *Sdic* copies surveyed. Red, high expression; black, intermediate expression; green, lower expression. Fifty-nine biological conditions were examined. The data were obtained in two different large-scale expression surveys (Graveley et al. 2011; Brown et al. 2014), which might differ in their power to detect lowly expressed transcripts, even in similar, although not identical, conditions. (C) Expression specificity, τ, upon considering all conditions. τ values range from 0 to 1, with higher values corresponding to more restricted expression and lower values to broader expression across conditions (Yanai et al. 2005). Log10 normalized expression values were used in the analyses. Examples of the detected reads in relevant conditions are provided in supplementary fig. S1.14, Supplementary Material online. CNS, central nervous system; hr, hour; Lx, larval stage x; PS, puff stage; WPP, white prepupae.

Table 1.1 Organization of the *Sdic* Region of *D. melanogaster* in Different Assemblies.



Assembly	Sequence Technology	Number of Scaffolds*	Number of *Sdic* Copies	Copy Order (T. . .*AnxB10* ←←. . .←← *sw*. . .C)	Region Size (kb)¶
BAC10C18[a]	Sanger	1	4	*AnxB10* – 1 – 2 – 3 – 4 – *sw*	30.742
R6[b]	Sanger	1	7	*AnxB10* – 1 – 2 – 3 – A – B – C – 4 – *sw*	53.701
Berlin[c]	SMRT	1	6	*AnxB10* – 1 – 4 – 3 – B – C – 2 – *sw*	45.959
PBcR[d]	SMRT	1	6	*AnxB10* – 1 – 4 – 3 – B – C – 2 – *sw*	46.387
FALCON[e]	SMRT	2	4 (*0012*)	*AnxB10* – 1 – 4 – 3 – B – *sw*	30.391
			3 (*0143*)	*sw* – 2 – C – 3. . ..	22.688
SLR[f]	Illumina TruSeq	6	ctg100000969823	. . .4 –? –? . . .	NA
			ctg100000969503	. . .? –? . . .	
			ctg100000969502	*sw* -? . . .	
			ctg100000964644 (RC)	*AnxB10* – 1 – 4. . .	
			ctg100000964565 (RC)	. . ..? –?	
			431	. . .? –? –? –? . . .	

NOTE.—SMRT, single-molecule real-time. A, *CG46275*; B, *CG46276*; and C, *CG46277*. T, telomere; C, centromere.

[a] Hoskins et al. (2007); Release 5; GenBank accession number AC011705.11. BLASTn analysis indicates that this BAC includes the region upstream of *sw* at one end and 47 nt of *AnxB10* that are absent in *AnxB10-like* at the other.

[b] dos Santos et al. (2015); Release 6; GenBank assembly accession number: GCA_000001215.4.

[c] Berlin et al. (2015); GenBank assembly accession number: GCA_000778455.1.

[d] Kim et al. (2014).

[e] S. Koren and C.S. Chin, unpublished data. Contig IDs are indicated in brackets.

[f] McCoy et al. (2014); GenBank assembly accession number: GCA_000705575.1.

*Upon BLASTn using the exonic sequences of *Sdic1* in Release 6.

¶From the first nucleotide at the 5′ of the TE part of the most upstream *Sdic* repeat through the last nucleotide at the 3′ UTR of the most downstream *Sdic* repeat.

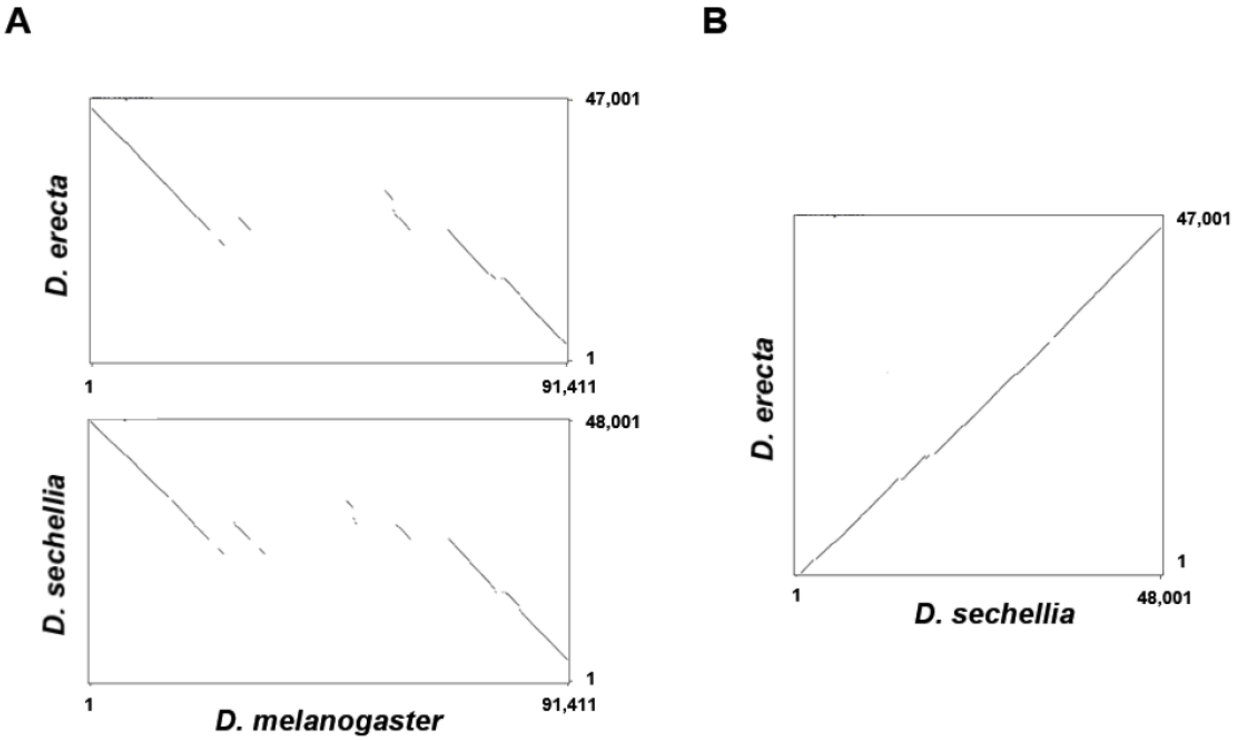
SUPPLEMENTARY TEXT

Of the four assemblies analyzed, two, a preassembly obtained using SMRT sequencing reads (S. Koren and C.S. Chin, unpublished data; see Material and Methods) and an independent assembly obtained using Illumina TruSeq Synthetic Long-Reads (SLRs) (McCoy, et al. 2014) provided a fragmented representation of the *Sdic* region. This fragmentation complicated both the annotation of the copies and the inference of their relative order, which was exacerbated by the absence of some diagnostic stretches of DNA in some of the copies (table 1.1). In the case of the SLRs, which are assembled from unique, small, barcoded sequence stretches obtained from an originally longer DNA fragment, one possibility is that reads containing the same portion of two different *Sdic* copies could be missassembled. This possibility seems to apply to the limited reconstruction of the *Mst77Y* region in *D. melanogaster* using this technology (Krsticevic, et al. 2015).

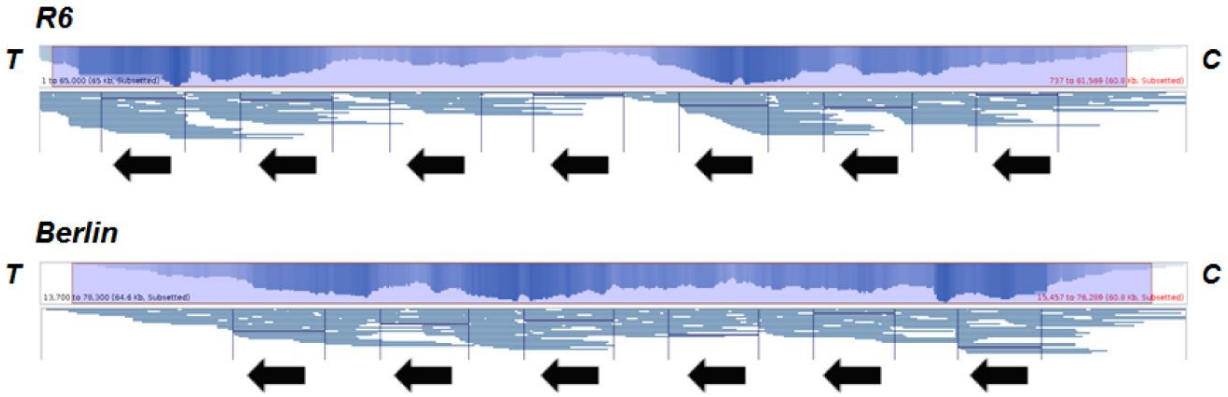
Relative to the use of SLRs as benchmark to validate the reconstruction of the *Sdic* region in the Release 6 and *Berlin* assemblies, the alignments between selected SLRs and the two assemblies under scrutiny were inspected in two complementary ways. In the first, we examined nucleotide mismatches and indels categorized according to their size (≤4 nt versus >4 nt) (Supplementary fig. S1.4A). In the second, we extracted the identity scores directly from the BLASR alignments, which can be found in the m4 output files. In this case, we created four frequency classes for the resulting identity scores found in the alignments against the two assemblies and evaluated whether the observed values could be obtained by chance alone (Supplementary fig. S1.4B). We concluded that the frequency classes of alignments involving lower sequence identity scores are significantly more populated by alignments between SLRs and

Release 6 than between SLRs and the *Berlin* assembly (randomization test of goodness-of-fit, P_{adj}<0.001).

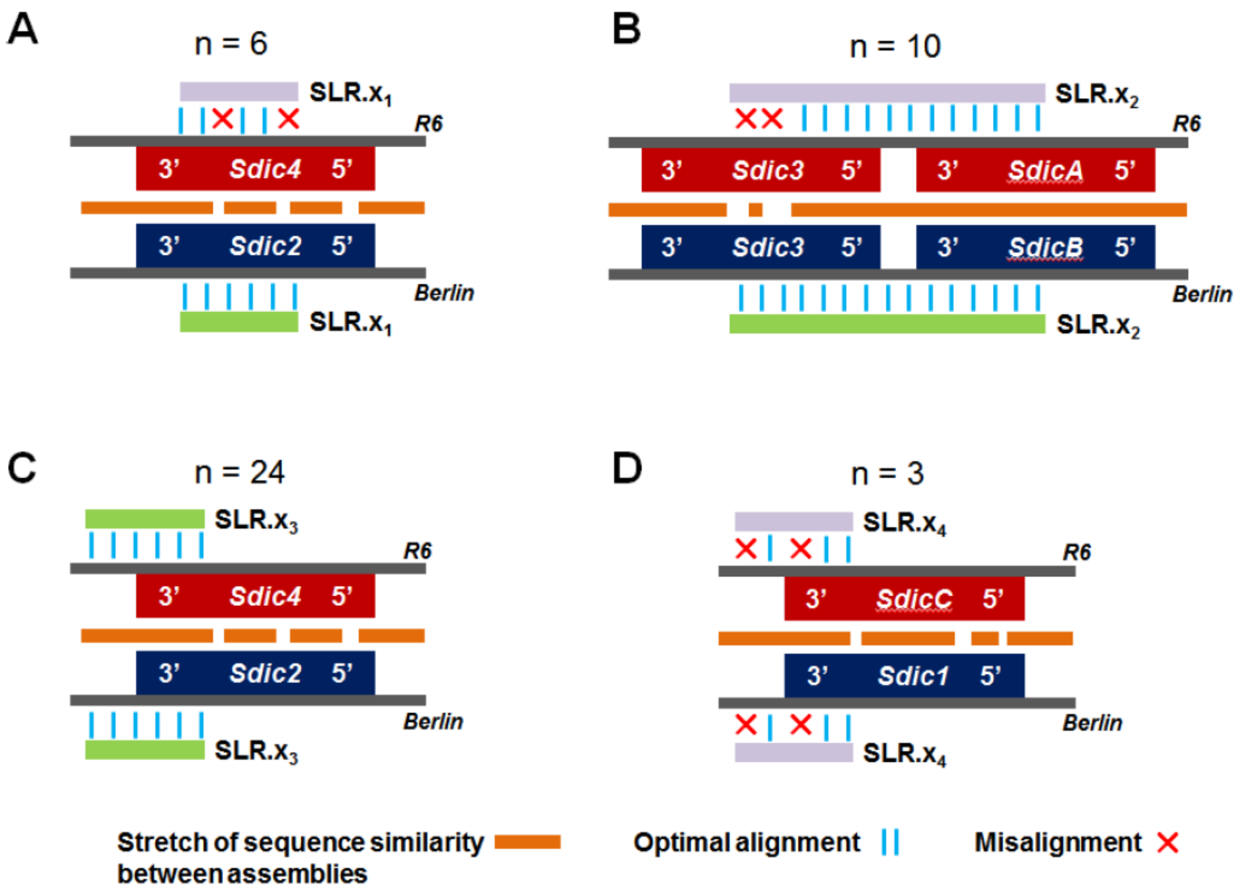
To further evaluate the accuracy of the reconstruction of the *Sdic* region in the Release 6 and *Berlin* assemblies, we examined the support that SLRs provided to two types of diagnostic sequence stretches. Specifically, we found no SLR supporting the distinctive junctions *Sdic1-Sdic2, Sdic2-Sdic3,* and *SdicC-Sdic4* of Release 6. Conversely, the junctions *Sdic1-Sdic4, Sdic4-Sdic3,* and *SdicC-Sdic2* present in the *Berlin* assembly were supported by 4, 4, and 2 SLRs, respectively. Further, we examined the support received by same-copy differences. Precisely, *Sdic1* and *Sdic4* appear to have exchanged a diagnostic 53 nt insert between the two copies in the two assemblies (supplementary table S1.4). In addition, an 18 nt insert and another 8 nt long section harboring 3 nt differences are swapped between the *Sdic1* copies of the two assemblies. We found 21 SLRs that supported the association of the mentioned diagnostic stretches in *Sdic1* (15) and *Sdic4* (6) as they appear in the *Berlin* assembly, while no reads supported these diagnostic stretches as they appear in Release 6.



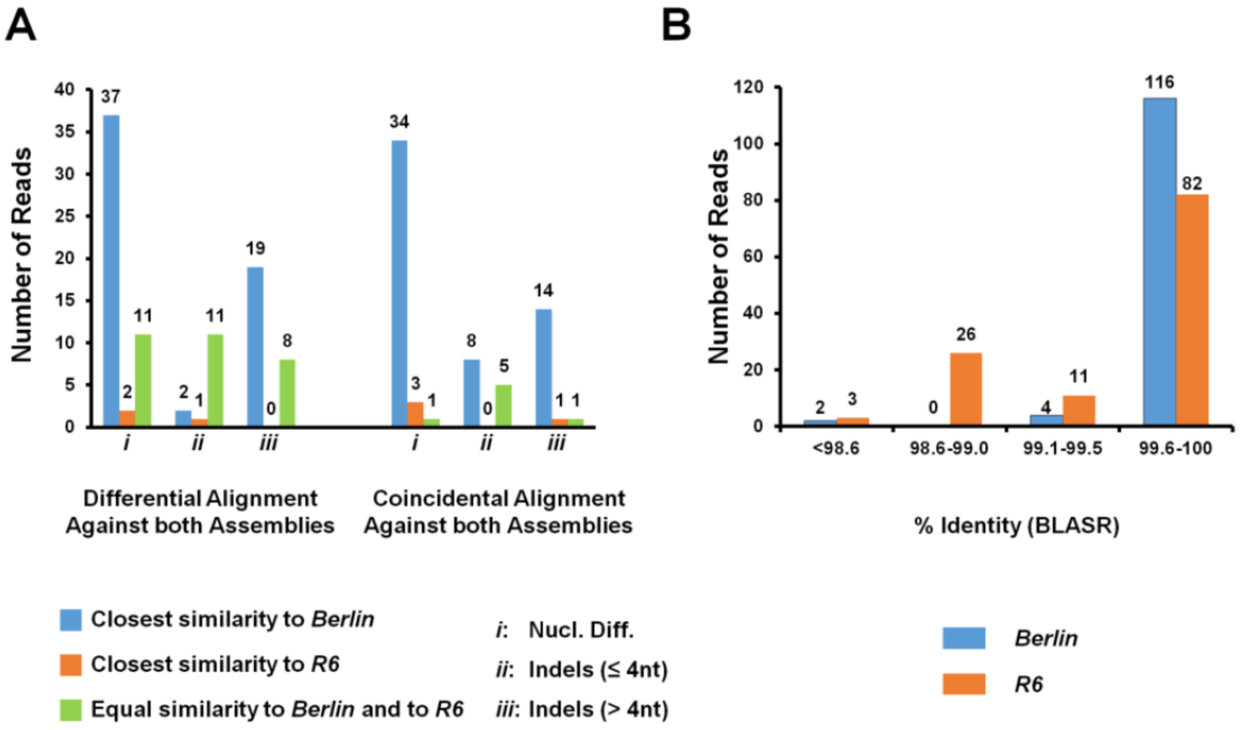
Supplementary Figure S1.1. Dot plot for the chromosomal region from *CG9572* to *CG17068* between *D. melanogaster, D. erecta,* and *D. sechellia.* *D. erecta* and *D. sechellia* also belong to the *D. melanogaster* species group, representing different phylogenetic distances in relation to *D. melanogaster,* and are collinear with *D. melanogaster* for this genomic region (von Grotthuss, et al. 2010). This collinearity is interrupted by the extra ~45 kb in the *D. melanogaster* genome that corresponds to the tandemly-repeated *Sdic* multigene family (A). In contrast, a marked collinearity can be observed for this region between *D. erecta* and *D. sechellia* (B). The dot plots were created with PipMaker (Schwartz, et al. 2000). The *D. melanogaster* sequence used was extracted from the assembly GCA_000778455.1 (Berlin, et al. 2015) while the sequences from *D. erecta* and *D. sechellia* were extracted from the Release 1.05 (scaffold_4690: 10170000..10217000) and Release 1.3 (scaffold_8: 2268000..2316000) respectively.



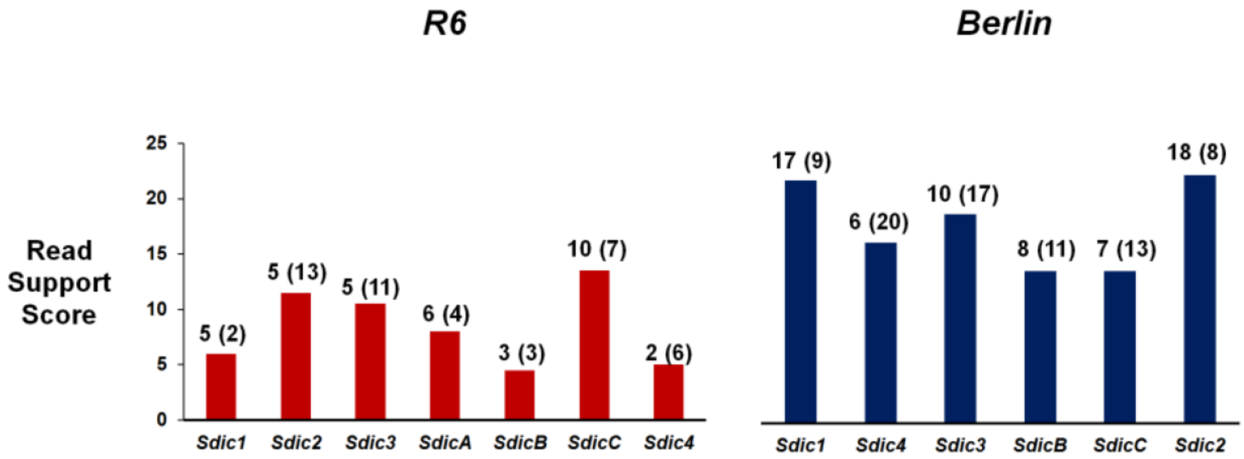
Supplementary Figure S1.2. Mapping of 122 Illumina TruSeq Synthetic Long Reads (SLRs) against the *Sdic* region. *R6*, Release 6; *Berlin*, assembly GCA_000778455.1. The position of the putative *Sdic* copies present in the two assemblies is denoted under each assembly with black arrows (fig. 1.1B for details about the identity of the copies). *T*, telomere; *C*, centromere.



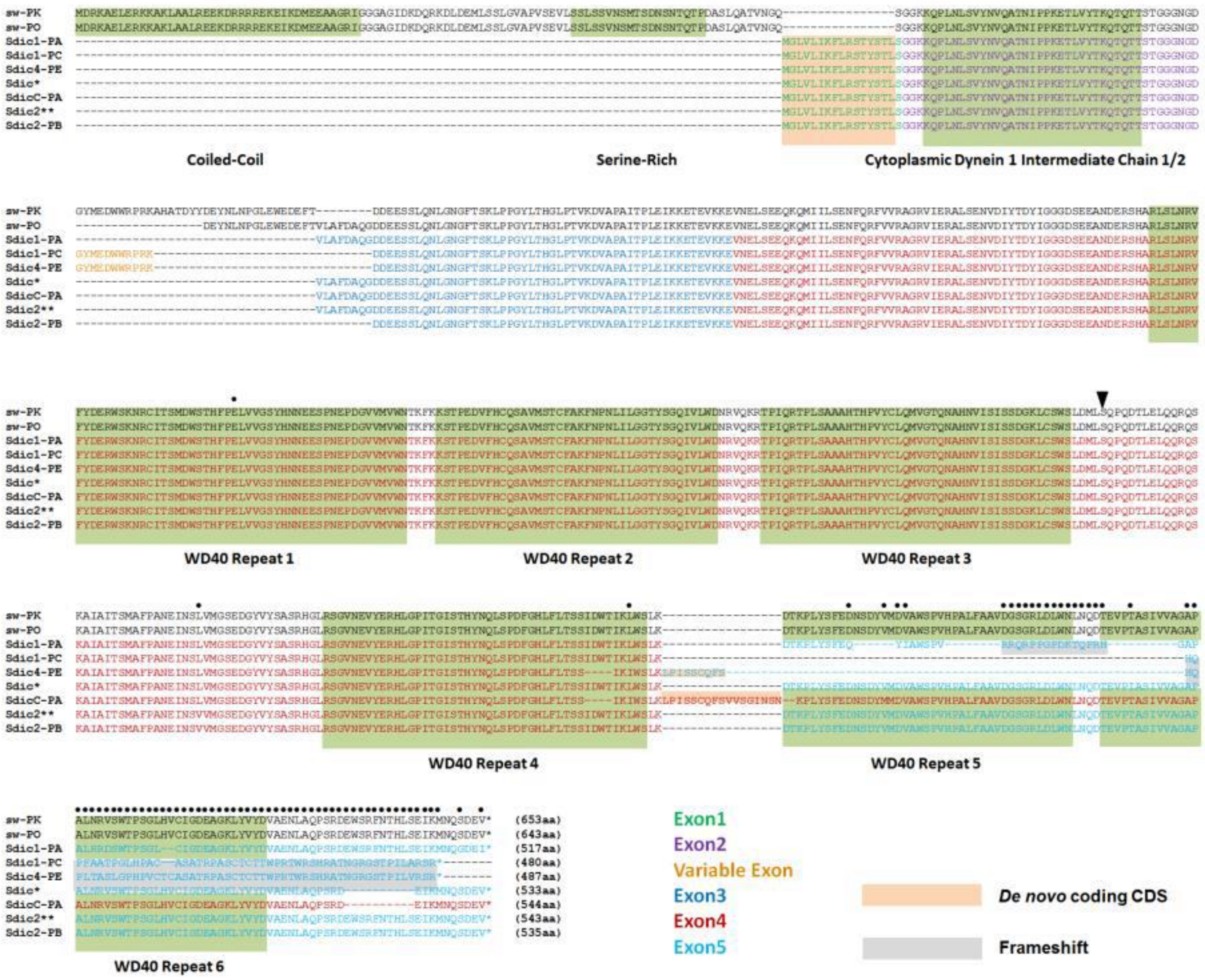
Supplementary Figure S1.3. Types of discrepancies seen in the alignment of Illumina TruSeq Synthetic Long Reads (SLRs) to two different assemblies when compared at the *Sdic* region. All types involve alignments with different *Sdic* repeats in the two assemblies and fall into three main categories; the number of reads falling into each category is shown on top. The first category includes reads that show more similarity with one of the assemblies than the other, including two types of reads: those encompassing stretches from a single *Sdic* repeat (A) and those encompassing stretches from two *Sdic* repeats (B). In type B, the discrepancy is ultimately associated with the different order of some of the *Sdic* repeats between assemblies. The second category includes reads that span a region of the repeat that is ≥99.9% identical between different copies in the two assemblies (C). The third category corresponds to reads that show an equally non-optimal alignment against both assemblies, aligning to two different *Sdic* repeats in the two assemblies (D). *R6*, Release 6; *Berlin*, assembly GCA_000778455.1. *x*i denotes a particular SLR.



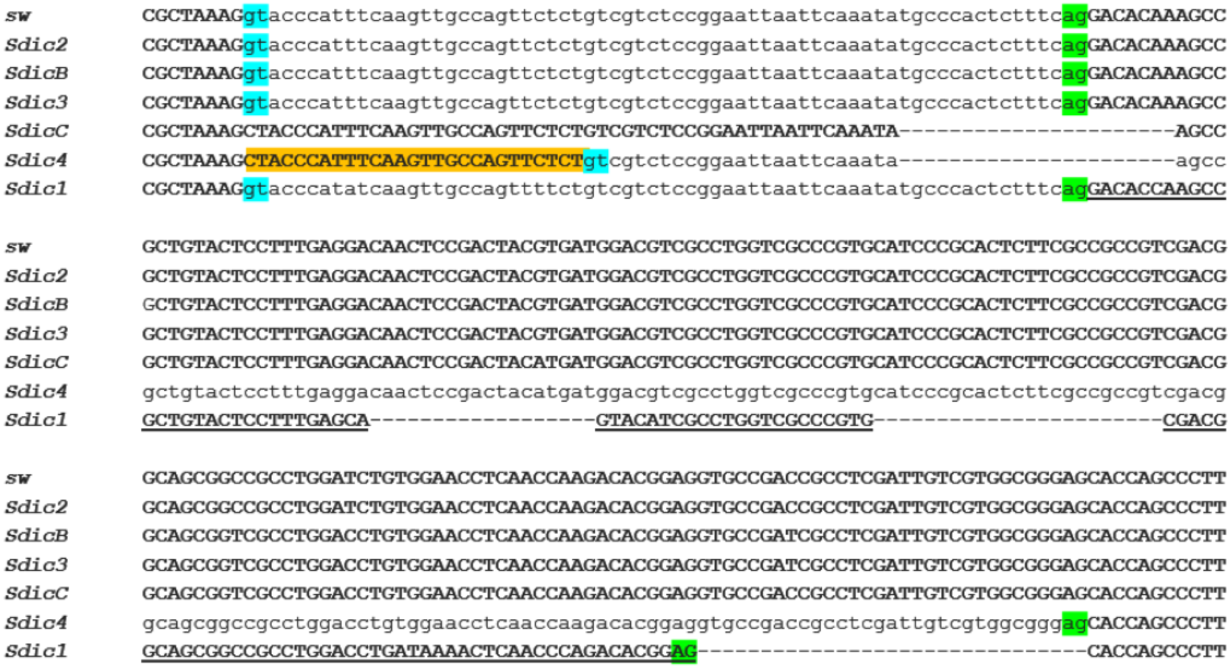
Supplementary Figure S1.4. Differential support of Illumina TruSeq Synthetic Long Reads (SLRs) to two existing reconstructions of the *Sdic* region. (A) Read categorization based on how optimal the read alignment is against the *Berlin* assembly and Release 6 when nucleotides mismatches and indels of different size are considered (i-iii). Overall, SLRs harbored substantially fewer differences when compared to the *Berlin* assembly than to Release 6, as indicated by the higher read number (blue). Whether or not, left and right respectively, the read mapped to the same location within the *Sdic* region in both assemblies had no impact. Some reads could be impacted by several types of sequence differences and a few others showed no preferential alignment for either assembly. (B) Read number distribution based on percent sequence identity in the alignments of SLRs against the *Berlin* assembly and Release 6. Overall, SLRs tend to show higher levels of sequence identity against *Berlin* than against Release 6 (see Supplementary text). Sequence identity values were extracted from the m4 alignment files. *R6*, Release 6; *Berlin*, assembly GCA_000778455.1.



Supplementary Figure S1.5. Differential coverage of the *Sdic* repeat region in two assemblies when aligned to Illumina TruSeq Synthetic Long Reads (SLRs) using BLASR. Only reads showing high quality alignments with at least one of the assemblies were considered. A high-quality alignment implies ≥99.9% sequence identity with a given reference assembly according to BLASR and an extension of the alignment of ≥99.9%. The number of SLRs supporting only a particular copy appears on top of each bin (*a*), with those supporting a second copy (*b*) in parenthesis. The read support score equals *a* + (*b*/2). A greater number of reads were shown to have high quality alignments against GCA_000778455.1 (*Berlin*) compared to Release 6 (*R6*). In addition, the *Sdic* copies display more even coverage across the *Sdic* region for the *Berlin* assembly than *R6*.

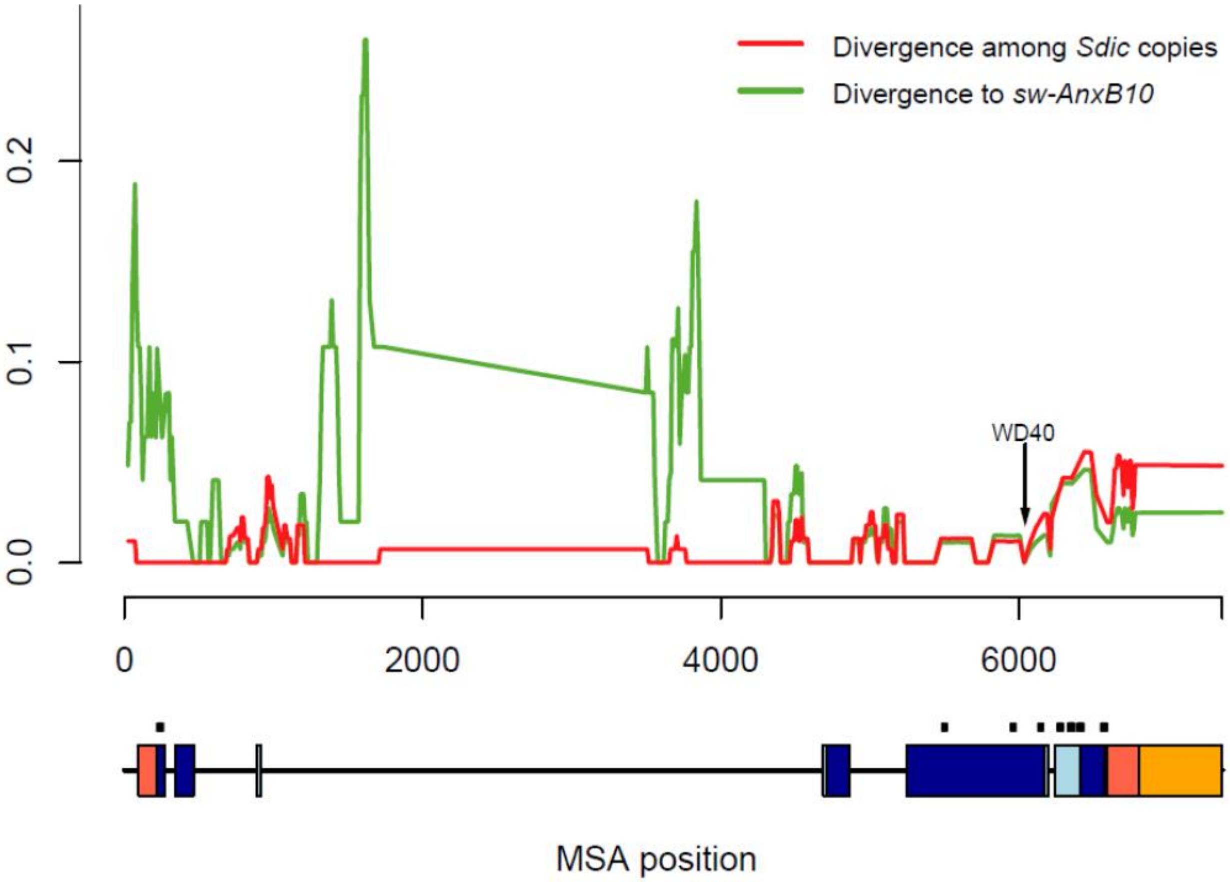


Supplementary Figure S1.6. Sdic protein sequence alignment. Amino acids stretches from different *Sdic* exons are color coded; protein motifs as delineated by INTERPRO (Mitchell, et al. 2015) are highlighted in green boxes. All Sdic variants include one *cytoplasmic dynein 1 intermediate chain 1/2* domain and 4 or 6 WD40 domains. The three most downstream WD40 motifs can include differences among variants or be absent altogether leading to shorter variants of SDIC; Sdic1-PC and Sdic4-PE are just 480 and 487 amino acids long, respectively. Solid dots indicate positions with variable amino acids across sequences, using two sw isoforms as a reference. These two splice variants isoforms display all the sequence positions that could be aligned with Sdic. *, identical amino acid sequence for variants Sdic3-PE, Sdic3-PF, Sdic3-PG, and SdicB-PA. **, identical amino acid sequence for variants Sdic2-PA and Sdic2-PC. The length of each protein variant is listed at the end of its sequence. Salmon box, amino acids corresponding to ancestrally intronic sequences in *sw* that are now coding in particular *Sdic* transcripts. Grey box, amino acid stretches associated with frameshift mutations. Arrowhead, amino acid position associated with a synonymous nucleotide difference between *Berlin* and *R6* assemblies in *Sdic2* only; this is the only noticeable difference within the coding fraction of the individual *Sdic* copies between the two assemblies.

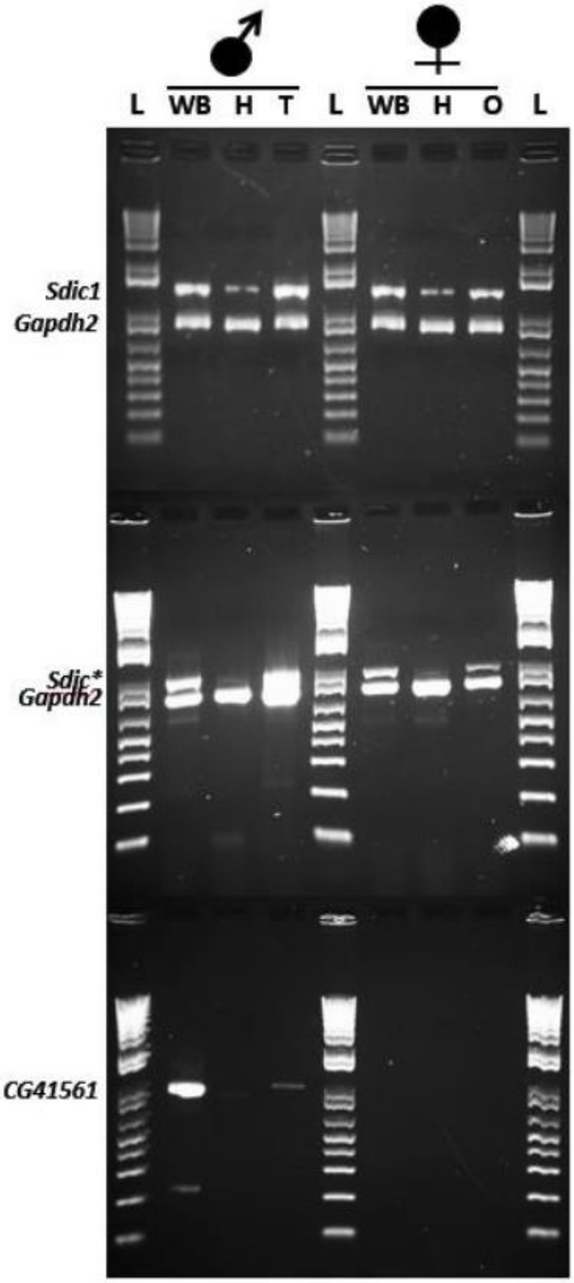


```
sw       CGCTAAAGgtacccatttcaagttgccagttctctgtcgtctccggaattaattcaaatatgcccactctttcagGACACAAAGCC
Sdic2    CGCTAAAGgtacccatttcaagttgccagttctctgtcgtctccggaattaattcaaatatgcccactctttcagGACACAAAGCC
SdicB    CGCTAAAGgtacccatttcaagttgccagttctctgtcgtctccggaattaattcaaatatgcccactctttcagGACACAAAGCC
Sdic3    CGCTAAAGgtacccatttcaagttgccagttctctgtcgtctccggaattaattcaaatatgcccactctttcagGACACAAAGCC
SdicC    CGCTAAAGCTACCCATTTCAAGTTGCCAGTTCTCTGTCGTCTCCGGAATTAATTCAAATA----------------------AGCC
Sdic4    CGCTAAAGCTACCCATTTCAAGTTGCCAGTTCTCTgtcgtctccggaattaattcaaata----------------------agcc
Sdic1    CGCTAAAGgtacccatatcaagttgccagtttctgtcgtctccggaattaattcaaatatgcccactctttcagGACACCAAGCC

sw       GCTGTACTCCTTTGAGGACAACTCCGACTACGTGATGGACGTCGCCTGGTCGCCCGTGCATCCCGCACTCTTCGCCGCCGTCGACG
Sdic2    GCTGTACTCCTTTGAGGACAACTCCGACTACGTGATGGACGTCGCCTGGTCGCCCGTGCATCCCGCACTCTTCGCCGCCGTCGACG
SdicB    GCTGTACTCCTTTGAGGACAACTCCGACTACGTGATGGACGTCGCCTGGTCGCCCGTGCATCCCGCACTCTTCGCCGCCGTCGACG
Sdic3    GCTGTACTCCTTTGAGGACAACTCCGACTACGTGATGGACGTCGCCTGGTCGCCCGTGCATCCCGCACTCTTCGCCGCCGTCGACG
SdicC    GCTGTACTCCTTTGAGGACAACTCCGACTACATGATGGACGTCGCCTGGTCGCCCGTGCATCCCGCACTCTTCGCCGCCGTCGACG
Sdic4    gctgtactcctttgaggacaactccgactacatgatggacgtcgcctggtcgcccgtgcatcccgcactcttcgccgccgtcgacg
Sdic1    GCTGTACTCCTTTGAGCA------------------GTACATCGCCTGGTCGCCCGTG-----------------------CGACG

sw       GCAGCGGCCGCCTGGATCTGTGGAACCTCAACCAAGACACGGAGGTGCCGACCGCCTCGATTGTCGTGGCGGGAGCACCAGCCCTT
Sdic2    GCAGCGGCCGCCTGGATCTGTGGAACCTCAACCAAGACACGGAGGTGCCGACCGCCTCGATTGTCGTGGCGGGAGCACCAGCCCTT
SdicB    GCAGCGGTCGCCTGGACCTGTGGAACCTCAACCAAGACACGGAGGTGCCGATCGCCTCGATTGTCGTGGCGGGAGCACCAGCCCTT
Sdic3    GCAGCGGTCGCCTGGACCTGTGGAACCTCAACCAAGACACGGAGGTGCCGATCGCCTCGATTGTCGTGGCGGGAGCACCAGCCCTT
SdicC    GCAGCGGTCGCCTGGACCTGTGGAACCTCAACCAAGACACGGAGGTGCCGACCGCCTCGATTGTCGTGGCGGGAGCACCAGCCCTT
Sdic4    gcagcggccgcctggacctgtggaacctcaaccaagacacggaggtgccgaccgcctcgattgtcgtggcgggagCACCAGCCCTT
Sdic1    GCAGCGGCCGCCTGGACCTGATAAAACTCAACCCAGACACGGAG-------------------------------CACCAGCCCTT
```

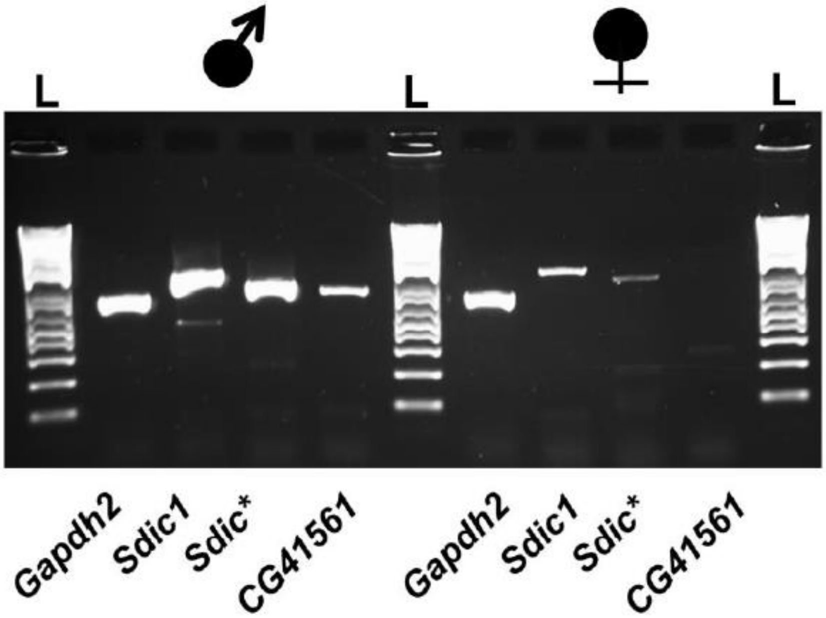
Supplementary Figure S1.7. Mutations at splice donor and alternative acceptor sites have altered the coding fractions of *SdicC, Sdic4*, and *Sdic1*. For *Sdic4*, the splice donor site was lost, causing exon 4 to extend until a splice donor site 27 nt downstream (orange box), shifting the reading frame. For *SdicC*, the splice donor site was also lost by mutation, along with the acceptor site being lost to a deletion, resulting in the whole intron being converted into coding sequence. For *Sdic1*, there are alternative splice acceptor sites that differentiate the transcripts *Sdic1-RA* and *Sdic1-RC* by the presence of 100 nt of coding region, which yields proteins with different reading frames. Sdic1-PA stems from three deletions in exon 5 (18, 23, and 31 nt) that resulted in a frameshift for 15 amino acids and a loss of 24 amino acids. Together, these events resulted in an expansion of the intron between exons 4 and 5 to 167 and 190 nt in Sdic1-RC and Sdic4-RE, respectively. Uppercase, coding; lowercase, intron; underlined, *Sdic1* variable region of exon 5; blue and green, splice donor and acceptor sites, respectively.



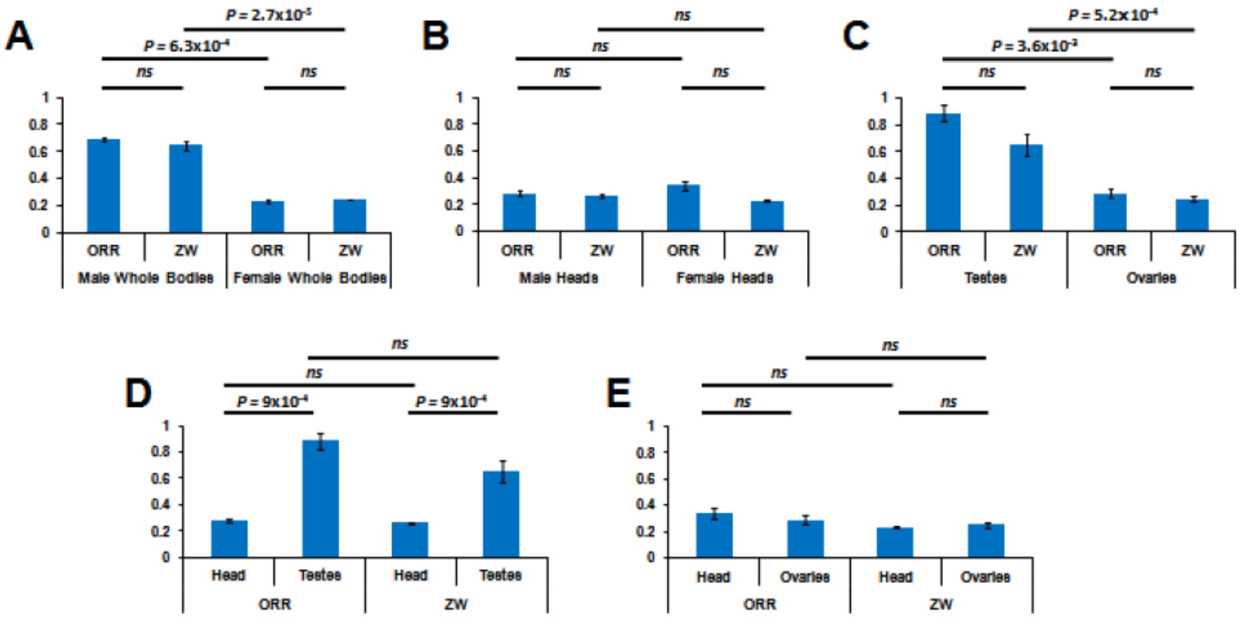
Supplementary Figure S1.8. Sliding window plot showing the patterns of nucleotide variability along the multiple sequence alignment (MSA) of the *Sdic* copies. The red and green lines represent the nucleotide divergence among the *Sdic* copies, and between *Sdic* and the *sw-AnxB10* composite, respectively. In the 5' end, all *Sdic* copies are inferred to experience gene conversion, reducing the between-copy-variation (red) relative to *sw-AnxB10* (green). Conversely, in the 3' end, *sw-AnxB10* exchanges DNA with all *Sdic* copies except *Sdic1*, which undergoes adaptive diversification (fig. 1.2A and supplementary table S1.7). This reduces the divergence to *sw-AnxB10*, to the same level that exists among the *Sdic* copies. The arrow points to the end of the fourth WD40 domain (see supplementary fig. S1.6), where a shift in nucleotide variability patterns is observed. Dark blue boxes denote stretches that are protein-coding in all *Sdic* copies, while light blue boxes indicate regions present in alternative transcript isoforms. Likewise, dark and light orange boxes indicate the fractions of the 5' and 3' UTRs shared by all (*i.e.* the constitutive fraction) or a subset of copies, respectively. Solid dots indicate the location of non-synonymous mutations.



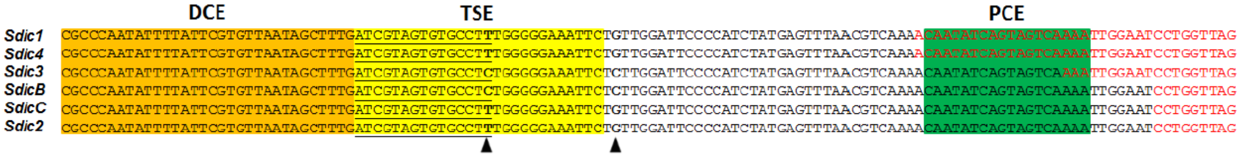
Supplementary Figure S1.9. Expression profile of *Sdic* in Oregon-R. Both an amplicon specific to *Sdic1* transcripts and an amplicon shared by *Sdic4* and *SdicC* transcripts (*Sdic**) revealed expression in ovaries and both male and female heads. This demonstrates that these *Sdic* copies are neither male specific nor limited to the reproductive system. DNA corresponding to each *Sdic* amplicon was cloned and Sanger sequenced from RT-PCR reactions using heads (but for *Sdic** in females), ovaries, and testes, corroborating the presumed identities. In the case of *Sdic1*, the seven clones sequenced all confirmed the existence of the *Sdic1-RA* transcript; none validated the existence of *Sdic1-RC*. *Gapdh2* was used as a control for the integrity of the cDNA template. *CG41561*, a *Y*-linked gene, was used to control for contamination in the female samples. According to FlyBase (dos Santos, et al. 2015), *CG41561* is expressed only in males, primarily in testes, but not in the head. L, 1Kb+ Ladder; WB, whole body; H, head; T, testes; and O, ovaries.



Supplementary Figure S1.10. *Sdic* expression in Zimbabwe-109 whole bodies. Female expression of *Sdic1* and *Sdic4* and/or *SdicC* (*Sdic**) was also confirmed in Zimbabwe-109, demonstrating that female expression of *Sdic* is likely common in all *D. melanogaster* strains. *Gapdh2* was used as a control for the integrity of the cDNA template. *CG41561*, a *Y*-linked gene, was used to control for contamination in the female sample. L, 1Kb+ ladder.



Supplementary Figure S1.11. Relative expression levels of *Sdic1* across samples as assayed by qRT-PCR. (A) whole bodies; (B) heads; (C) sex organs; (D) within males; (E) within females. Horizontal lines indicate the contrast evaluated. *P*-values were calculated using the moderate *t*-test as in LIMMA (Smyth 2004) and adjusted for multiple corrections (Benjamini and Hochberg 1995). *ns*, non-statistically significant differences ($P > 0.05$). Normalized Ct values were transformed (Material and Methods) such that 1 and 0 signify the highest and the lowest expression level, respectively. Error bars denote the standard error of the mean. Three biological replicates were used for each strain by sex by tissue type combination. Although *Sdic1* is expressed in the ovaries and heads of *D. melanogaster* adults, the highest expression levels are found in testes. See supplementary table S1.8 for the expression difference values from each comparison, both for *Sdic1* and *Sdic4*. ORR, Oregon-R; ZW, Zimbabwe-109.


DCE
TSE
PCE
Sdic1 CGCCCAATATTTTATTCGTGTTAATAGCTTTGATCGTAGTGTGCCTTTGGGGAAATTCTGTTGGATTCCCCATCTATGAGTTTAACGTCAAAACAATATCAGTAGTCAAAATTGGAATCCTGGTTAG
Sdic4 CGCCCAATATTTTATTCGTGTTAATAGCTTTGATCGTAGTGTGCCTTTGGGGAAATTCTGTTGGATTCCCCATCTATGAGTTTAACGTCAAAACAATATCAGTAGTCAAAATTGGAATCCTGGTTAG
Sdic3 CGCCCAATATTTTATTCGTGTTAATAGCTTTGATCGTAGTGTGCCTCTGGGGAAATTCTCTTGGATTCCCCATCTATGAGTTTAACGTCAAAACAATATCAGTAGTCAAAATTGGAATCCTGGTTAG
SdicB CGCCCAATATTTTATTCGTGTTAATAGCTTTGATCGTAGTGTGCCTCTGGGGAAATTCTCTTGGATTCCCCATCTATGAGTTTAACGTCAAAACAATATCAGTAGTCAAAATTGGAATCCTGGTTAG
SdicC CGCCCAATATTTTATTCGTGTTAATAGCTTTGATCGTAGTGTGCCTTTGGGGAAATTCTGTTGGATTCCCCATCTATGAGTTTAACGTCAAAACAATATCAGTAGTCAAAATTGGAATCCTGGTTAG
Sdic2 CGCCCAATATTTTATTCGTGTTAATAGCTTTGATCGTAGTGTGCCTTTGGGGAAATTCTGTTGGATTCCCCATCTATGAGTTTAACGTCAAAACAATATCAGTAGTCAAAATTGGAATCCTGGTTAG

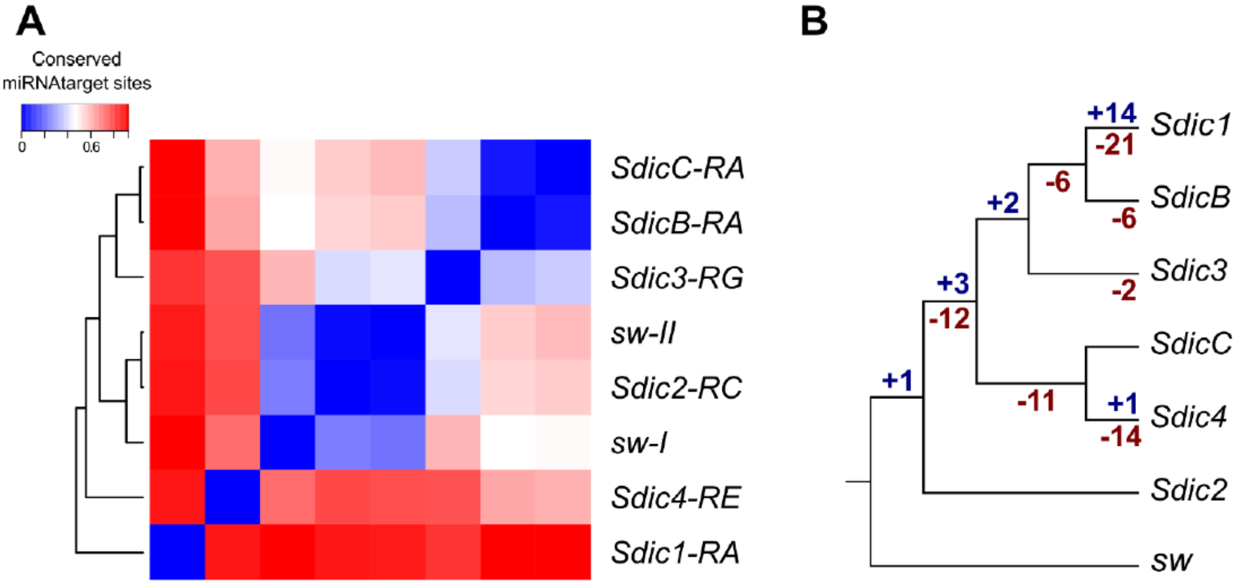
Supplementary Figure S1.12. Promoter sequence across *Sdic* copies. The *Sdic* promoter region is composed of three color-coded elements: a distal core element (DCE), a testes-like specific promoter element (TSE), and a proximal core element (PCE). Two different promoter variants exist in *Sdic*, differing by two nucleotides: one in the TSE and another in the 35bp spacer sequence. The testes-specific promoter was deemed as such due to its close sequence similarity with the previously delineated TSE motif in the *βTub85D* gene (Michiels, et al. 1989; Nurminsky, et al. 1998). This motif encompasses a TATA box and a 14 nt long stretch (5'-ATCGTAGTAGCCTA-3') that confers testis-expression specificity to *βTub85D* (Michiels, et al. 1989). Underlined, the equivalent stretch in *Sdic*, which contains one extra nucleotide and two differences in relation to the stretch in *βTub85D*. Variable sites are in bold and indicated by an arrowhead. The variable nucleotide within the TSE distinguishes *Sdic3* and *SdicB* from the rest of the copies. The putative 5'UTR (red) boundaries are delineated according to FlyBase (dos Santos, et al. 2015).

https://academic.oup.com/mbe/article/34/1/51/2452321#supplementary-data

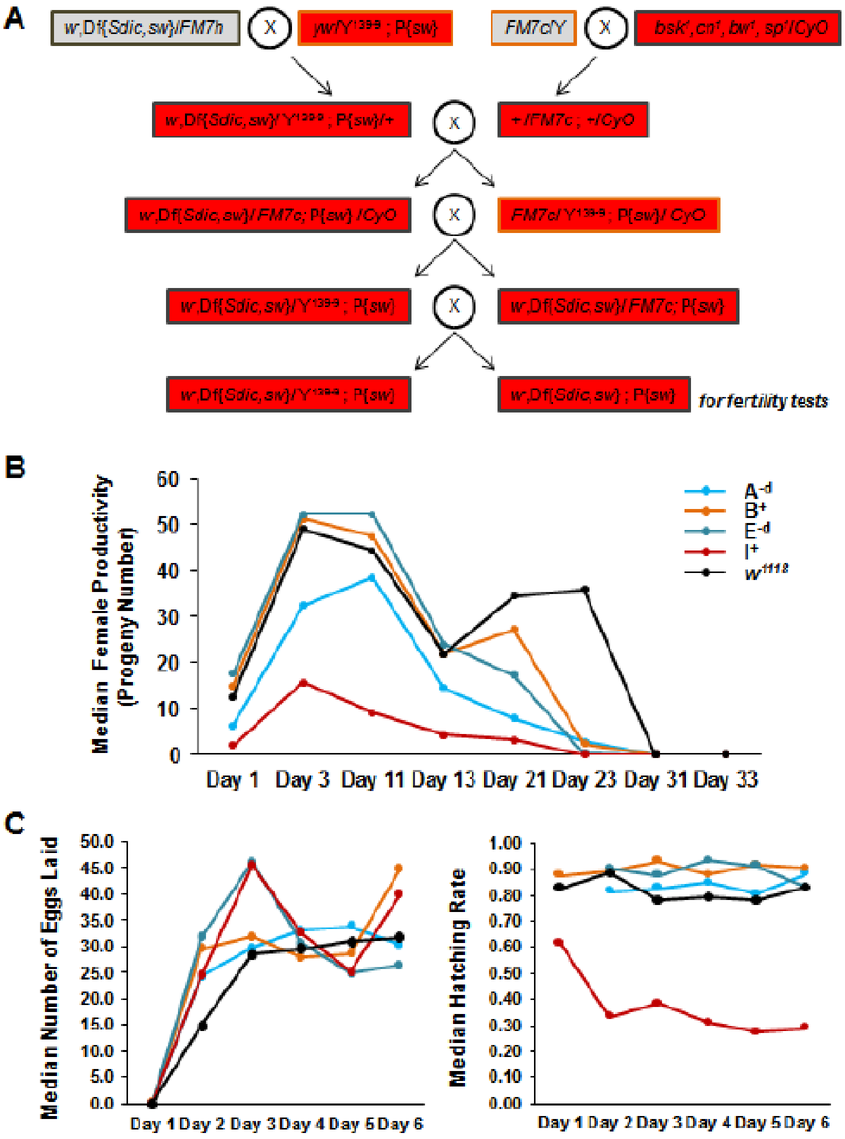
Supplementary Figure S1.13. Distribution along the *Sdic* transcripts of the diagnostic motifs and PCR priming sites used for detecting expression. Each motif used for detecting transcripts via RNA-seq data is color-coded and labeled. Each primer is boxed in blue (orange indicates overlap). Boundaries between exons are marked with an arrow and |. Exon number is shown above (e, constant exon; v, variable exon). The sequences shown are as in the *Berlin* assembly. See supplementary tables S1.10 and S1.16 for sequence motifs and primers respectively.

https://academic.oup.com/mbe/article/34/1/51/2452321#supplementary-data

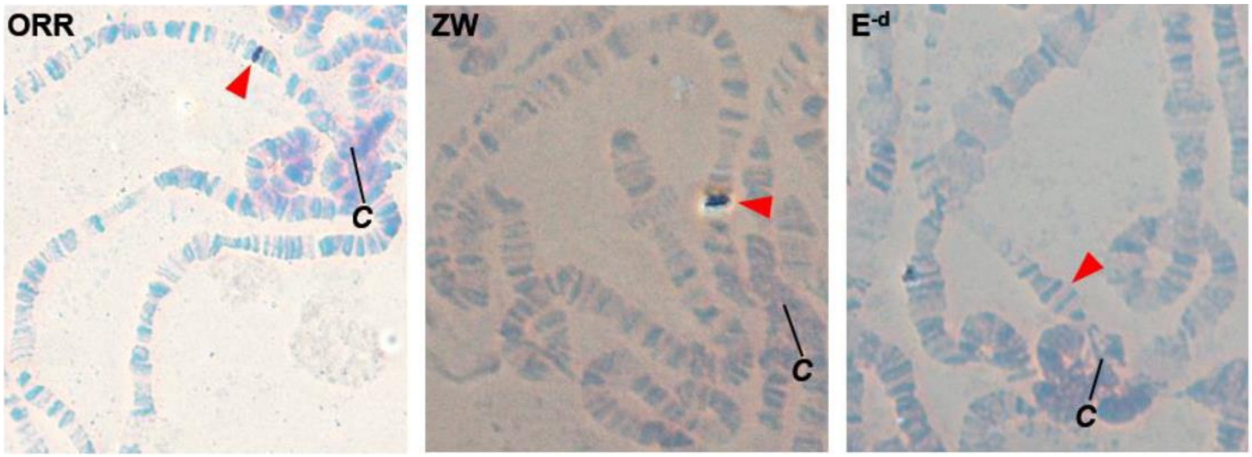
Supplementary Figure S1.14. Example of alignments between diagnostic motif corresponding to particular *Sdic* copies and RNA-seq reads. The results for four *Sdic* copies (A-D) across some of the biological conditions surveyed are shown. These conditions include: testis from four-day-old mated males; ovaries from four-day-old virgin females; central nervous system from two-day old pupae; imaginal discs from third instar larvae, wandering stage; four-day-old pupae; and five-day-old males. Diagnostic motifs appear at the top of each alignment. supplementary table S1.10 shows the precise location of the core motif that is informative of the expression of a particular copy or transcript. The library and numerical ID of reads showing perfect alignments against the diagnostic motifs are indicated on the left.



Supplementary Figure S1.15. Comparison of microRNA target sites across *Sdic* and *sw* 3'UTRs. (A) Heatmap showing pairwise similarities in microRNA target sites between transcripts. Dark blue indicates full conservation of orthologous sites. Genes were clustered with UPGMA according to binary distances between orthologous microRNA target sites. The 15 transcripts of *sw* exhibit one of two possible 3'UTRs, here denoted as *I* and *II*. (B) Pattern of gain (blue) and loss (red) of microRNA target sites across the *Sdic* phylogeny (fig. 1.2C), according to a maximum parsimony reconstruction of events.

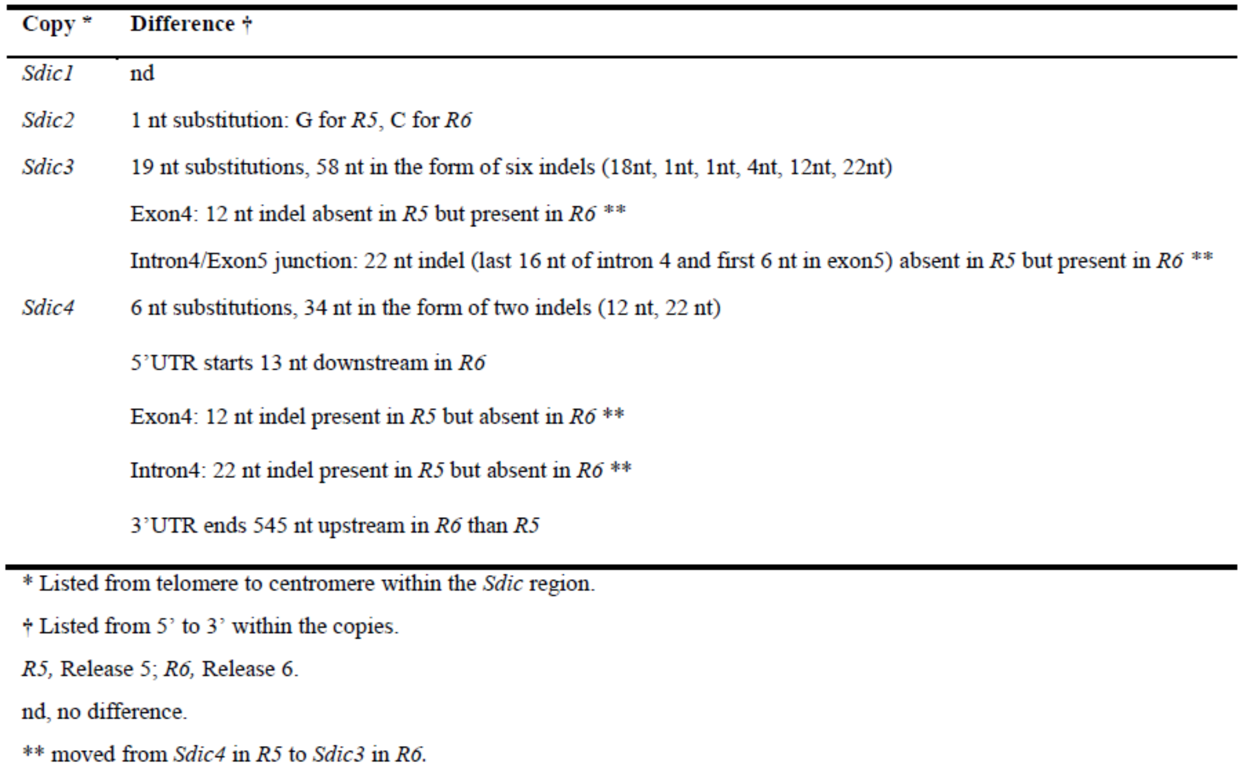


Supplementary Figure S1.16. Assessing the impact of the *Sdic* region on female fitness. (A) Cross scheme performed to generate the females used in fertility assays. These females are homozygous for the deletion of the *Sdic* region (A^{-d} and E^{-d}). Females of these strains are also homozygous for the *sw* transgene, making the number of *sw* copies identical to females from B^{+}, I^{+}, and w^{1118} strains, which were used as controls in downstream analyses. The original strains A^{-} and E^{-} were generated simultaneously through the same experimental procedure as the B^{+} and I^{+} strains, respectively, with the difference being that the deletion occurred in the former two but not in the latter two strains (Yeh, et al. 2012). (B) Productivity of females with and without the *Sdic* region (supplementary table S1.12). Only data from days 1, 3, 11, 13, 21, and 23 were considered in downstream analyses since no individuals emerged on days 31 and 33. (C) Comparison of two additional parameters of female fitness among females with and without the *Sdic* region. Left, number of eggs laid per female (supplementary table S1.13). Right, egg hatching rate (supplementary table S1.14).



Supplementary Figure S1.17. Physical mapping of the *Sdic* region in relevant strains. *In situ* hybridization of a *Sdic* specific probe to the polytene chromosomes of strains with (Oregon-R, Zimbabwe-109) and without (E-d) the *Sdic* region at *X*(19C1). Red arrowhead, *X*(19C1). Unlike wildtype strains, the engineered strain E- shows no evidence of a hybridization signal, corroborating that the profound changes in the annotation of the *Sdic* region did not affect the precise location of the *P* elements used in engineering the deletion of the region (Yeh, et al. 2012). Therefore, the elimination of all existing copies of *Sdic* at *X*(19C1) in previously engineered strains make them suitable for the phenotypic tests performed here and previously (Yeh, et al. 2012). *C*, centromere. ORR, Oregon-R; ZW, Zimbabwe-109.

Supplementary Table S1.1. Sequence differences for the same *Sdic* copy between the *R5* and *R6* assemblies



Copy *	Difference †
Sdic1	nd
Sdic2	1 nt substitution: G for *R5*, C for *R6*
Sdic3	19 nt substitutions, 58 nt in the form of six indels (18nt, 1nt, 1nt, 4nt, 12nt, 22nt)
	Exon4: 12 nt indel absent in *R5* but present in *R6* **
	Intron4/Exon5 junction: 22 nt indel (last 16 nt of intron 4 and first 6 nt in exon5) absent in *R5* but present in *R6* **
Sdic4	6 nt substitutions, 34 nt in the form of two indels (12 nt, 22 nt)
	5'UTR starts 13 nt downstream in *R6*
	Exon4: 12 nt indel present in *R5* but absent in *R6* **
	Intron4: 22 nt indel present in *R5* but absent in *R6* **
	3'UTR ends 545 nt upstream in *R6* than *R5*

* Listed from telomere to centromere within the *Sdic* region.

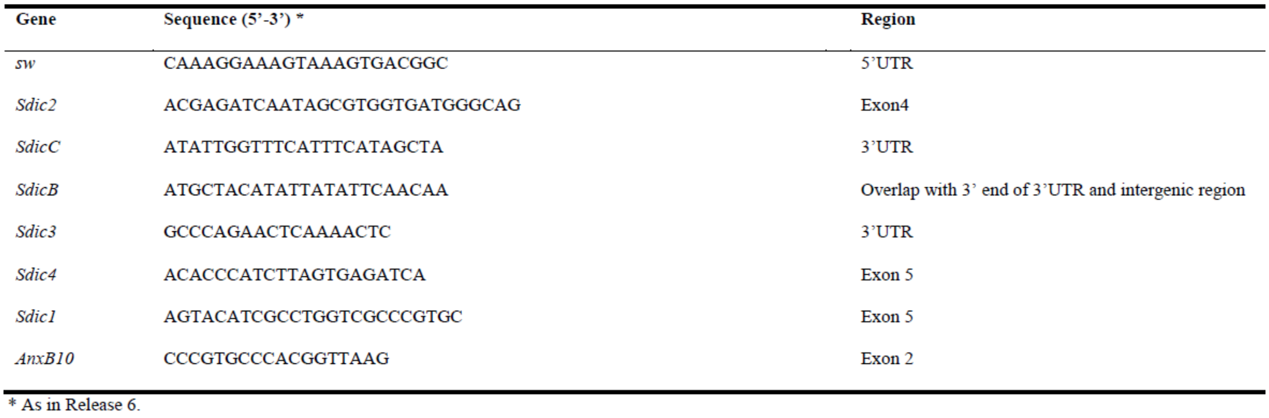
† Listed from 5' to 3' within the copies.

R5, Release 5; *R6,* Release 6.

nd, no difference.

** moved from *Sdic4* in *R5* to *Sdic3* in *R6*.

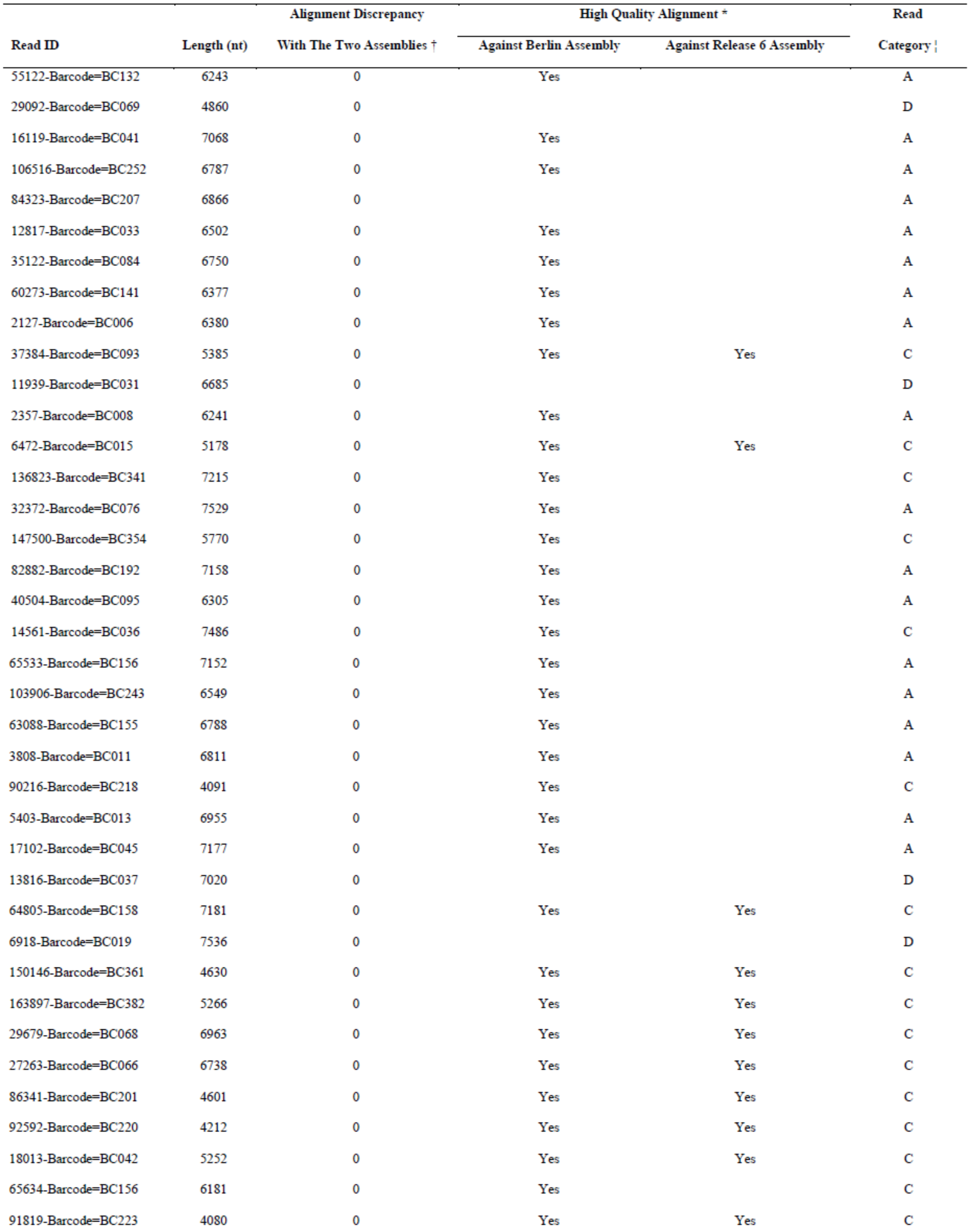
Supplementary Table S1.2. Diagnostic motifs used to detect gene locations in the *Sdic* region across assemblies



Gene	Sequence (5’-3’) *	Region
sw	CAAAGGAAAGTAAAGTGACGGC	5’UTR
Sdic2	ACGAGATCAATAGCGTGGTGATGGGCAG	Exon4
SdicC	ATATTGGTTTCATTTCATAGCTA	3’UTR
SdicB	ATGCTACATATTATATTCAACAA	Overlap with 3’ end of 3’UTR and intergenic region
Sdic3	GCCCAGAACTCAAAACTC	3’UTR
Sdic4	ACACCCATCTTAGTGAGATCA	Exon 5
Sdic1	AGTACATCGCCTGGTCGCCCGTGC	Exon 5
AnxB10	CCCGTGCCCACGGTTAAG	Exon 2

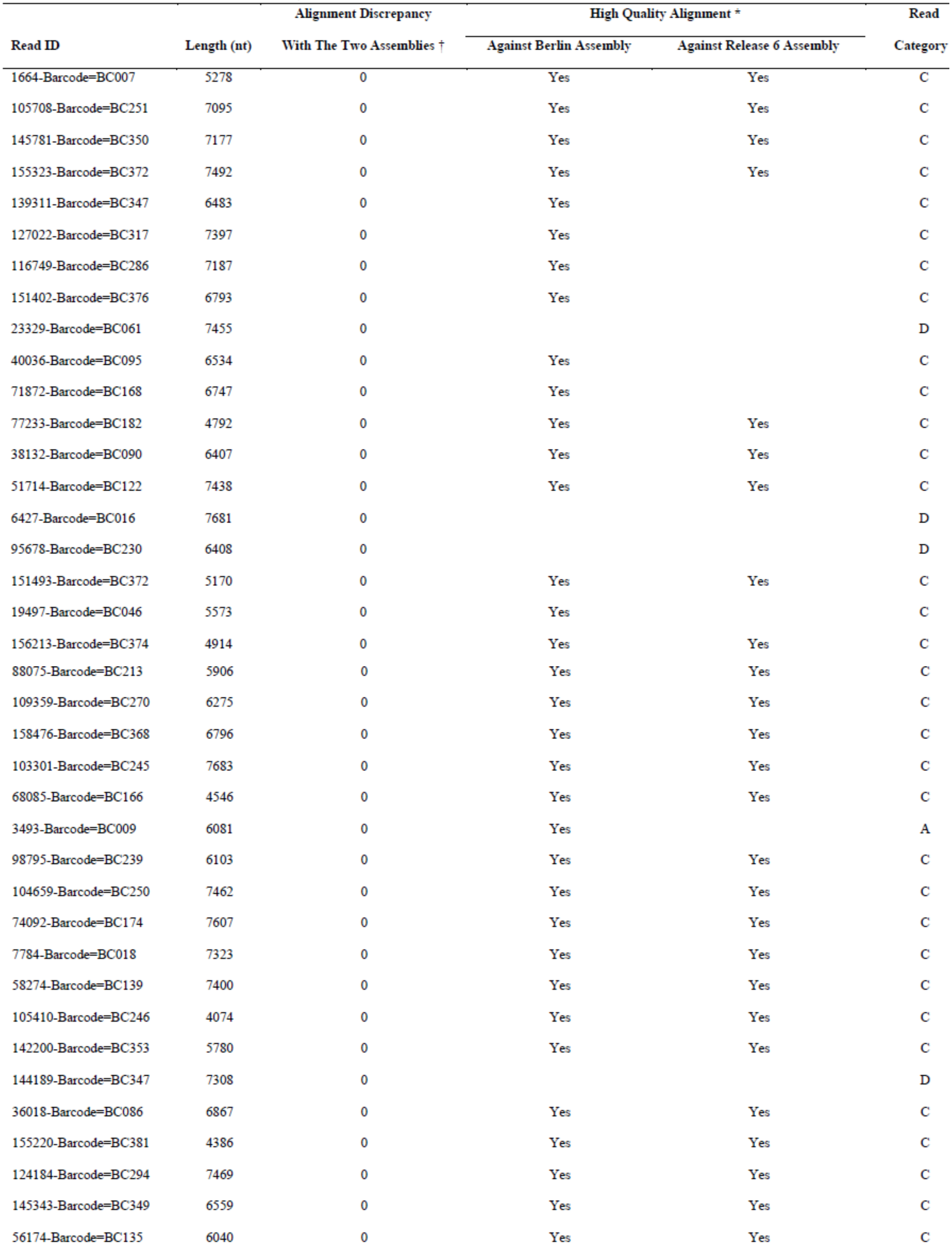
* As in Release 6.

Supplementary Table S1.3. Subset of SLRs used to evaluate the accuracy of the reconstruction of the *Sdic* region between two assemblies



		Alignment Discrepancy	High Quality Alignment *		Read
Read ID	Length (nt)	With The Two Assemblies †	Against Berlin Assembly	Against Release 6 Assembly	Category ¦
55122-Barcode=BC132	6243	0	Yes		A
29092-Barcode=BC069	4860	0			D
16119-Barcode=BC041	7068	0	Yes		A
106516-Barcode=BC252	6787	0	Yes		A
84323-Barcode=BC207	6866	0			A
12817-Barcode=BC033	6502	0	Yes		A
35122-Barcode=BC084	6750	0	Yes		A
60273-Barcode=BC141	6377	0	Yes		A
2127-Barcode=BC006	6380	0	Yes		A
37384-Barcode=BC093	5385	0	Yes	Yes	C
11939-Barcode=BC031	6685	0			D
2357-Barcode=BC008	6241	0	Yes		A
6472-Barcode=BC015	5178	0	Yes	Yes	C
136823-Barcode=BC341	7215	0	Yes		C
32372-Barcode=BC076	7529	0	Yes		A
147500-Barcode=BC354	5770	0	Yes		C
82882-Barcode=BC192	7158	0	Yes		A
40504-Barcode=BC095	6305	0	Yes		A
14561-Barcode=BC036	7486	0	Yes		C
65533-Barcode=BC156	7152	0	Yes		A
103906-Barcode=BC243	6549	0	Yes		A
63088-Barcode=BC155	6788	0	Yes		A
3808-Barcode=BC011	6811	0	Yes		A
90216-Barcode=BC218	4091	0	Yes		C
5403-Barcode=BC013	6955	0	Yes		A
17102-Barcode=BC045	7177	0	Yes		A
13816-Barcode=BC037	7020	0			D
64805-Barcode=BC158	7181	0	Yes	Yes	C
6918-Barcode=BC019	7536	0			D
150146-Barcode=BC361	4630	0	Yes	Yes	C
163897-Barcode=BC382	5266	0	Yes	Yes	C
29679-Barcode=BC068	6963	0	Yes	Yes	C
27263-Barcode=BC066	6738	0	Yes	Yes	C
86341-Barcode=BC201	4601	0	Yes	Yes	C
92592-Barcode=BC220	4212	0	Yes	Yes	C
18013-Barcode=BC042	5252	0	Yes	Yes	C
65634-Barcode=BC156	6181	0	Yes		C
91819-Barcode=BC223	4080	0	Yes	Yes	C

Supplementary Table S1.3. Subset of SLRs used to evaluate the accuracy of the reconstruction of the *Sdic* region between two assemblies

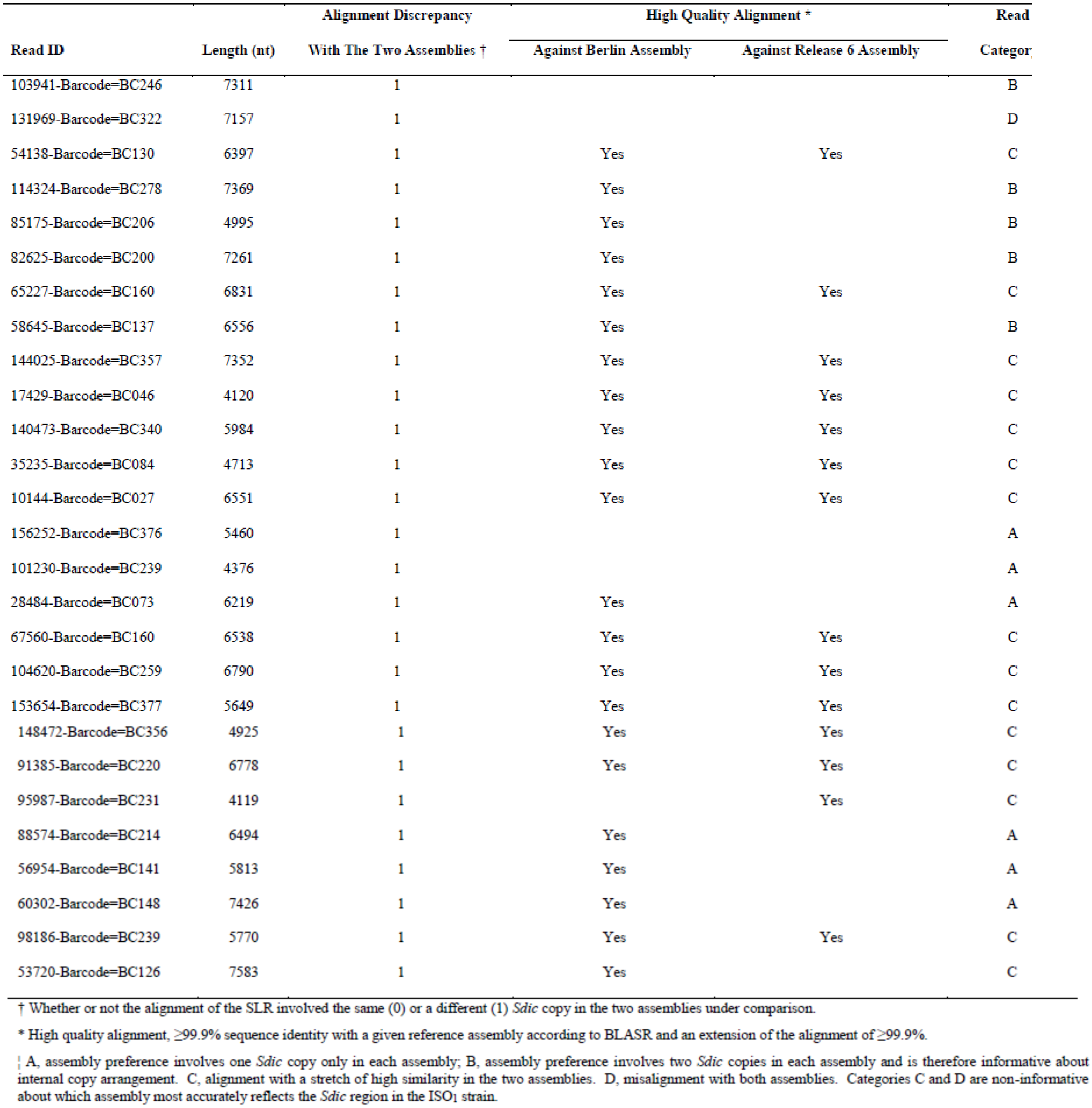


		Alignment Discrepancy	High Quality Alignment *		Read
Read ID	Length (nt)	With The Two Assemblies †	Against Berlin Assembly	Against Release 6 Assembly	Category
1664-Barcode=BC007	5278	0	Yes	Yes	C
105708-Barcode=BC251	7095	0	Yes	Yes	C
145781-Barcode=BC350	7177	0	Yes	Yes	C
155323-Barcode=BC372	7492	0	Yes	Yes	C
139311-Barcode=BC347	6483	0	Yes		C
127022-Barcode=BC317	7397	0	Yes		C
116749-Barcode=BC286	7187	0	Yes		C
151402-Barcode=BC376	6793	0	Yes		C
23329-Barcode=BC061	7455	0			D
40036-Barcode=BC095	6534	0	Yes		C
71872-Barcode=BC168	6747	0	Yes		C
77233-Barcode=BC182	4792	0	Yes	Yes	C
38132-Barcode=BC090	6407	0	Yes	Yes	C
51714-Barcode=BC122	7438	0	Yes	Yes	C
6427-Barcode=BC016	7681	0			D
95678-Barcode=BC230	6408	0			D
151493-Barcode=BC372	5170	0	Yes	Yes	C
19497-Barcode=BC046	5573	0	Yes		C
156213-Barcode=BC374	4914	0	Yes	Yes	C
88075-Barcode=BC213	5906	0	Yes	Yes	C
109359-Barcode=BC270	6275	0	Yes	Yes	C
158476-Barcode=BC368	6796	0	Yes	Yes	C
103301-Barcode=BC245	7683	0	Yes	Yes	C
68085-Barcode=BC166	4546	0	Yes	Yes	C
3493-Barcode=BC009	6081	0	Yes		A
98795-Barcode=BC239	6103	0	Yes	Yes	C
104659-Barcode=BC250	7462	0	Yes	Yes	C
74092-Barcode=BC174	7607	0	Yes	Yes	C
7784-Barcode=BC018	7323	0	Yes	Yes	C
58274-Barcode=BC139	7400	0	Yes	Yes	C
105410-Barcode=BC246	4074	0	Yes	Yes	C
142200-Barcode=BC353	5780	0	Yes	Yes	C
144189-Barcode=BC347	7308	0			D
36018-Barcode=BC086	6867	0	Yes	Yes	C
155220-Barcode=BC381	4386	0	Yes	Yes	C
124184-Barcode=BC294	7469	0	Yes	Yes	C
145343-Barcode=BC349	6559	0	Yes	Yes	C
56174-Barcode=BC135	6040	0	Yes	Yes	C

Supplementary Table S1.3. Subset of SLRs used to evaluate the accuracy of the reconstruction of the *Sdic* region between two assemblies

		Alignment Discrepancy	High Quality Alignment *		Read
Read ID	Length (nt)	With The Two Assemblies †	Against Berlin Assembly	Against Release 6 Assembly	Category
13107-Barcode=BC034	6442	0	Yes	Yes	C
119101-Barcode=BC297	7014	0	Yes	Yes	C
70108-Barcode=BC167	7599	0	Yes	Yes	C
11386-Barcode=BC028	4937	1	Yes	Yes	C
57446-Barcode=BC138	5557	1	Yes		B
151577-Barcode=BC364	6676	1	Yes		B
73312-Barcode=BC173	4914	1		Yes	B
52133-Barcode=BC125	7499	1	Yes		B
131712-Barcode=BC326	6866	1	Yes		C
133393-Barcode=BC314	6878	1	Yes		C
10606-Barcode=BC025	5440	1	Yes		B
120275-Barcode=BC285	7205	1			D
117586-Barcode=BC292	6347	1	Yes		C
28137-Barcode=BC067	7201	1			D
133572-Barcode=BC333	7218	1	Yes	Yes	C
58008-Barcode=BC136	7651	1	Yes		C
2837-Barcode=BC008	7096	1	Yes		C
94918-Barcode=BC229	6282	1	Yes	Yes	C
10365-Barcode=BC026	6755	1	Yes	Yes	C
1664-Barcode=BC007	5278	0	Yes	Yes	C
105708-Barcode=BC251	7095	0	Yes	Yes	C
145781-Barcode=BC350	7177	0	Yes	Yes	C
155323-Barcode=BC372	7492	0	Yes	Yes	C
139311-Barcode=BC347	6483	0	Yes		C
127022-Barcode=BC317	7397	0	Yes		C
116749-Barcode=BC286	7187	0	Yes		C
151402-Barcode=BC376	6793	0	Yes		C
23329-Barcode=BC061	7455	0			D
40036-Barcode=BC095	6534	0	Yes		C
71872-Barcode=BC168	6747	0	Yes		C
77233-Barcode=BC182	4792	0	Yes	Yes	C
38132-Barcode=BC090	6407	0	Yes	Yes	C
51714-Barcode=BC122	7438	0	Yes	Yes	C
6427-Barcode=BC016	7681	0			D
95678-Barcode=BC230	6408	0			D
151493-Barcode=BC372	5170	0	Yes	Yes	C
19497-Barcode=BC046	5573	0	Yes		C
156213-Barcode=BC374	4914	0	Yes	Yes	C

Supplementary Table S1.3. Subset of SLRs used to evaluate the accuracy of the reconstruction of the *Sdic* region between two assemblies



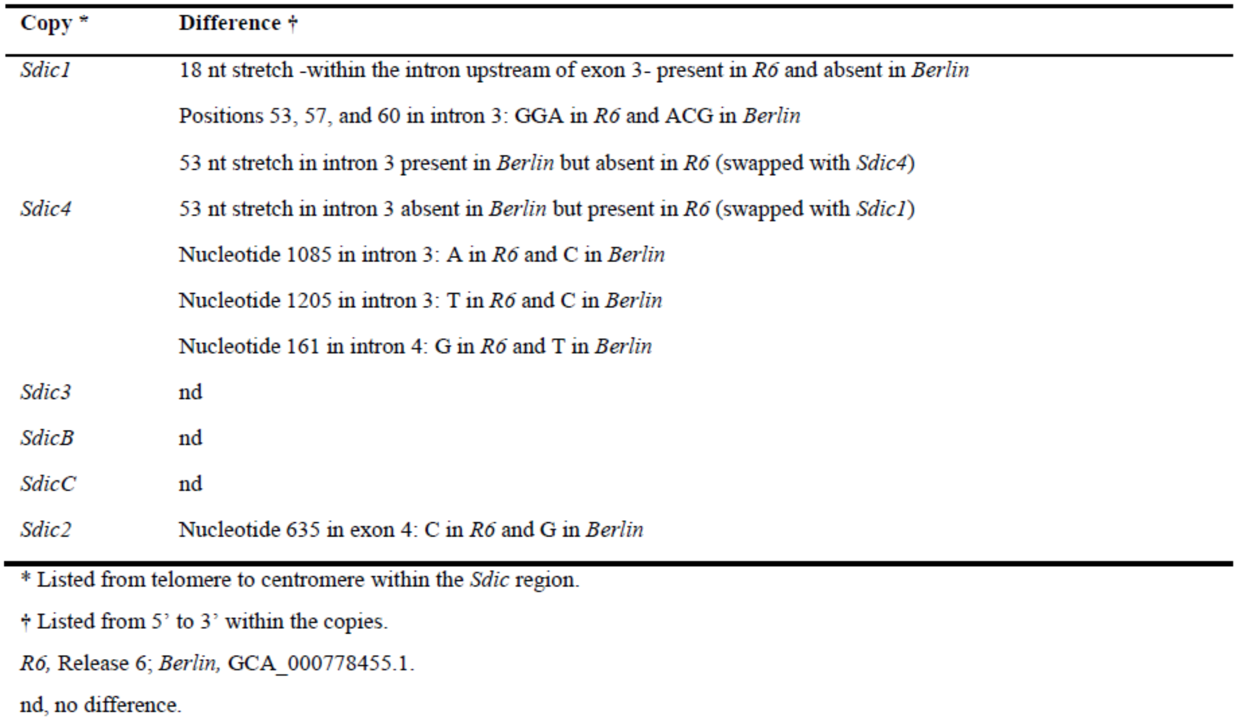
		Alignment Discrepancy	High Quality Alignment *		Read
Read ID	Length (nt)	With The Two Assemblies †	Against Berlin Assembly	Against Release 6 Assembly	Categor
103941-Barcode=BC246	7311	1			B
131969-Barcode=BC322	7157	1			D
54138-Barcode=BC130	6397	1	Yes	Yes	C
114324-Barcode=BC278	7369	1	Yes		B
85175-Barcode=BC206	4995	1	Yes		B
82625-Barcode=BC200	7261	1	Yes		B
65227-Barcode=BC160	6831	1	Yes	Yes	C
58645-Barcode=BC137	6556	1	Yes		B
144025-Barcode=BC357	7352	1	Yes	Yes	C
17429-Barcode=BC046	4120	1	Yes	Yes	C
140473-Barcode=BC340	5984	1	Yes	Yes	C
35235-Barcode=BC084	4713	1	Yes	Yes	C
10144-Barcode=BC027	6551	1	Yes	Yes	C
156252-Barcode=BC376	5460	1			A
101230-Barcode=BC239	4376	1			A
28484-Barcode=BC073	6219	1	Yes		A
67560-Barcode=BC160	6538	1	Yes	Yes	C
104620-Barcode=BC259	6790	1	Yes	Yes	C
153654-Barcode=BC377	5649	1	Yes	Yes	C
148472-Barcode=BC356	4925	1	Yes	Yes	C
91385-Barcode=BC220	6778	1	Yes	Yes	C
95987-Barcode=BC231	4119	1		Yes	C
88574-Barcode=BC214	6494	1	Yes		A
56954-Barcode=BC141	5813	1	Yes		A
60302-Barcode=BC148	7426	1	Yes		A
98186-Barcode=BC239	5770	1	Yes	Yes	C
53720-Barcode=BC126	7583	1	Yes		C

† Whether or not the alignment of the SLR involved the same (0) or a different (1) *Sdic* copy in the two assemblies under comparison.

* High quality alignment, ≥99.9% sequence identity with a given reference assembly according to BLASR and an extension of the alignment of ≥99.9%.

¦ A, assembly preference involves one *Sdic* copy only in each assembly; B, assembly preference involves two *Sdic* copies in each assembly and is therefore informative about internal copy arrangement. C, alignment with a stretch of high similarity in the two assemblies. D, misalignment with both assemblies. Categories C and D are non-informative about which assembly most accurately reflects the *Sdic* region in the ISO$_1$ strain.

Supplementary Table S1.4. Sequence differences for the same *Sdic* copy between the *R6* and the *Berlin* assemblies



Copy *	Difference †
Sdic1	18 nt stretch -within the intron upstream of exon 3- present in *R6* and absent in *Berlin*
	Positions 53, 57, and 60 in intron 3: GGA in *R6* and ACG in *Berlin*
	53 nt stretch in intron 3 present in *Berlin* but absent in *R6* (swapped with *Sdic4*)
Sdic4	53 nt stretch in intron 3 absent in *Berlin* but present in *R6* (swapped with *Sdic1*)
	Nucleotide 1085 in intron 3: A in *R6* and C in *Berlin*
	Nucleotide 1205 in intron 3: T in *R6* and C in *Berlin*
	Nucleotide 161 in intron 4: G in *R6* and T in *Berlin*
Sdic3	nd
SdicB	nd
SdicC	nd
Sdic2	Nucleotide 635 in exon 4: C in *R6* and G in *Berlin*

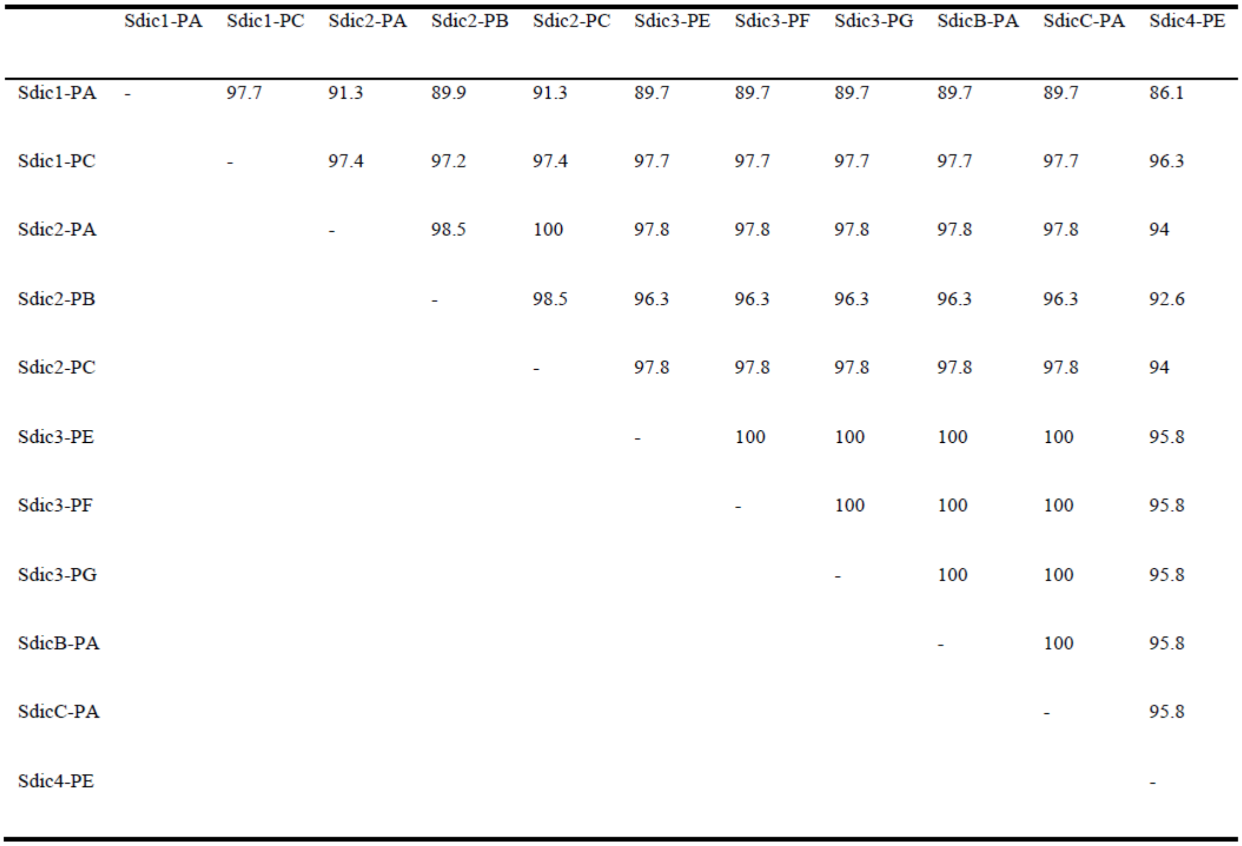
* Listed from telomere to centromere within the *Sdic* region.

† Listed from 5' to 3' within the copies.

R6, Release 6; *Berlin*, GCA_000778455.1.

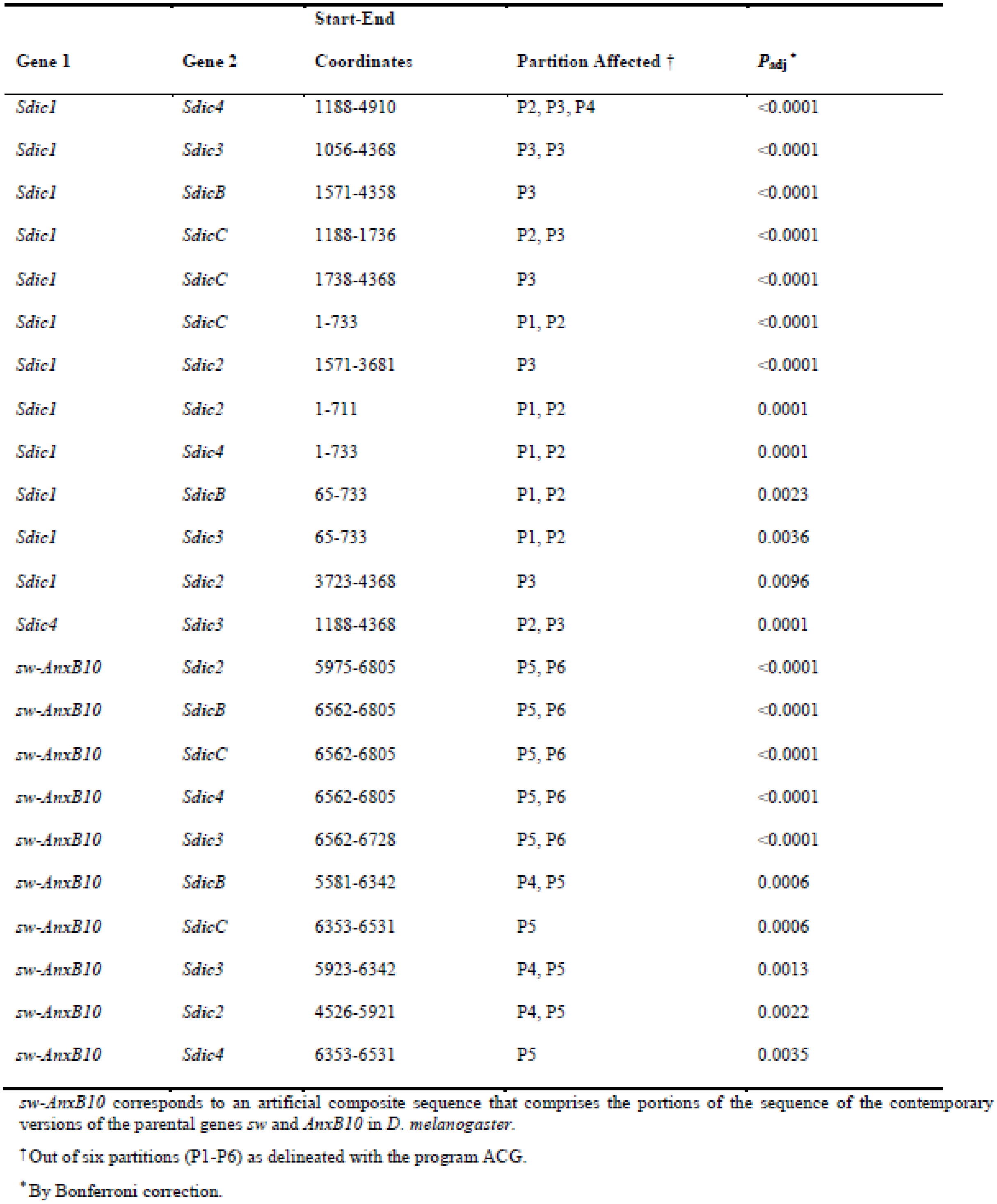
nd, no difference.

Supplementary Table S1.5. Percent similarity at the protein level between different pairs of Sdic variants



	Sdic1-PA	Sdic1-PC	Sdic2-PA	Sdic2-PB	Sdic2-PC	Sdic3-PE	Sdic3-PF	Sdic3-PG	SdicB-PA	SdicC-PA	Sdic4-PE
Sdic1-PA	-	97.7	91.3	89.9	91.3	89.7	89.7	89.7	89.7	89.7	86.1
Sdic1-PC		-	97.4	97.2	97.4	97.7	97.7	97.7	97.7	97.7	96.3
Sdic2-PA			-	98.5	100	97.8	97.8	97.8	97.8	97.8	94
Sdic2-PB				-	98.5	96.3	96.3	96.3	96.3	96.3	92.6
Sdic2-PC					-	97.8	97.8	97.8	97.8	97.8	94
Sdic3-PE						-	100	100	100	100	95.8
Sdic3-PF							-	100	100	100	95.8
Sdic3-PG								-	100	100	95.8
SdicB-PA									-	100	95.8
SdicC-PA										-	95.8
Sdic4-PE											-

Supplementary Table S1.6. Gene conversion events between gene pairs predicted by GENECONV



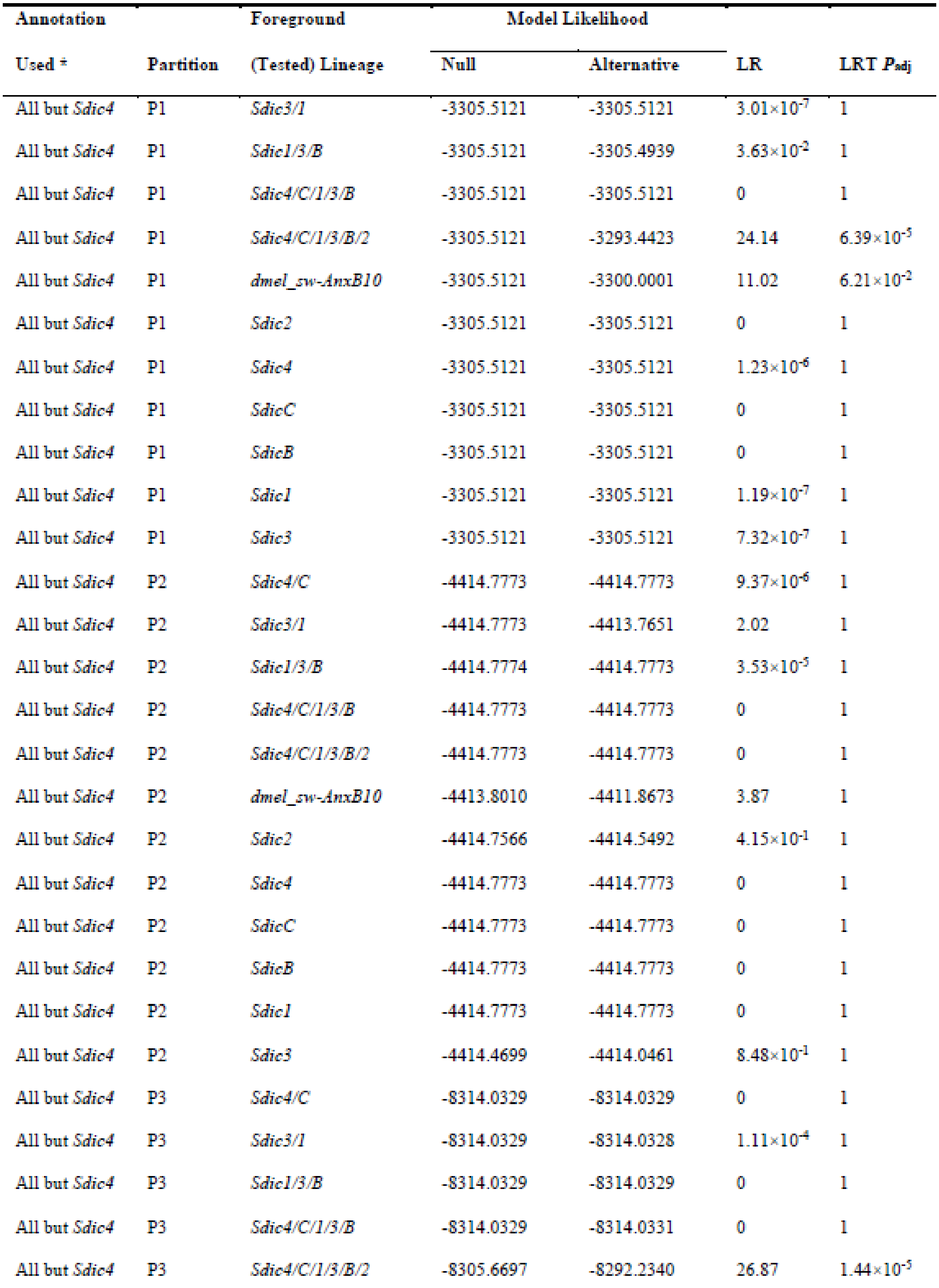
Gene 1	Gene 2	Start-End Coordinates	Partition Affected †	P_{adj} *
Sdic1	*Sdic4*	1188-4910	P2, P3, P4	<0.0001
Sdic1	*Sdic3*	1056-4368	P3, P3	<0.0001
Sdic1	*SdicB*	1571-4358	P3	<0.0001
Sdic1	*SdicC*	1188-1736	P2, P3	<0.0001
Sdic1	*SdicC*	1738-4368	P3	<0.0001
Sdic1	*SdicC*	1-733	P1, P2	<0.0001
Sdic1	*Sdic2*	1571-3681	P3	<0.0001
Sdic1	*Sdic2*	1-711	P1, P2	0.0001
Sdic1	*Sdic4*	1-733	P1, P2	0.0001
Sdic1	*SdicB*	65-733	P1, P2	0.0023
Sdic1	*Sdic3*	65-733	P1, P2	0.0036
Sdic1	*Sdic2*	3723-4368	P3	0.0096
Sdic4	*Sdic3*	1188-4368	P2, P3	0.0001
sw-AnxB10	*Sdic2*	5975-6805	P5, P6	<0.0001
sw-AnxB10	*SdicB*	6562-6805	P5, P6	<0.0001
sw-AnxB10	*SdicC*	6562-6805	P5, P6	<0.0001
sw-AnxB10	*Sdic4*	6562-6805	P5, P6	<0.0001
sw-AnxB10	*Sdic3*	6562-6728	P5, P6	<0.0001
sw-AnxB10	*SdicB*	5581-6342	P4, P5	0.0006
sw-AnxB10	*SdicC*	6353-6531	P5	0.0006
sw-AnxB10	*Sdic3*	5923-6342	P4, P5	0.0013
sw-AnxB10	*Sdic2*	4526-5921	P4, P5	0.0022
sw-AnxB10	*Sdic4*	6353-6531	P5	0.0035

sw-AnxB10 corresponds to an artificial composite sequence that comprises the portions of the sequence of the contemporary versions of the parental genes *sw* and *AnxB10* in *D. melanogaster*.

† Out of six partitions (P1-P6) as delineated with the program ACG.

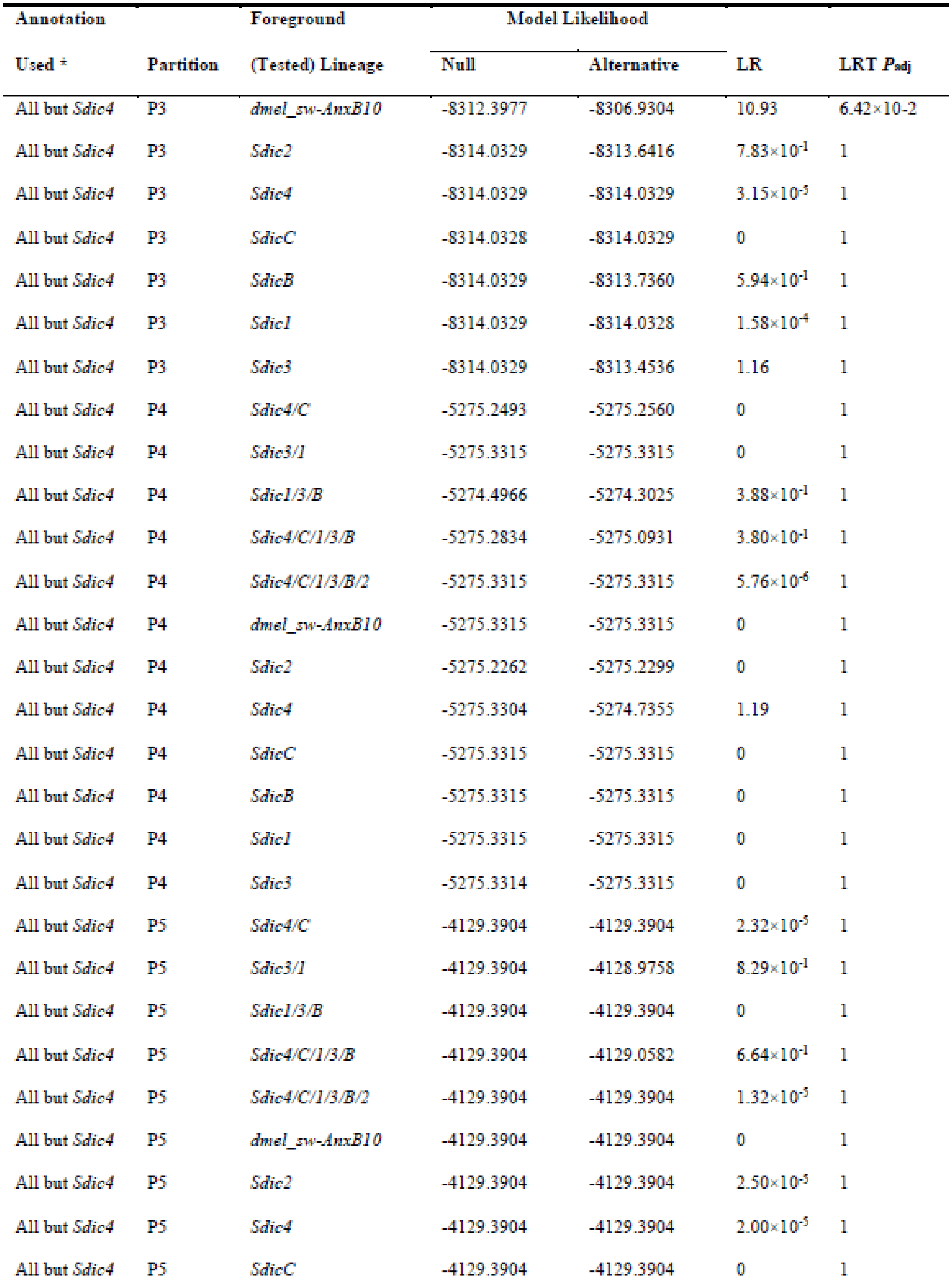
* By Bonferroni correction.

Supplementary Table S1.7. Assessment of the action of positive selection across *Sdic* partitions



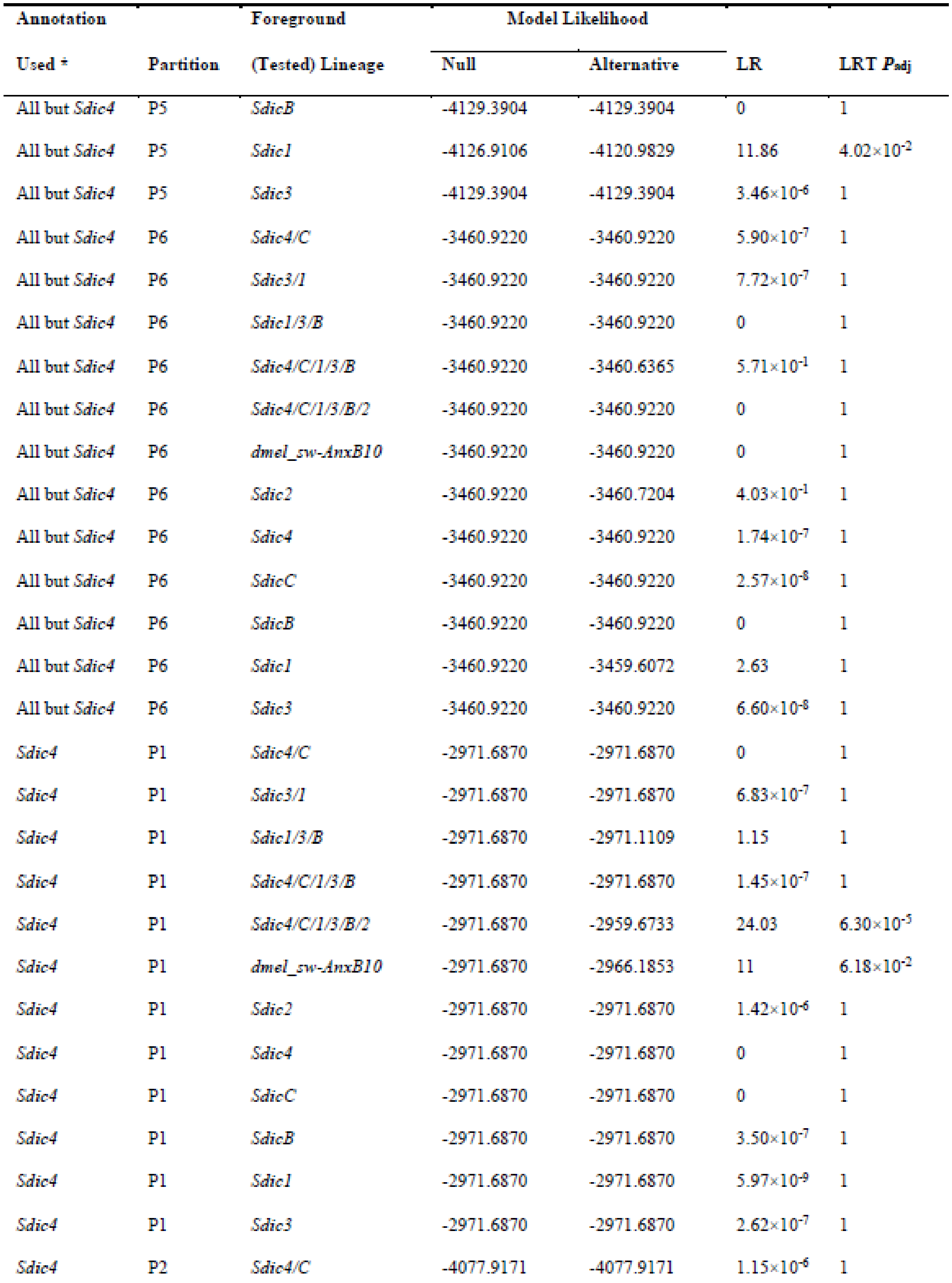
Annotation		Foreground	Model Likelihood			
Used *	Partition	(Tested) Lineage	Null	Alternative	LR	LRT P_{adj}
All but *Sdic4*	P1	*Sdic3/1*	-3305.5121	-3305.5121	3.01×10^{-7}	1
All but *Sdic4*	P1	*Sdic1/3/B*	-3305.5121	-3305.4939	3.63×10^{-2}	1
All but *Sdic4*	P1	*Sdic4/C/1/3/B*	-3305.5121	-3305.5121	0	1
All but *Sdic4*	P1	*Sdic4/C/1/3/B/2*	-3305.5121	-3293.4423	24.14	6.39×10^{-5}
All but *Sdic4*	P1	*dmel_sw-AnxB10*	-3305.5121	-3300.0001	11.02	6.21×10^{-2}
All but *Sdic4*	P1	*Sdic2*	-3305.5121	-3305.5121	0	1
All but *Sdic4*	P1	*Sdic4*	-3305.5121	-3305.5121	1.23×10^{-6}	1
All but *Sdic4*	P1	*SdicC*	-3305.5121	-3305.5121	0	1
All but *Sdic4*	P1	*SdicB*	-3305.5121	-3305.5121	0	1
All but *Sdic4*	P1	*Sdic1*	-3305.5121	-3305.5121	1.19×10^{-7}	1
All but *Sdic4*	P1	*Sdic3*	-3305.5121	-3305.5121	7.32×10^{-7}	1
All but *Sdic4*	P2	*Sdic4/C*	-4414.7773	-4414.7773	9.37×10^{-6}	1
All but *Sdic4*	P2	*Sdic3/1*	-4414.7773	-4413.7651	2.02	1
All but *Sdic4*	P2	*Sdic1/3/B*	-4414.7774	-4414.7773	3.53×10^{-5}	1
All but *Sdic4*	P2	*Sdic4/C/1/3/B*	-4414.7773	-4414.7773	0	1
All but *Sdic4*	P2	*Sdic4/C/1/3/B/2*	-4414.7773	-4414.7773	0	1
All but *Sdic4*	P2	*dmel_sw-AnxB10*	-4413.8010	-4411.8673	3.87	1
All but *Sdic4*	P2	*Sdic2*	-4414.7566	-4414.5492	4.15×10^{-1}	1
All but *Sdic4*	P2	*Sdic4*	-4414.7773	-4414.7773	0	1
All but *Sdic4*	P2	*SdicC*	-4414.7773	-4414.7773	0	1
All but *Sdic4*	P2	*SdicB*	-4414.7773	-4414.7773	0	1
All but *Sdic4*	P2	*Sdic1*	-4414.7773	-4414.7773	0	1
All but *Sdic4*	P2	*Sdic3*	-4414.4699	-4414.0461	8.48×10^{-1}	1
All but *Sdic4*	P3	*Sdic4/C*	-8314.0329	-8314.0329	0	1
All but *Sdic4*	P3	*Sdic3/1*	-8314.0329	-8314.0328	1.11×10^{-4}	1
All but *Sdic4*	P3	*Sdic1/3/B*	-8314.0329	-8314.0329	0	1
All but *Sdic4*	P3	*Sdic4/C/1/3/B*	-8314.0329	-8314.0331	0	1
All but *Sdic4*	P3	*Sdic4/C/1/3/B/2*	-8305.6697	-8292.2340	26.87	1.44×10^{-5}

Supplementary Table S1.7. Assessment of the action of positive selection across *Sdic* partitions



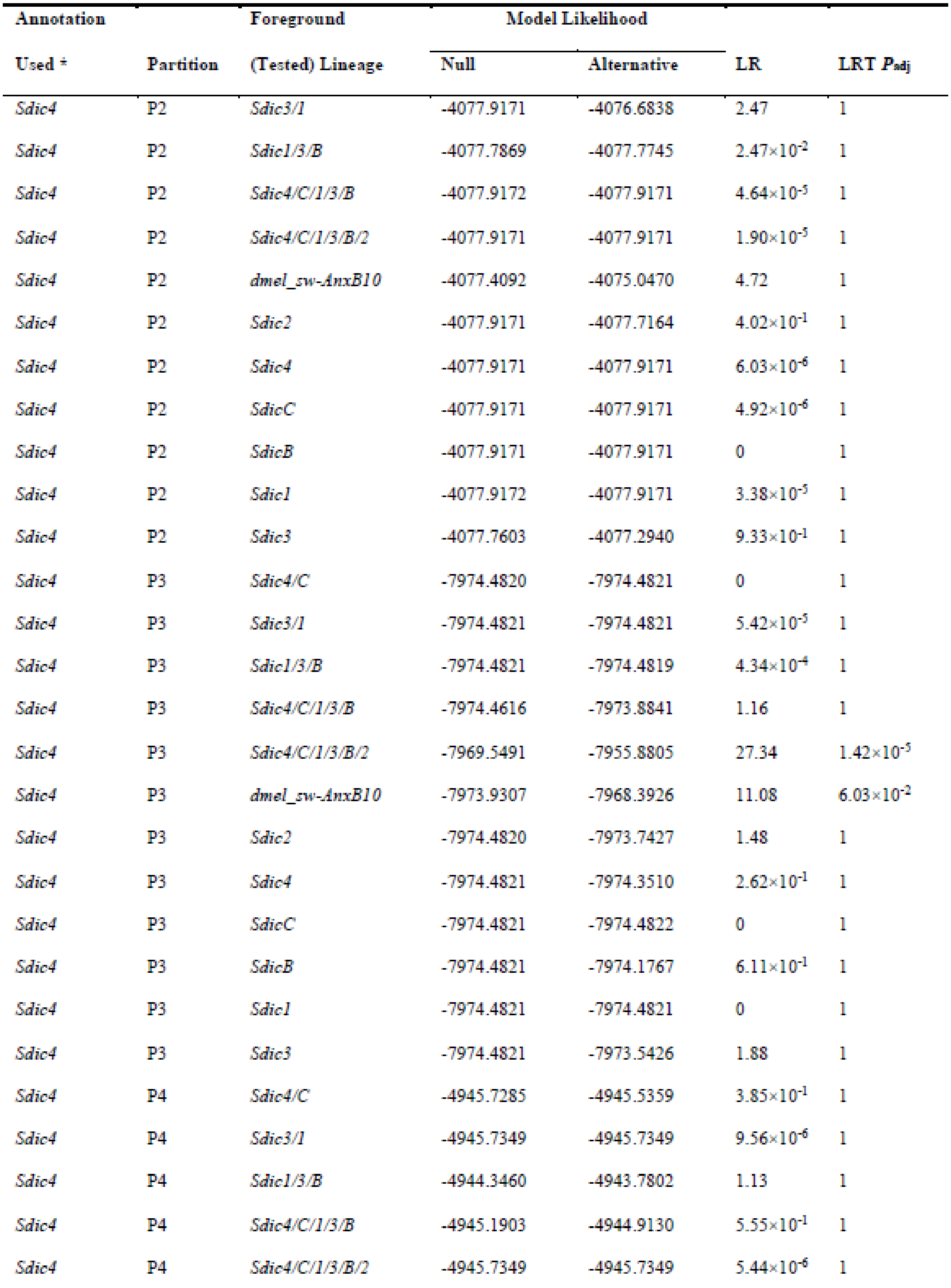
Annotation		Foreground	Model Likelihood			
Used *	Partition	(Tested) Lineage	Null	Alternative	LR	LRT P_{adj}
All but *Sdic4*	P3	*dmel_sw-AnxB10*	-8312.3977	-8306.9304	10.93	6.42×10-2
All but *Sdic4*	P3	*Sdic2*	-8314.0329	-8313.6416	7.83×10^{-1}	1
All but *Sdic4*	P3	*Sdic4*	-8314.0329	-8314.0329	3.15×10^{-5}	1
All but *Sdic4*	P3	*SdicC*	-8314.0328	-8314.0329	0	1
All but *Sdic4*	P3	*SdicB*	-8314.0329	-8313.7360	5.94×10^{-1}	1
All but *Sdic4*	P3	*Sdic1*	-8314.0329	-8314.0328	1.58×10^{-4}	1
All but *Sdic4*	P3	*Sdic3*	-8314.0329	-8313.4536	1.16	1
All but *Sdic4*	P4	*Sdic4/C*	-5275.2493	-5275.2560	0	1
All but *Sdic4*	P4	*Sdic3/1*	-5275.3315	-5275.3315	0	1
All but *Sdic4*	P4	*Sdic1/3/B*	-5274.4966	-5274.3025	3.88×10^{-1}	1
All but *Sdic4*	P4	*Sdic4/C/1/3/B*	-5275.2834	-5275.0931	3.80×10^{-1}	1
All but *Sdic4*	P4	*Sdic4/C/1/3/B/2*	-5275.3315	-5275.3315	5.76×10^{-6}	1
All but *Sdic4*	P4	*dmel_sw-AnxB10*	-5275.3315	-5275.3315	0	1
All but *Sdic4*	P4	*Sdic2*	-5275.2262	-5275.2299	0	1
All but *Sdic4*	P4	*Sdic4*	-5275.3304	-5274.7355	1.19	1
All but *Sdic4*	P4	*SdicC*	-5275.3315	-5275.3315	0	1
All but *Sdic4*	P4	*SdicB*	-5275.3315	-5275.3315	0	1
All but *Sdic4*	P4	*Sdic1*	-5275.3315	-5275.3315	0	1
All but *Sdic4*	P4	*Sdic3*	-5275.3314	-5275.3315	0	1
All but *Sdic4*	P5	*Sdic4/C*	-4129.3904	-4129.3904	2.32×10^{-5}	1
All but *Sdic4*	P5	*Sdic3/1*	-4129.3904	-4128.9758	8.29×10^{-1}	1
All but *Sdic4*	P5	*Sdic1/3/B*	-4129.3904	-4129.3904	0	1
All but *Sdic4*	P5	*Sdic4/C/1/3/B*	-4129.3904	-4129.0582	6.64×10^{-1}	1
All but *Sdic4*	P5	*Sdic4/C/1/3/B/2*	-4129.3904	-4129.3904	1.32×10^{-5}	1
All but *Sdic4*	P5	*dmel_sw-AnxB10*	-4129.3904	-4129.3904	0	1
All but *Sdic4*	P5	*Sdic2*	-4129.3904	-4129.3904	2.50×10^{-5}	1
All but *Sdic4*	P5	*Sdic4*	-4129.3904	-4129.3904	2.00×10^{-5}	1
All but *Sdic4*	P5	*SdicC*	-4129.3904	-4129.3904	0	1

Supplementary Table S1.7. Assessment of the action of positive selection across *Sdic* partitions



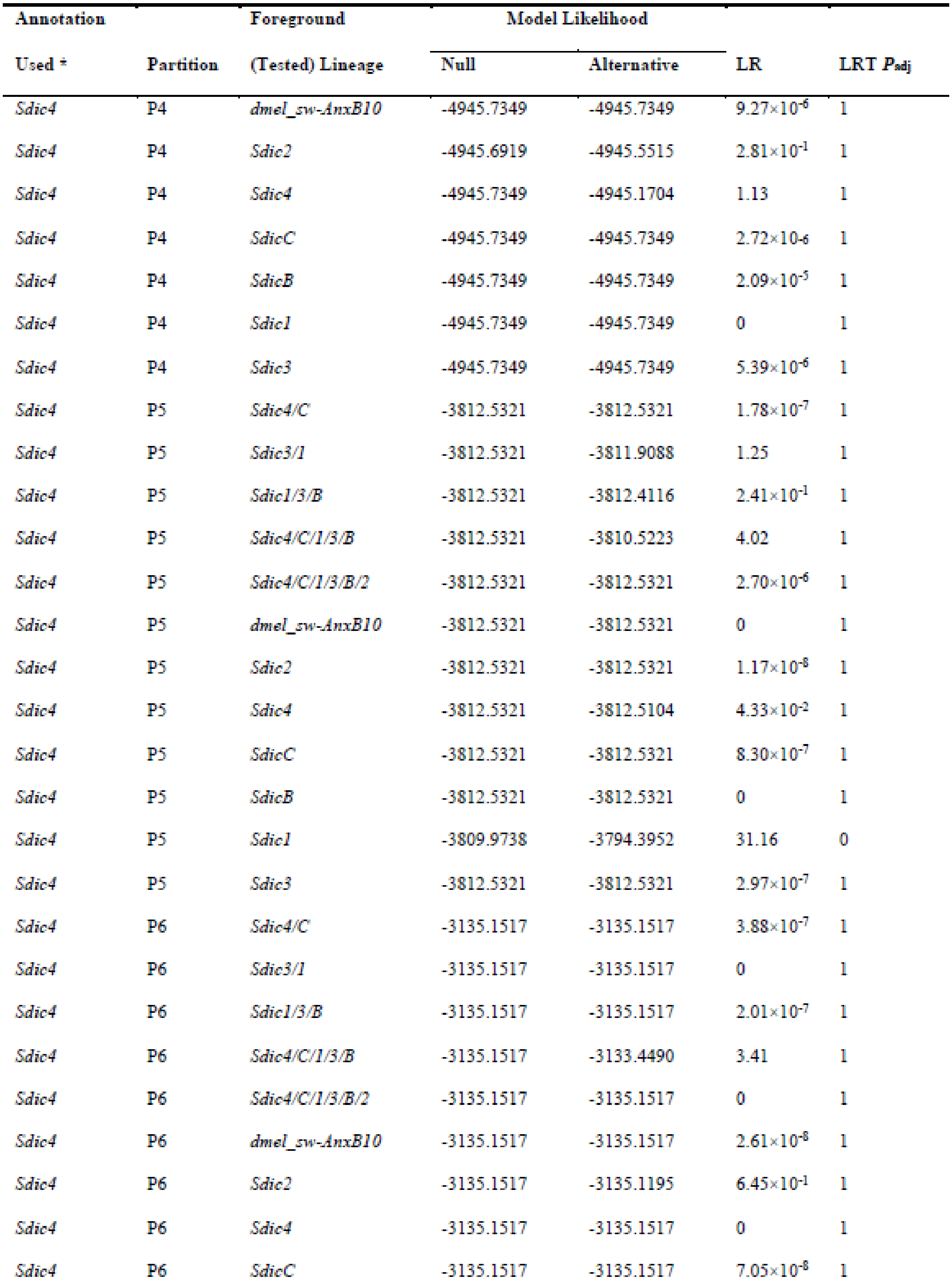
Annotation		Foreground	Model Likelihood			
Used *	Partition	(Tested) Lineage	Null	Alternative	LR	LRT P_{adj}
All but *Sdic4*	P5	*SdicB*	-4129.3904	-4129.3904	0	1
All but *Sdic4*	P5	*Sdic1*	-4126.9106	-4120.9829	11.86	4.02×10^{-2}
All but *Sdic4*	P5	*Sdic3*	-4129.3904	-4129.3904	3.46×10^{-6}	1
All but *Sdic4*	P6	*Sdic4/C*	-3460.9220	-3460.9220	5.90×10^{-7}	1
All but *Sdic4*	P6	*Sdic3/1*	-3460.9220	-3460.9220	7.72×10^{-7}	1
All but *Sdic4*	P6	*Sdic1/3/B*	-3460.9220	-3460.9220	0	1
All but *Sdic4*	P6	*Sdic4/C/1/3/B*	-3460.9220	-3460.6365	5.71×10^{-1}	1
All but *Sdic4*	P6	*Sdic4/C/1/3/B/2*	-3460.9220	-3460.9220	0	1
All but *Sdic4*	P6	*dmel_sw-AnxB10*	-3460.9220	-3460.9220	0	1
All but *Sdic4*	P6	*Sdic2*	-3460.9220	-3460.7204	4.03×10^{-1}	1
All but *Sdic4*	P6	*Sdic4*	-3460.9220	-3460.9220	1.74×10^{-7}	1
All but *Sdic4*	P6	*SdicC*	-3460.9220	-3460.9220	2.57×10^{-8}	1
All but *Sdic4*	P6	*SdicB*	-3460.9220	-3460.9220	0	1
All but *Sdic4*	P6	*Sdic1*	-3460.9220	-3459.6072	2.63	1
All but *Sdic4*	P6	*Sdic3*	-3460.9220	-3460.9220	6.60×10^{-8}	1
Sdic4	P1	*Sdic4/C*	-2971.6870	-2971.6870	0	1
Sdic4	P1	*Sdic3/1*	-2971.6870	-2971.6870	6.83×10^{-7}	1
Sdic4	P1	*Sdic1/3/B*	-2971.6870	-2971.1109	1.15	1
Sdic4	P1	*Sdic4/C/1/3/B*	-2971.6870	-2971.6870	1.45×10^{-7}	1
Sdic4	P1	*Sdic4/C/1/3/B/2*	-2971.6870	-2959.6733	24.03	6.30×10^{-5}
Sdic4	P1	*dmel_sw-AnxB10*	-2971.6870	-2966.1853	11	6.18×10^{-2}
Sdic4	P1	*Sdic2*	-2971.6870	-2971.6870	1.42×10^{-6}	1
Sdic4	P1	*Sdic4*	-2971.6870	-2971.6870	0	1
Sdic4	P1	*SdicC*	-2971.6870	-2971.6870	0	1
Sdic4	P1	*SdicB*	-2971.6870	-2971.6870	3.50×10^{-7}	1
Sdic4	P1	*Sdic1*	-2971.6870	-2971.6870	5.97×10^{-9}	1
Sdic4	P1	*Sdic3*	-2971.6870	-2971.6870	2.62×10^{-7}	1
Sdic4	P2	*Sdic4/C*	-4077.9171	-4077.9171	1.15×10^{-6}	1

Supplementary Table S1.7. Assessment of the action of positive selection across *Sdic* partitions



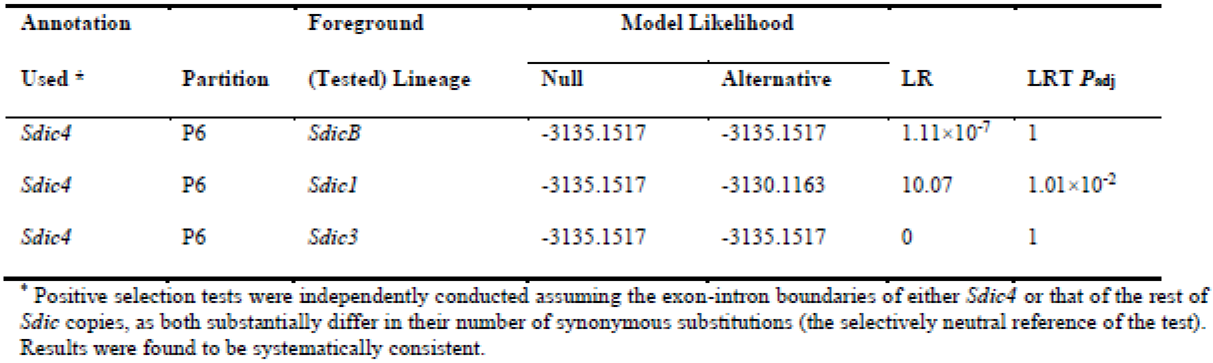
Annotation		Foreground	Model Likelihood			
Used *	Partition	(Tested) Lineage	Null	Alternative	LR	LRT P_{adj}
Sdic4	P2	*Sdic3/1*	-4077.9171	-4076.6838	2.47	1
Sdic4	P2	*Sdic1/3/B*	-4077.7869	-4077.7745	2.47×10^{-2}	1
Sdic4	P2	*Sdic4/C/1/3/B*	-4077.9172	-4077.9171	4.64×10^{-5}	1
Sdic4	P2	*Sdic4/C/1/3/B/2*	-4077.9171	-4077.9171	1.90×10^{-5}	1
Sdic4	P2	*dmel_sw-AnxB10*	-4077.4092	-4075.0470	4.72	1
Sdic4	P2	*Sdic2*	-4077.9171	-4077.7164	4.02×10^{-1}	1
Sdic4	P2	*Sdic4*	-4077.9171	-4077.9171	6.03×10^{-6}	1
Sdic4	P2	*SdicC*	-4077.9171	-4077.9171	4.92×10^{-6}	1
Sdic4	P2	*SdicB*	-4077.9171	-4077.9171	0	1
Sdic4	P2	*Sdic1*	-4077.9172	-4077.9171	3.38×10^{-5}	1
Sdic4	P2	*Sdic3*	-4077.7603	-4077.2940	9.33×10^{-1}	1
Sdic4	P3	*Sdic4/C*	-7974.4820	-7974.4821	0	1
Sdic4	P3	*Sdic3/1*	-7974.4821	-7974.4821	5.42×10^{-5}	1
Sdic4	P3	*Sdic1/3/B*	-7974.4821	-7974.4819	4.34×10^{-4}	1
Sdic4	P3	*Sdic4/C/1/3/B*	-7974.4616	-7973.8841	1.16	1
Sdic4	P3	*Sdic4/C/1/3/B/2*	-7969.5491	-7955.8805	27.34	1.42×10^{-5}
Sdic4	P3	*dmel_sw-AnxB10*	-7973.9307	-7968.3926	11.08	6.03×10^{-2}
Sdic4	P3	*Sdic2*	-7974.4820	-7973.7427	1.48	1
Sdic4	P3	*Sdic4*	-7974.4821	-7974.3510	2.62×10^{-1}	1
Sdic4	P3	*SdicC*	-7974.4821	-7974.4822	0	1
Sdic4	P3	*SdicB*	-7974.4821	-7974.1767	6.11×10^{-1}	1
Sdic4	P3	*Sdic1*	-7974.4821	-7974.4821	0	1
Sdic4	P3	*Sdic3*	-7974.4821	-7973.5426	1.88	1
Sdic4	P4	*Sdic4/C*	-4945.7285	-4945.5359	3.85×10^{-1}	1
Sdic4	P4	*Sdic3/1*	-4945.7349	-4945.7349	9.56×10^{-6}	1
Sdic4	P4	*Sdic1/3/B*	-4944.3460	-4943.7802	1.13	1
Sdic4	P4	*Sdic4/C/1/3/B*	-4945.1903	-4944.9130	5.55×10^{-1}	1
Sdic4	P4	*Sdic4/C/1/3/B/2*	-4945.7349	-4945.7349	5.44×10^{-6}	1

Supplementary Table S1.7. Assessment of the action of positive selection across *Sdic* partitions



Annotation		Foreground	Model Likelihood			
Used *	Partition	(Tested) Lineage	Null	Alternative	LR	LRT P_{adj}
Sdic4	P4	*dmel_sw-AnxB10*	-4945.7349	-4945.7349	9.27×10^{-6}	1
Sdic4	P4	*Sdic2*	-4945.6919	-4945.5515	2.81×10^{-1}	1
Sdic4	P4	*Sdic4*	-4945.7349	-4945.1704	1.13	1
Sdic4	P4	*SdicC*	-4945.7349	-4945.7349	2.72×10-6	1
Sdic4	P4	*SdicB*	-4945.7349	-4945.7349	2.09×10^{-5}	1
Sdic4	P4	*Sdic1*	-4945.7349	-4945.7349	0	1
Sdic4	P4	*Sdic3*	-4945.7349	-4945.7349	5.39×10^{-6}	1
Sdic4	P5	*Sdic4/C*	-3812.5321	-3812.5321	1.78×10^{-7}	1
Sdic4	P5	*Sdic3/1*	-3812.5321	-3811.9088	1.25	1
Sdic4	P5	*Sdic1/3/B*	-3812.5321	-3812.4116	2.41×10^{-1}	1
Sdic4	P5	*Sdic4/C/1/3/B*	-3812.5321	-3810.5223	4.02	1
Sdic4	P5	*Sdic4/C/1/3/B/2*	-3812.5321	-3812.5321	2.70×10^{-6}	1
Sdic4	P5	*dmel_sw-AnxB10*	-3812.5321	-3812.5321	0	1
Sdic4	P5	*Sdic2*	-3812.5321	-3812.5321	1.17×10^{-8}	1
Sdic4	P5	*Sdic4*	-3812.5321	-3812.5104	4.33×10^{-2}	1
Sdic4	P5	*SdicC*	-3812.5321	-3812.5321	8.30×10^{-7}	1
Sdic4	P5	*SdicB*	-3812.5321	-3812.5321	0	1
Sdic4	P5	*Sdic1*	-3809.9738	-3794.3952	31.16	0
Sdic4	P5	*Sdic3*	-3812.5321	-3812.5321	2.97×10^{-7}	1
Sdic4	P6	*Sdic4/C*	-3135.1517	-3135.1517	3.88×10^{-7}	1
Sdic4	P6	*Sdic3/1*	-3135.1517	-3135.1517	0	1
Sdic4	P6	*Sdic1/3/B*	-3135.1517	-3135.1517	2.01×10^{-7}	1
Sdic4	P6	*Sdic4/C/1/3/B*	-3135.1517	-3133.4490	3.41	1
Sdic4	P6	*Sdic4/C/1/3/B/2*	-3135.1517	-3135.1517	0	1
Sdic4	P6	*dmel_sw-AnxB10*	-3135.1517	-3135.1517	2.61×10^{-8}	1
Sdic4	P6	*Sdic2*	-3135.1517	-3135.1195	6.45×10^{-1}	1
Sdic4	P6	*Sdic4*	-3135.1517	-3135.1517	0	1
Sdic4	P6	*SdicC*	-3135.1517	-3135.1517	7.05×10^{-8}	1

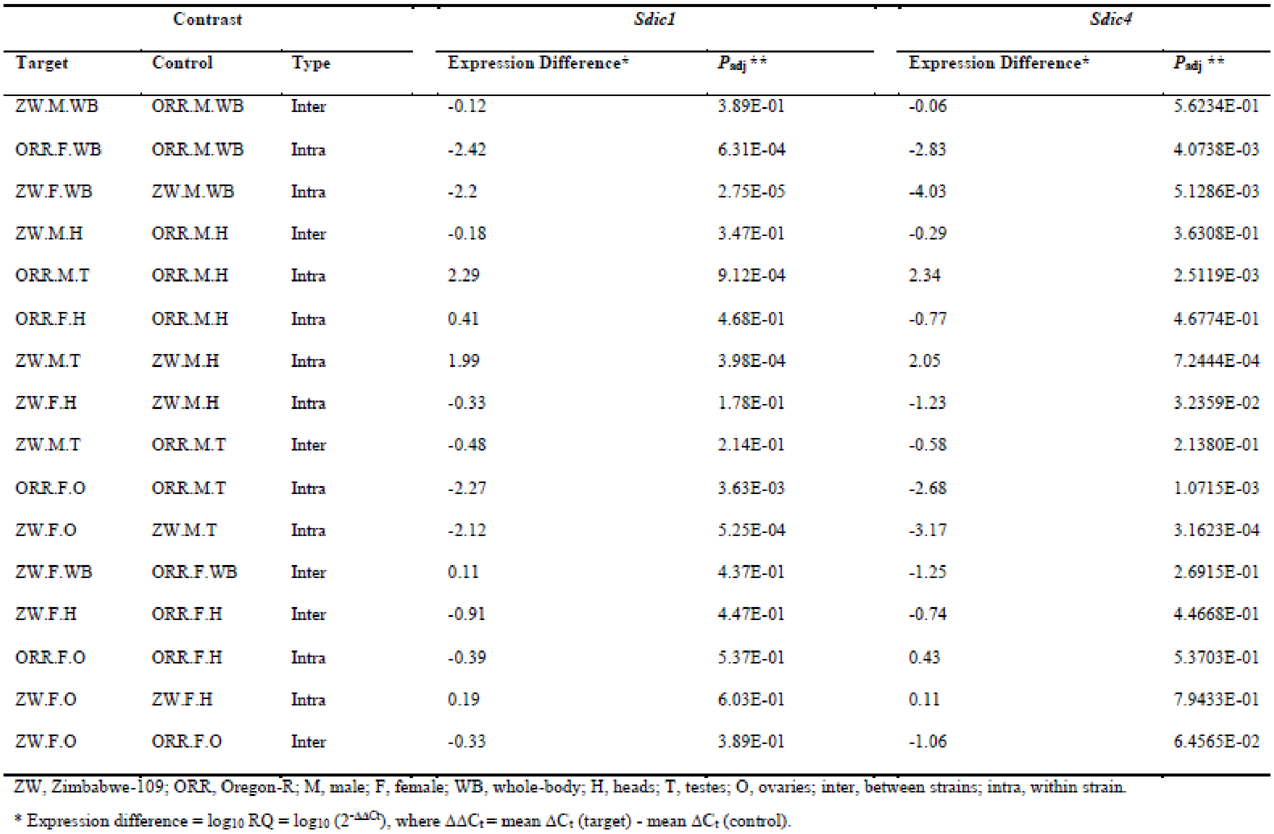
Supplementary Table S1.7. Assessment of the action of positive selection across *Sdic* partitions



Annotation		Foreground	Model Likelihood			
Used *	Partition	(Tested) Lineage	Null	Alternative	LR	LRT P_{adj}
Sdic4	P6	*SdicB*	-3135.1517	-3135.1517	1.11×10^{-7}	1
Sdic4	P6	*Sdic1*	-3135.1517	-3130.1163	10.07	1.01×10^{-2}
Sdic4	P6	*Sdic3*	-3135.1517	-3135.1517	0	1

* Positive selection tests were independently conducted assuming the exon-intron boundaries of either *Sdic4* or that of the rest of *Sdic* copies, as both substantially differ in their number of synonymous substitutions (the selectively neutral reference of the test). Results were found to be systematically consistent.

Supplementary Table S1.8. Expression differences according to qRT-PCR assays for *Sdic1* and *Sdic4* across different biological samples and strains

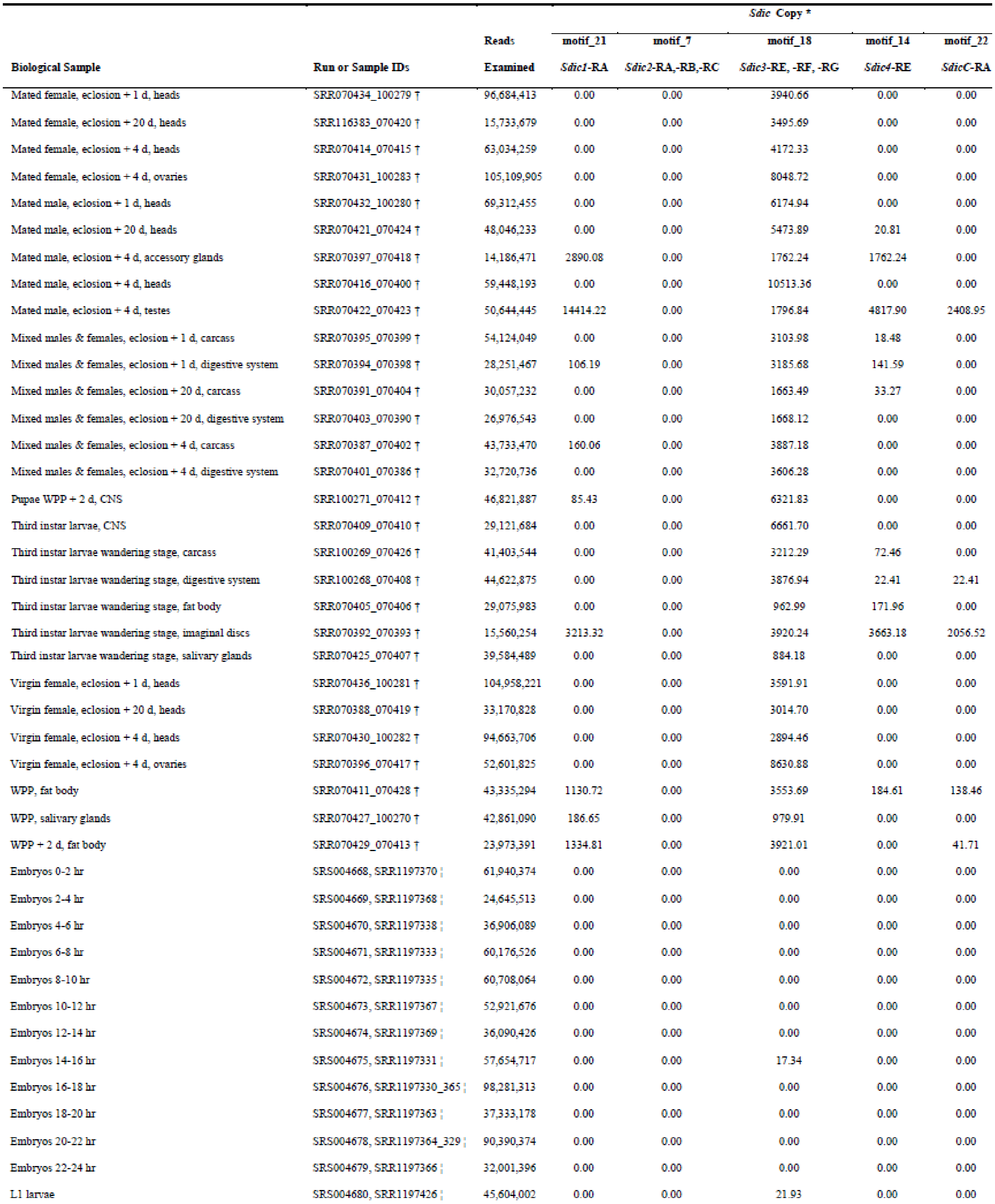


Contrast			*Sdic1*		*Sdic4*	
Target	**Control**	**Type**	**Expression Difference***	P_{adj} **	**Expression Difference***	P_{adj} **
ZW.M.WB	ORR.M.WB	Inter	-0.12	3.89E-01	-0.06	5.6234E-01
ORR.F.WB	ORR.M.WB	Intra	-2.42	6.31E-04	-2.83	4.0738E-03
ZW.F.WB	ZW.M.WB	Intra	-2.2	2.75E-05	-4.03	5.1286E-03
ZW.M.H	ORR.M.H	Inter	-0.18	3.47E-01	-0.29	3.6308E-01
ORR.M.T	ORR.M.H	Intra	2.29	9.12E-04	2.34	2.5119E-03
ORR.F.H	ORR.M.H	Intra	0.41	4.68E-01	-0.77	4.6774E-01
ZW.M.T	ZW.M.H	Intra	1.99	3.98E-04	2.05	7.2444E-04
ZW.F.H	ZW.M.H	Intra	-0.33	1.78E-01	-1.23	3.2359E-02
ZW.M.T	ORR.M.T	Inter	-0.48	2.14E-01	-0.58	2.1380E-01
ORR.F.O	ORR.M.T	Intra	-2.27	3.63E-03	-2.68	1.0715E-03
ZW.F.O	ZW.M.T	Intra	-2.12	5.25E-04	-3.17	3.1623E-04
ZW.F.WB	ORR.F.WB	Inter	0.11	4.37E-01	-1.25	2.6915E-01
ZW.F.H	ORR.F.H	Inter	-0.91	4.47E-01	-0.74	4.4668E-01
ORR.F.O	ORR.F.H	Intra	-0.39	5.37E-01	0.43	5.3703E-01
ZW.F.O	ZW.F.H	Intra	0.19	6.03E-01	0.11	7.9433E-01
ZW.F.O	ORR.F.O	Inter	-0.33	3.89E-01	-1.06	6.4565E-02

ZW, Zimbabwe-109; ORR, Oregon-R; M, male; F, female; WB, whole-body; H, heads; T, testes; O, ovaries; inter, between strains; intra, within strain.

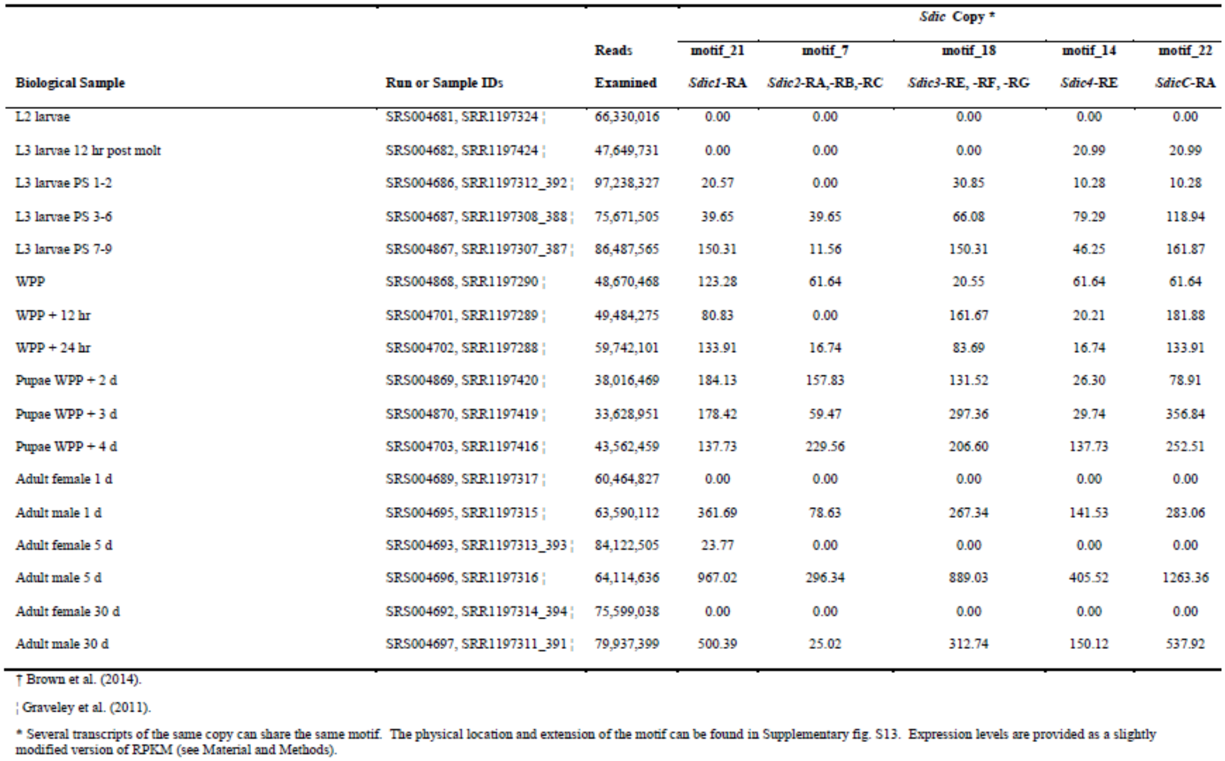
* Expression difference = $\log_{10}$ RQ = $\log_{10}$ ($2^{-\Delta\Delta C_t}$), where $\Delta\Delta C_t$ = mean ΔC_t (target) - mean ΔC_t (control).

Supplementary Table S1.9. Expression levels of five *Sdic* copies across 59 biological conditions as assayed by RNA-seq



			Sdic Copy *				
		Reads	motif_21	motif_7	motif_18	motif_14	motif_22
Biological Sample	Run or Sample IDs	Examined	*Sdic1*-RA	*Sdic2*-RA,-RB,-RC	*Sdic3*-RE, -RF, -RG	*Sdic4*-RE	*SdicC*-RA
Mated female, eclosion + 1 d, heads	SRR070434_100279 †	96,684,413	0.00	0.00	3940.66	0.00	0.00
Mated female, eclosion + 20 d, heads	SRR116383_070420 †	15,733,679	0.00	0.00	3495.69	0.00	0.00
Mated female, eclosion + 4 d, heads	SRR070414_070415 †	63,034,259	0.00	0.00	4172.33	0.00	0.00
Mated female, eclosion + 4 d, ovaries	SRR070431_100283 †	105,109,905	0.00	0.00	8048.72	0.00	0.00
Mated male, eclosion + 1 d, heads	SRR070432_100280 †	69,312,455	0.00	0.00	6174.94	0.00	0.00
Mated male, eclosion + 20 d, heads	SRR070421_070424 †	48,046,233	0.00	0.00	5473.89	20.81	0.00
Mated male, eclosion + 4 d, accessory glands	SRR070397_070418 †	14,186,471	2890.08	0.00	1762.24	1762.24	0.00
Mated male, eclosion + 4 d, heads	SRR070416_070400 †	59,448,193	0.00	0.00	10513.36	0.00	0.00
Mated male, eclosion + 4 d, testes	SRR070422_070423 †	50,644,445	14414.22	0.00	1796.84	4817.90	2408.95
Mixed males & females, eclosion + 1 d, carcass	SRR070395_070399 †	54,124,049	0.00	0.00	3103.98	18.48	0.00
Mixed males & females, eclosion + 1 d, digestive system	SRR070394_070398 †	28,251,467	106.19	0.00	3185.68	141.59	0.00
Mixed males & females, eclosion + 20 d, carcass	SRR070391_070404 †	30,057,232	0.00	0.00	1663.49	33.27	0.00
Mixed males & females, eclosion + 20 d, digestive system	SRR070403_070390 †	26,976,543	0.00	0.00	1668.12	0.00	0.00
Mixed males & females, eclosion + 4 d, carcass	SRR070387_070402 †	43,733,470	160.06	0.00	3887.18	0.00	0.00
Mixed males & females, eclosion + 4 d, digestive system	SRR070401_070386 †	32,720,736	0.00	0.00	3606.28	0.00	0.00
Pupae WPP + 2 d, CNS	SRR100271_070412 †	46,821,887	85.43	0.00	6321.83	0.00	0.00
Third instar larvae, CNS	SRR070409_070410 †	29,121,684	0.00	0.00	6661.70	0.00	0.00
Third instar larvae wandering stage, carcass	SRR100269_070426 †	41,403,544	0.00	0.00	3212.29	72.46	0.00
Third instar larvae wandering stage, digestive system	SRR100268_070408 †	44,622,875	0.00	0.00	3876.94	22.41	22.41
Third instar larvae wandering stage, fat body	SRR070405_070406 †	29,075,983	0.00	0.00	962.99	171.96	0.00
Third instar larvae wandering stage, imaginal discs	SRR070392_070393 †	15,560,254	3213.32	0.00	3920.24	3663.18	2056.52
Third instar larvae wandering stage, salivary glands	SRR070425_070407 †	39,584,489	0.00	0.00	884.18	0.00	0.00
Virgin female, eclosion + 1 d, heads	SRR070436_100281 †	104,958,221	0.00	0.00	3591.91	0.00	0.00
Virgin female, eclosion + 20 d, heads	SRR070388_070419 †	33,170,828	0.00	0.00	3014.70	0.00	0.00
Virgin female, eclosion + 4 d, heads	SRR070430_100282 †	94,663,706	0.00	0.00	2894.46	0.00	0.00
Virgin female, eclosion + 4 d, ovaries	SRR070396_070417 †	52,601,825	0.00	0.00	8630.88	0.00	0.00
WPP, fat body	SRR070411_070428 †	43,335,294	1130.72	0.00	3553.69	184.61	138.46
WPP, salivary glands	SRR070427_100270 †	42,861,090	186.65	0.00	979.91	0.00	0.00
WPP + 2 d, fat body	SRR070429_070413 †	23,973,391	1334.81	0.00	3921.01	0.00	41.71
Embryos 0-2 hr	SRS004668, SRR1197370 ‡	61,940,374	0.00	0.00	0.00	0.00	0.00
Embryos 2-4 hr	SRS004669, SRR1197368 ‡	24,645,513	0.00	0.00	0.00	0.00	0.00
Embryos 4-6 hr	SRS004670, SRR1197338 ‡	36,906,089	0.00	0.00	0.00	0.00	0.00
Embryos 6-8 hr	SRS004671, SRR1197333 ‡	60,176,526	0.00	0.00	0.00	0.00	0.00
Embryos 8-10 hr	SRS004672, SRR1197335 ‡	60,708,064	0.00	0.00	0.00	0.00	0.00
Embryos 10-12 hr	SRS004673, SRR1197367 ‡	52,921,676	0.00	0.00	0.00	0.00	0.00
Embryos 12-14 hr	SRS004674, SRR1197369 ‡	36,090,426	0.00	0.00	0.00	0.00	0.00
Embryos 14-16 hr	SRS004675, SRR1197331 ‡	57,654,717	0.00	0.00	17.34	0.00	0.00
Embryos 16-18 hr	SRS004676, SRR1197330_365 ‡	98,281,313	0.00	0.00	0.00	0.00	0.00
Embryos 18-20 hr	SRS004677, SRR1197363 ‡	37,333,178	0.00	0.00	0.00	0.00	0.00
Embryos 20-22 hr	SRS004678, SRR1197364_329 ‡	90,390,374	0.00	0.00	0.00	0.00	0.00
Embryos 22-24 hr	SRS004679, SRR1197366 ‡	32,001,396	0.00	0.00	0.00	0.00	0.00
L1 larvae	SRS004680, SRR1197426 ‡	45,604,002	0.00	0.00	21.93	0.00	0.00

Supplementary Table S1.9. Expression levels of five *Sdic* copies across 59 biological conditions as assayed by RNA-seq



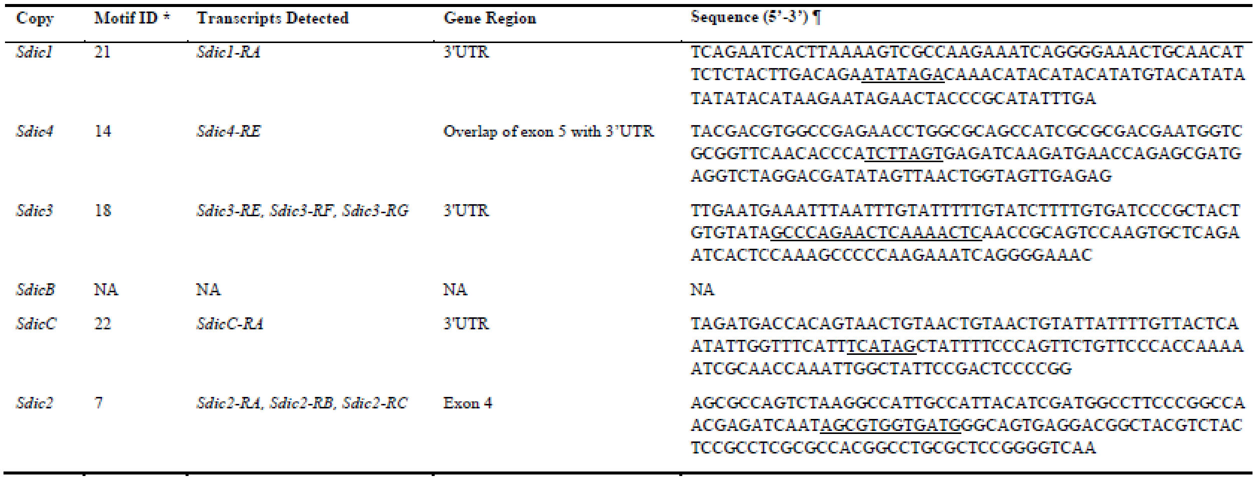
		Reads	*Sdic* Copy *				
			motif_21	motif_7	motif_18	motif_14	motif_22
Biological Sample	**Run or Sample IDs**	**Examined**	***Sdic1*-RA**	***Sdic2*-RA,-RB,-RC**	***Sdic3*-RE, -RF, -RG**	***Sdic4*-RE**	***SdicC*-RA**
L2 larvae	SRS004681, SRR1197324 ¦	66,330,016	0.00	0.00	0.00	0.00	0.00
L3 larvae 12 hr post molt	SRS004682, SRR1197424 ¦	47,649,731	0.00	0.00	0.00	20.99	20.99
L3 larvae PS 1-2	SRS004686, SRR1197312_392 ¦	97,238,327	20.57	0.00	30.85	10.28	10.28
L3 larvae PS 3-6	SRS004687, SRR1197308_388 ¦	75,671,505	39.65	39.65	66.08	79.29	118.94
L3 larvae PS 7-9	SRS004867, SRR1197307_387 ¦	86,487,565	150.31	11.56	150.31	46.25	161.87
WPP	SRS004868, SRR1197290 ¦	48,670,468	123.28	61.64	20.55	61.64	61.64
WPP + 12 hr	SRS004701, SRR1197289 ¦	49,484,275	80.83	0.00	161.67	20.21	181.88
WPP + 24 hr	SRS004702, SRR1197288 ¦	59,742,101	133.91	16.74	83.69	16.74	133.91
Pupae WPP + 2 d	SRS004869, SRR1197420 ¦	38,016,469	184.13	157.83	131.52	26.30	78.91
Pupae WPP + 3 d	SRS004870, SRR1197419 ¦	33,628,951	178.42	59.47	297.36	29.74	356.84
Pupae WPP + 4 d	SRS004703, SRR1197416 ¦	43,562,459	137.73	229.56	206.60	137.73	252.51
Adult female 1 d	SRS004689, SRR1197317 ¦	60,464,827	0.00	0.00	0.00	0.00	0.00
Adult male 1 d	SRS004695, SRR1197315 ¦	63,590,112	361.69	78.63	267.34	141.53	283.06
Adult female 5 d	SRS004693, SRR1197313_393 ¦	84,122,505	23.77	0.00	0.00	0.00	0.00
Adult male 5 d	SRS004696, SRR1197316 ¦	64,114,636	967.02	296.34	889.03	405.52	1263.36
Adult female 30 d	SRS004692, SRR1197314_394 ¦	75,599,038	0.00	0.00	0.00	0.00	0.00
Adult male 30 d	SRS004697, SRR1197311_391 ¦	79,937,399	500.39	25.02	312.74	150.12	537.92

† Brown et al. (2014).

¦ Graveley et al. (2011).

* Several transcripts of the same copy can share the same motif. The physical location and extension of the motif can be found in Supplementary fig. S13. Expression levels are provided as a slightly modified version of RPKM (see Material and Methods).

Supplementary Table S1.10. Diagnostic motifs used for expression profiling of different *Sdic* copies in *Berlin* assembly using RNA-seq data

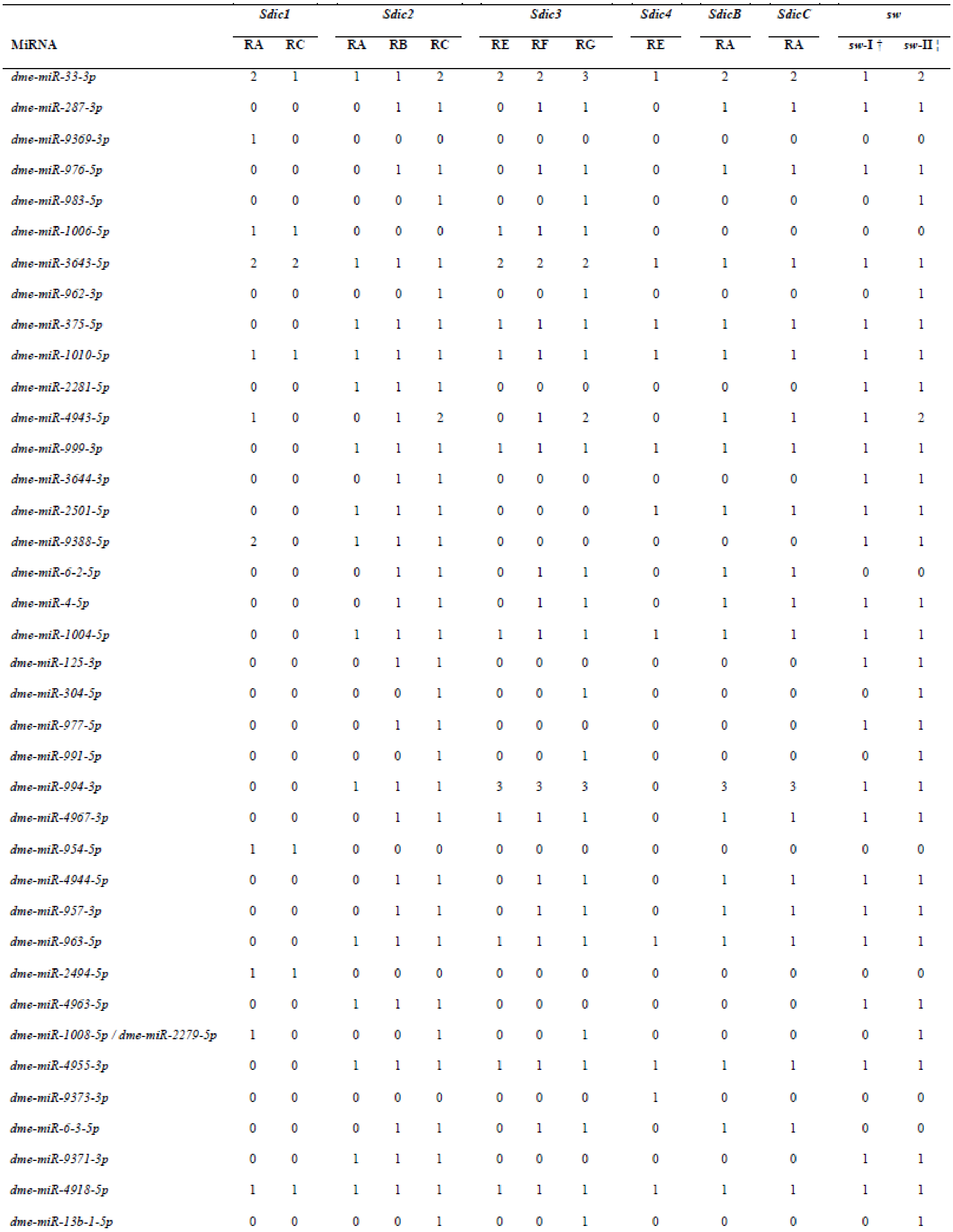


Copy	Motif ID *	Transcripts Detected	Gene Region	Sequence (5'-3') ¶
Sdic1	21	*Sdic1-RA*	3'UTR	TCAGAATCACTTAAAAGTCGCCAAGAAATCAGGGGAAACTGCAACATTCTCTACTTGACAGAATATAGACAAACATACATACATATGTACATATATATATACATAAGAATAGAACTACCCGCATATTTGA
Sdic4	14	*Sdic4-RE*	Overlap of exon 5 with 3'UTR	TACGACGTGGCCGAGAACCTGGCGCAGCCATCGCGCGACGAATGGTCGCGGTTCAACACCCATCTTAGTGAGATCAAGATGAACCAGAGCGATGAGGTCTAGGACGATATAGTTAACTGGTAGTTGAGAG
Sdic3	18	*Sdic3-RE, Sdic3-RF, Sdic3-RG*	3'UTR	TTGAATGAAATTTAATTTGTATTTTTGTATCTTTTGTGATCCCGCTACTGTGTATAGCCCAGAACTCAAAACTCAACCGCAGTCCAAGTGCTCAGAATCACTCCAAAGCCCCCAAGAAATCAGGGGAAAC
SdicB	NA	NA	NA	NA
SdicC	22	*SdicC-RA*	3'UTR	TAGATGACCACAGTAACTGTAACTGTAACTGTATTATTTTGTTACTCAATATTGGTTTCATTTCATAGCTATTTTCCCAGTTCTGTTCCCACCAAAAATCGCAACCAAATTGGCTATTCCGACTCCCCGG
Sdic2	7	*Sdic2-RA, Sdic2-RB, Sdic2-RC*	Exon 4	AGCGCCAGTCTAAGGCCATTGCCATTACATCGATGGCCTTCCCGGCCAACGAGATCAATAGCGTGGTGATGGGCAGTGAGGACGGCTACGTCTACTCCGCCTCGCGCCACGGCCTGCGCTCCGGGGTCAA

* As in Supplementary fig. S13.

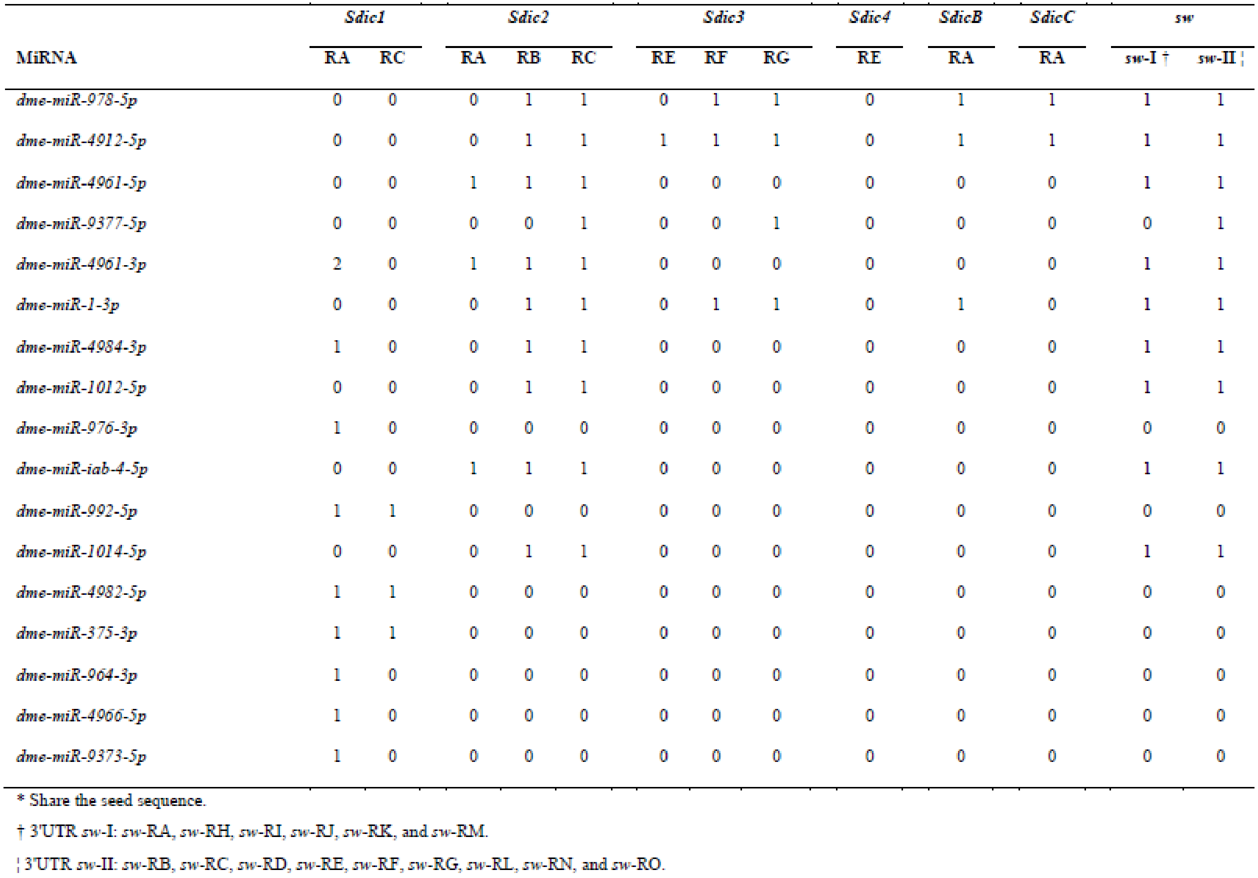
¶ Underlined, the core motif that is informative of the expression of a particular copy or transcript.

Supplementary Table S1.11. MiRNA binding sites in the 3’UTRs of different *Sdic* and *sw* transcripts



	Sdic1		*Sdic2*			*Sdic3*			*Sdic4*	*SdicB*	*SdicC*	*sw*	
MiRNA	RA	RC	RA	RB	RC	RE	RF	RG	RE	RA	RA	*sw*-I †	*sw*-II ‡
dme-miR-33-3p	2	1	1	1	2	2	2	3	1	2	2	1	2
dme-miR-287-3p	0	0	0	1	1	0	1	1	0	1	1	1	1
dme-miR-9369-3p	1	0	0	0	0	0	0	0	0	0	0	0	0
dme-miR-976-5p	0	0	0	1	1	0	1	1	0	1	1	1	1
dme-miR-983-5p	0	0	0	0	1	0	0	1	0	0	0	0	1
dme-miR-1006-5p	1	1	0	0	0	1	1	1	0	0	0	0	0
dme-miR-3643-5p	2	2	1	1	1	2	2	2	1	1	1	1	1
dme-miR-962-3p	0	0	0	0	1	0	0	1	0	0	0	0	1
dme-miR-375-5p	0	0	1	1	1	1	1	1	1	1	1	1	1
dme-miR-1010-5p	1	1	1	1	1	1	1	1	1	1	1	1	1
dme-miR-2281-5p	0	0	1	1	1	0	0	0	0	0	0	1	1
dme-miR-4943-5p	1	0	0	1	2	0	1	2	0	1	1	1	2
dme-miR-999-3p	0	0	1	1	1	1	1	1	1	1	1	1	1
dme-miR-3644-3p	0	0	0	1	1	0	0	0	0	0	0	1	1
dme-miR-2501-5p	0	0	1	1	1	0	0	0	1	1	1	1	1
dme-miR-9388-5p	2	0	1	1	1	0	0	0	0	0	0	1	1
dme-miR-6-2-5p	0	0	0	1	1	0	1	1	0	1	1	0	0
dme-miR-4-5p	0	0	0	1	1	0	1	1	0	1	1	1	1
dme-miR-1004-5p	0	0	1	1	1	1	1	1	1	1	1	1	1
dme-miR-125-3p	0	0	0	1	1	0	0	0	0	0	0	1	1
dme-miR-304-5p	0	0	0	0	1	0	0	1	0	0	0	0	1
dme-miR-977-5p	0	0	0	1	1	0	0	0	0	0	0	1	1
dme-miR-991-5p	0	0	0	0	1	0	0	1	0	0	0	0	1
dme-miR-994-3p	0	0	1	1	1	3	3	3	0	3	3	1	1
dme-miR-4967-3p	0	0	0	1	1	1	1	1	0	1	1	1	1
dme-miR-954-5p	1	1	0	0	0	0	0	0	0	0	0	0	0
dme-miR-4944-5p	0	0	0	1	1	0	1	1	0	1	1	1	1
dme-miR-957-3p	0	0	0	1	1	0	1	1	0	1	1	1	1
dme-miR-963-5p	0	0	1	1	1	1	1	1	1	1	1	1	1
dme-miR-2494-5p	1	1	0	0	0	0	0	0	0	0	0	0	0
dme-miR-4963-5p	0	0	1	1	1	0	0	0	0	0	0	1	1
dme-miR-1008-5p / dme-miR-2279-5p	1	0	0	0	1	0	0	1	0	0	0	0	1
dme-miR-4955-3p	0	0	1	1	1	1	1	1	1	1	1	1	1
dme-miR-9373-3p	0	0	0	0	0	0	0	0	1	0	0	0	0
dme-miR-6-3-5p	0	0	0	1	1	0	1	1	0	1	1	0	0
dme-miR-9371-3p	0	0	1	1	1	0	0	0	0	0	0	1	1
dme-miR-4918-5p	1	1	1	1	1	1	1	1	1	1	1	1	1
dme-miR-13b-1-5p	0	0	0	0	1	0	0	1	0	0	0	0	1

Supplementary Table S1.11. MiRNA binding sites in the 3’UTRs of different *Sdic* and *sw* transcripts



	Sdic1		*Sdic2*			*Sdic3*			*Sdic4*	*SdicB*	*SdicC*	*sw*	
MiRNA	RA	RC	RA	RB	RC	RE	RF	RG	RE	RA	RA	*sw*-I †	*sw*-II ¦
dme-miR-978-5p	0	0	0	1	1	0	1	1	0	1	1	1	1
dme-miR-4912-5p	0	0	0	1	1	1	1	1	0	1	1	1	1
dme-miR-4961-5p	0	0	1	1	1	0	0	0	0	0	0	1	1
dme-miR-9377-5p	0	0	0	0	1	0	0	1	0	0	0	0	1
dme-miR-4961-3p	2	0	1	1	1	0	0	0	0	0	0	1	1
dme-miR-1-3p	0	0	0	1	1	0	1	1	0	1	0	1	1
dme-miR-4984-3p	1	0	0	1	1	0	0	0	0	0	0	1	1
dme-miR-1012-5p	0	0	0	1	1	0	0	0	0	0	0	1	1
dme-miR-976-3p	1	0	0	0	0	0	0	0	0	0	0	0	0
dme-miR-iab-4-5p	0	0	1	1	1	0	0	0	0	0	0	1	1
dme-miR-992-5p	1	1	0	0	0	0	0	0	0	0	0	0	0
dme-miR-1014-5p	0	0	0	1	1	0	0	0	0	0	0	1	1
dme-miR-4982-5p	1	1	0	0	0	0	0	0	0	0	0	0	0
dme-miR-375-3p	1	1	0	0	0	0	0	0	0	0	0	0	0
dme-miR-964-3p	1	0	0	0	0	0	0	0	0	0	0	0	0
dme-miR-4966-5p	1	0	0	0	0	0	0	0	0	0	0	0	0
dme-miR-9373-5p	1	0	0	0	0	0	0	0	0	0	0	0	0

* Share the seed sequence.

† 3'UTR *sw*-I: *sw*-RA, *sw*-RH, *sw*-RI, *sw*-RJ, *sw*-RK, and *sw*-RM.

¦ 3'UTR *sw*-II: *sw*-RB, *sw*-RC, *sw*-RD, *sw*-RE, *sw*-RF, *sw*-RG, *sw*-RL, *sw*-RN, and *sw*-RO.

Supplementary Table S1.12. Test for differences in productivity per female among female types

	Female Type *						Multiple Pairwise
Timepoint	A^{-d}	B^{+}	E^{-d}	I^{+}	w^{1118}	*P* † (*n*)	Comparison §
Day 1	8.715, (5.550, 11.885)	15.368, (12.200, 18.537)	18.174, (15.000, 21.343)	2.449, (-0.790, 5.687)	16.853, (13.750, 19.959)	<0.0001 (23-25)	E^{-} B^{+} w^{1118} A^{-} I^{+}
Day 3	30.278, (23.381, 37.175)	48.653, (41.756, 55.550)	57.840, (50.943, 64.737)	16.797, (9.752, 23.842)	50.667, (43.909, 57.424)	<0.0001 (23-25)	E^{-} B^{+} w^{1118} A^{-} I^{+}
Day 11	40.467, (29.350, 51.583)	50.356, (39.760, 60.955)	56.210, (45.840, 66.576)	9.949, (-0.420, 20.315)	44.127, (34.180, 54.069)	<0.0001 (20-25)	E^{-} B^{+} w^{1118} A^{-} I^{+}
Day 13	14.142, (8.949, 19.335)	22.689, (17.738, 27.641)	26.044, (21.201, 30.886)	5.688, (0.846, 10.531)	23.873, (19.229, 28.518)	<0.0001 (20-25)	E^{-} B^{+} w^{1118} A^{-} I^{+}
Day 21	16.461, (3.740, 29.184)	29.617, (17.890, 41.347)	20.523, (9.340, 31.707)	6.570, (-5.460, 18.605)	44.449, (33.510, 55.388)	0.0011 (17-23)	w^{1118} B^{+} E^{-} A^{-} I^{+}
Day 23	11.202, (-3.520, 25.923)	21.176, (8.190, 34.158)	15.720, (3.980, 27.463)	4.439, (-8.200, 17.075)	42.603, (30.580, 54.623)	0.0028 (14-22)	w^{1118} B^{+} A^{-} E^{-} I^{+}

* Mean, 95% CI (lower boundary, upper boundary).

† According to Kruskal-Wallis test; sample size range across female types is shown in parenthesis.

§ Female types are ranked from higher to lower median (left to right). Female types not showing statistical significant differences in pairwise Steel-Dwass tests at $P<0.05$ are shown as jointly underlined.

Supplementary Table S1.13. Test for differences in egg hatching rate among female types

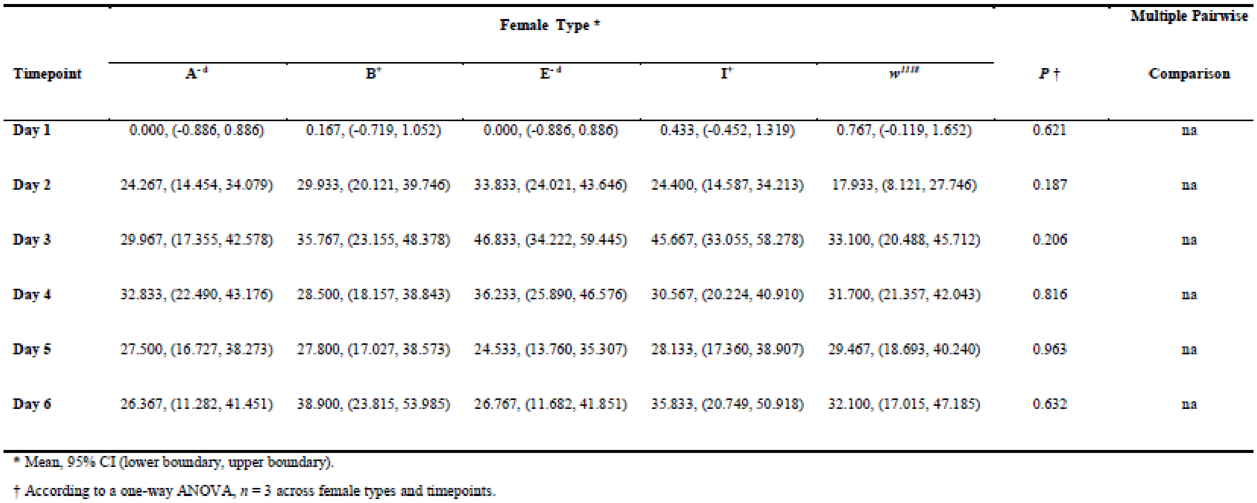
	Female Type *						Multiple Pairwise
Timepoint	A^{-d}	B^{+}	E^{-d}	I^{+}	w^{1118}	*P* †	Comparison §
Day 1	na	0.875, (-0.713, 2.463)	na	0.620	0.830	na	na
Day 2	0.853, (0.623, 1.084)	0.910, (0.802, 1.018)	0.903, (0.889, 0.918)	0.340, (0.274, 0.406)	0.900, (0.810, 0.990)	0.107	na
Day 3	0.860, (0.567, 1.153)	0.927, (0.839, 1.014)	0.850, (0.671, 1.029)	0.380, (0.305, 0.455)	0.803, (0.610, 0.996)	0.062	na
Day 4	0.853, (0.568, 1.139)	0.883, (0.672, 1.095)	0.903, (0.789, 1.018)	0.310, (0.260, 0.360)	0.810, (0.701, 0.918)	0.076	na
Day 5	0.830, (0.722, 0.938)	0.770, (0.124, 1.416)	0.867, (0.495, 1.239)	0.263, (0.169, 0.357)	0.770, (0.680, 0.860)	0.094	na
Day 6	0.903, (0.758, 1.049)	0.893, (0.842, 0.945)	0.857, (0.742, 0.971)	0.267, (0.166, 0.367)	0.787, (0.510, 1.064)	0.048	$\underline{B^{+}\ A^{-}\ E^{-}\ w^{1118}\ I^{+}}$

* Mean, 95% CI (lower boundary, upper boundary).

† According to Kruskal-Wallis test, *n* = 3 across female types and timepoints, except for Day 1 for which the sample size was 0, 2, 0, 1, and 1 respectively.

§ Female types are ranked from higher to lower median (left to right). Female types not showing statistical significant differences in pairwise Steel-Dwass tests at $P<0.05$ are shown as jointly underlined.

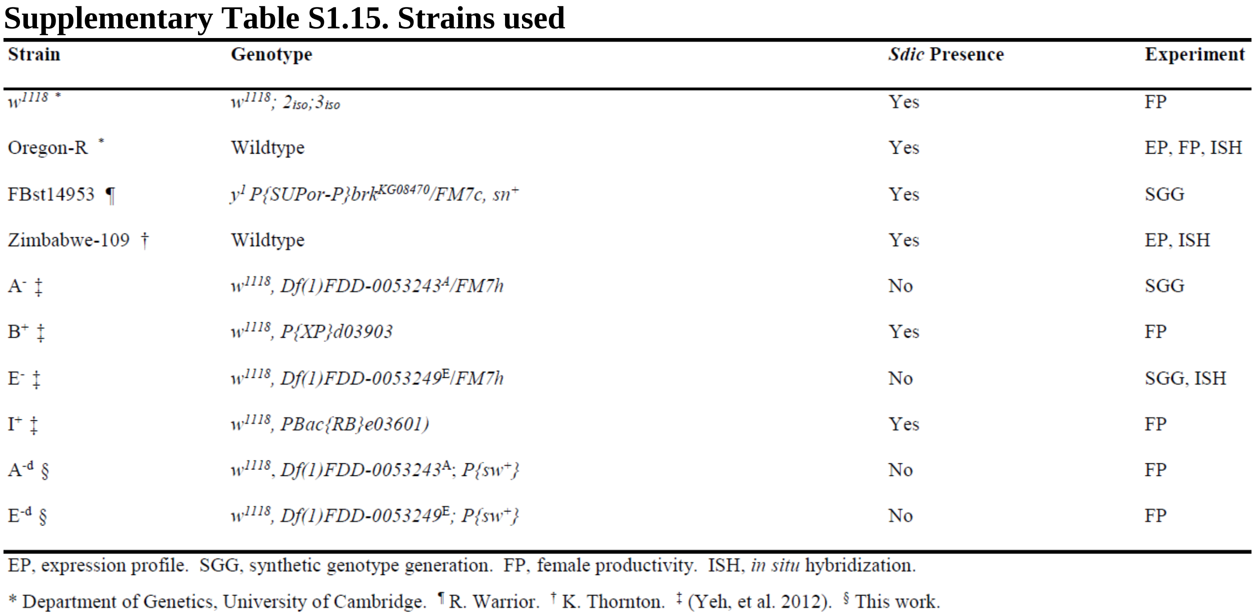
Supplementary Table S1.14. Test for differences in number of eggs per female among female types



	Female Type *						Multiple Pairwise
Timepoint	A^{-d}	B^{+}	E^{-d}	I^{+}	*w^{1118}*	*P* †	Comparison
Day 1	0.000, (-0.886, 0.886)	0.167, (-0.719, 1.052)	0.000, (-0.886, 0.886)	0.433, (-0.452, 1.319)	0.767, (-0.119, 1.652)	0.621	na
Day 2	24.267, (14.454, 34.079)	29.933, (20.121, 39.746)	33.833, (24.021, 43.646)	24.400, (14.587, 34.213)	17.933, (8.121, 27.746)	0.187	na
Day 3	29.967, (17.355, 42.578)	35.767, (23.155, 48.378)	46.833, (34.222, 59.445)	45.667, (33.055, 58.278)	33.100, (20.488, 45.712)	0.206	na
Day 4	32.833, (22.490, 43.176)	28.500, (18.157, 38.843)	36.233, (25.890, 46.576)	30.567, (20.224, 40.910)	31.700, (21.357, 42.043)	0.816	na
Day 5	27.500, (16.727, 38.273)	27.800, (17.027, 38.573)	24.533, (13.760, 35.307)	28.133, (17.360, 38.907)	29.467, (18.693, 40.240)	0.963	na
Day 6	26.367, (11.282, 41.451)	38.900, (23.815, 53.985)	26.767, (11.682, 41.851)	35.833, (20.749, 50.918)	32.100, (17.015, 47.185)	0.632	na

* Mean, 95% CI (lower boundary, upper boundary).

† According to a one-way ANOVA, *n* = 3 across female types and timepoints.



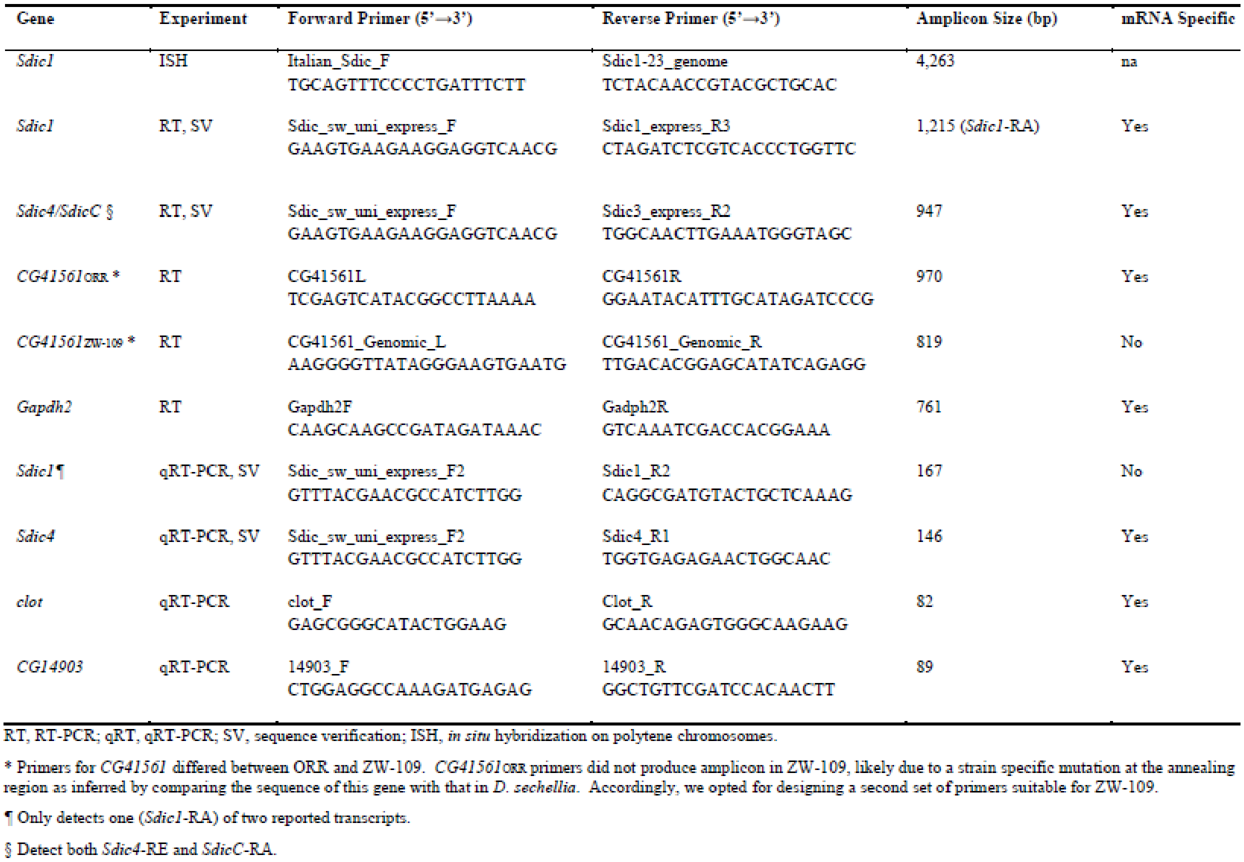
Supplementary Table S1.15. Strains used

Strain	Genotype	*Sdic* Presence	Experiment
w^{1118} *	w^{1118}; 2_{iso};3_{iso}	Yes	FP
Oregon-R *	Wildtype	Yes	EP, FP, ISH
FBst14953 ¶	*y^1 P{SUPor-P}brk^{KG08470}/FM7c, sn$^+$*	Yes	SGG
Zimbabwe-109 †	Wildtype	Yes	EP, ISH
A^- ‡	*w^{1118}, Df(1)FDD-0053243^A/FM7h*	No	SGG
B^+ ‡	*w^{1118}, P{XP}d03903*	Yes	FP
E^- ‡	*w^{1118}, Df(1)FDD-0053249^E/FM7h*	No	SGG, ISH
I^+ ‡	*w^{1118}, PBac{RB}e03601)*	Yes	FP
A^{-d} §	*w^{1118}, Df(1)FDD-0053243^A; P{sw$^+$}*	No	FP
E^{-d} §	*w^{1118}, Df(1)FDD-0053249^E; P{sw$^+$}*	No	FP

EP, expression profile. SGG, synthetic genotype generation. FP, female productivity. ISH, *in situ* hybridization.

* Department of Genetics, University of Cambridge. ¶ R. Warrior. † K. Thornton. ‡ (Yeh, et al. 2012). § This work.

Supplementary Table S1.16. Primers used and amplicons generated



Gene	Experiment	Forward Primer (5’→3’)	Reverse Primer (5’→3’)	Amplicon Size (bp)	mRNA Specific
Sdic1	ISH	Italian_Sdic_F TGCAGTTTCCCCTGATTTCTT	Sdic1-23_genome TCTACAACCGTACGCTGCAC	4,263	na
Sdic1	RT, SV	Sdic_sw_uni_express_F GAAGTGAAGAAGGAGGTCAACG	Sdic1_express_R3 CTAGATCTCGTCACCCTGGTTC	1,215 (*Sdic1*-RA)	Yes
Sdic4/SdicC §	RT, SV	Sdic_sw_uni_express_F GAAGTGAAGAAGGAGGTCAACG	Sdic3_express_R2 TGGCAACTTGAAATGGGTAGC	947	Yes
CG41561$_{ORR}$ *	RT	CG41561L TCGAGTCATACGGCCTTAAAA	CG41561R GGAATACATTTGCATAGATCCCG	970	Yes
CG41561$_{ZW-109}$ *	RT	CG41561_Genomic_L AAGGGGTTATAGGGAAGTGAATG	CG41561_Genomic_R TTGACACGGAGCATATCAGAGG	819	No
Gapdh2	RT	Gapdh2F CAAGCAAGCCGATAGATAAAC	Gadph2R GTCAAATCGACCACGGAAA	761	Yes
Sdic1 ¶	qRT-PCR, SV	Sdic_sw_uni_express_F2 GTTTACGAACGCCATCTTGG	Sdic1_R2 CAGGCGATGTACTGCTCAAAG	167	No
Sdic4	qRT-PCR, SV	Sdic_sw_uni_express_F2 GTTTACGAACGCCATCTTGG	Sdic4_R1 TGGTGAGAGAACTGGCAAC	146	Yes
clot	qRT-PCR	clot_F GAGCGGGCATACTGGAAG	Clot_R GCAACAGAGTGGGCAAGAAG	82	Yes
CG14903	qRT-PCR	14903_F CTGGAGGCCAAAGATGAGAG	14903_R GGCTGTTCGATCCACAACTT	89	Yes

RT, RT-PCR; qRT, qRT-PCR; SV, sequence verification; ISH, *in situ* hybridization on polytene chromosomes.

* Primers for *CG41561* differed between ORR and ZW-109. *CG41561*$_{ORR}$ primers did not produce amplicon in ZW-109, likely due to a strain specific mutation at the annealing region as inferred by comparing the sequence of this gene with that in *D. sechellia*. Accordingly, we opted for designing a second set of primers suitable for ZW-109.

¶ Only detects one (*Sdic1*-RA) of two reported transcripts.

§ Detect both *Sdic4*-RE and *SdicC*-RA.

SUPPLEMENTARY REFERENCES

Benjamini Y, Hochberg Y 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* 57: 289-300.

Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* 33: 623-630.

dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, Emmert DB, Gelbart WM, FlyBase C 2015. FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res* 43: D690-697.

Krsticevic FJ, Schrago CG, Carvalho AB 2015. Long-Read Single Molecule Sequencing To Resolve Tandem Gene Copies: The *Mst77Y* Region on the *Drosophila melanogaster Y* Chromosome. *G3* (Bethesda).

McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, Petrov DA, Fiston-Lavier AS 2014. Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly repetitive transposable elements. *PLoS One* 9: e106689.

Michiels F, Gasch A, Kaltschmidt B, Renkawitz-Pohl R 1989. A 14 bp promoter element directs the testis specificity of the *Drosophila beta 2 tubulin* gene. *Embo J* 8: 1559-1565.

Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S*, et al.* 2015. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res* 43: D213-221.

Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL 1998. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature* 396: 572-575.

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W 2000. PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res* 10: 577-586.

Smyth GK 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3: Article3.

von Grotthuss M, Ashburner M, Ranz JM 2010. Fragile regions and not functional constraints predominate in shaping gene organization in the genus *Drosophila*. *Genome Res* 20: 1084-1096.

Yeh SD, Do T, Chan C, Cordova A, Carranza F, Yamamoto EA, Abbassi M, Gandasetiawan KA, Librado P, Damia E*, et al.* 2012. Functional evidence that a recently evolved *Drosophila* sperm-specific gene boosts sperm competition. *Proc Natl Acad Sci U S A* 109: 2043-2048.

CHAPTER 2

Understanding the early evolutionary stages of a tandem *Drosophila melanogaster*-specific gene family: A structural and functional population study

ABSTRACT

Gene families underlie genetic innovation and phenotypic diversification. However, our understanding of the early genomic and functional evolution of tandemly arranged gene families remains incomplete as paralog sequence similarity hinders their accurate characterization. The *Drosophila melanogaster*-specific gene family *Sdic* is tandemly repeated and impacts sperm competition. We scrutinized *Sdic* in 20 geographically diverse populations using reference-quality genome assemblies, read-depth methodologies, and qPCR, finding that ~90% of the individuals harbor 3–7 copies as well as evidence of population differentiation. In strains with reliable gene annotations, copy number variation (CNV) and differential transposable element insertions distinguish one structurally distinct version of the *Sdic* region per strain. All 31 annotated copies featured protein-coding potential and, based on the protein variant encoded, were categorized into 13 paratypes differing in their 3' ends, with 3–5 paratypes coexisting in any strain examined. Despite widespread gene conversion, the only copy present in all strains has functionally diverged at both coding and regulatory levels under positive selection. Contrary to artificial tandem duplications of the *Sdic* region that resulted in increased male expression, CNV in cosmopolitan strains did not correlate with expression levels, likely as a result of differential genome modifier composition. Duplicating the region did not enhance sperm competitiveness, suggesting a fitness cost at high expression levels or a plateau effect. Beyond facilitating a minimally optimal expression level, *Sdic* CNV acts as a catalyst of protein and regulatory diversity, showcasing a possible evolutionary path recently formed tandem multigene families can follow toward long-term consolidation in eukaryotic genomes.

INTRODUCTION

Structural variants have been largely overlooked in genetic variation surveys, limiting our understanding on the genetic basis of phenotypic change (Feyereisen et al. 2015; Huddleston and Eichler 2016; Chakraborty et al. 2019). Structural variants include >50-nt-long duplications and deletions, transpositions, inversions, and translocations. Complex genomic regions, those that exhibit unusually high levels of structural variation often in the form multiple copies of particular, high identity sequences generated by some kind of duplicative mechanism, are predominantly affected by this oversight. Accordingly, these regions are often grossly misassembled or absent altogether in reference genome assemblies (Hollox 2012; Ranz and Clifton 2019). This in turn precludes their accurate genomic and functional characterization, which is relevant given the close interplay between these regions, evolutionary change, and disease (Dennis and Eichler 2016). This interplay arises from the proclivity of complex genomic regions to structural remodeling (Hurles 2004; Hollox 2012), often resulting in marked copy number variation (CNV) patterns for the encompassed genes (Sudmant et al. 2010; Jiang et al. 2012; Carpenter et al. 2015) and in the formation of new gene entities with chimeric or defective features (Dennis et al. 2012; Nuttle et al. 2016; Fiddes et al. 2018). Despite the potential of these genomic regions to impact the phenotype and organismal fitness (Hollox 2008; Jugulam et al. 2014; Chakraborty et al. 2019), our understanding of how they evolve remains largely incomplete.

To date, most complex genomic regions characterized molecularly have been linked to traits associated with viability and fecundity (Dennis et al. 2017; Chakraborty et al. 2019) as opposed with reproductive success, that is, to traits targeted by sexual selection rather than by natural selection (Darwin 1871). A form of sexual selection, sperm competition, biases fertilization at the postcopulatory level in numerous species groups (Parker 1970; Birkhead 1998). Among the few genetic factors known to affect sperm competition (Civetta and Ranz 2019), there is one that resides within a complex region of the *Drosophila melanogaster* euchromatin: the tandem multigene family *Sdic*. *Sdic* is absent in the rest of the genus *Drosophila*, having originated at some point in the *D. melanogaster* lineage after diverging from the *simulans* clade ~1.4 Ma (Nurminsky et al. 1998; Obbard et al. 2012).

The original *Sdic* gene resulted from a segmental duplication on the *X* chromosome spanning two adjacent genes, *sw* and *AnxB10*, which fused through a set of deletions while accommodating multiple nucleotide substitutions. Subsequently, this chimeric entity underwent a tandem expansion (Nurminsky et al. 1998). The repetitive nature of *Sdic* and the high sequence similarity among the resident paralogs make this region prone to recurrent nonallelic homologous recombination (NAHR) events, that is, unequal crossing over, which should result in contractions and expansions of the tandem array (Hastings et al. 2009). Thus, the organization of the *Sdic* region in the *D. melanogaster* reference strain, which includes six copies of a repeat unit, spanning in total ~46 kb (Clifton et al. 2017), might just be a nonrepresentative state within the actual breadth in copy number (CN) in natural populations. In fact, the CN distribution at the *Sdic* region is unknown, as are the occurrence of other structural changes (e.g., transposable element—TE—insertions) and the frequency of structurally distinct versions of the region. Also unknown is the extent to which *Sdic* CNV can impact expression levels, as often assumed after tandem duplication events (Kondrashov et al. 2002; Kondrashov 2010), or can act as a catalyst for protein diversification (Traherne et al. 2010), or both. In fact, without this crucial information, it is not feasible to determine whether putative expression changes mirroring alterations in *Sdic* CN actually impact sperm competitive ability. Further, it is unclear whether the patterns of gene conversion and overall sequence conservation documented across the *Sdic* copies in the reference strain hold in strains representing other populations of *D. melanogaster*. Overall, *Sdic*, offers the opportunity to investigate different levels of change and their consequences at the early stages of a recently expanded multigene family, which has been typically neglected despite its importance to understand the fate of gene duplicates and the origin of new gene functions (Kondrashov 2010; Katju and Bergthorsson 2013; Long et al. 2013; Cardoso-Moreira et al. 2016; Naseeb et al. 2017; Rogers et al. 2017).

We have analyzed the *Sdic* region at the genetic, functional, and phenotypic levels using two panels of strains with diverse geographical origin, including the ancestral sub-Saharan distribution range of *D. melanogaster* (Begun and Aquadro 1993), and other synthetic strains harboring complete duplications of the *Sdic* region. We aim at: 1) gauging the breadth of *Sdic* CNV in different parts of the world using the annotation of the region in reference-quality genome assemblies, qPCR assays, and read-depth algorithms suitable for analyzing Illumina sequencing reads; 2) evaluating the role of positive selection in explaining the sequence evolution at the coding and noncoding levels of this tandemly arranged multigene family, as well as the relevance of gene conversion; 3) determining by qRT–PCR assays the extent to which CNV translates into expression variation in natural populations and genome-edited strains that allow control of genomic background differences; and 4) testing whether increased *Sdic* expression correlates with varying sperm competitive ability using different genetic modifications of the *Sdic* region. While answering some of these questions, we also found that a fraction of reference-quality assemblies generated with single-molecule real-time (SMRT) and Nanopore sequencing technologies still do not faithfully recapitulate the organization of the *Sdic* region.

RESULTS

Naturally Occurring CNV in the *Sdic* Region

To generate a global portrait of *Sdic* CNV in *D. melanogaster*, we examined two different panels of strains. First, we focused on a panel of 15 strains (eight from the Americas; two from Africa; and five from Eurasia and the Middle East; supplementary tables S2.1 and S2.2, Supplementary Material online) for which female-derived reference-quality assemblies have been generated (Chakraborty et al. 2018, 2019). These assemblies offer the opportunity to parse patterns of additional structural variation, including inversions and TE insertions, in addition to calibrate two other approaches to estimate CNV: qPCR and read-depth analysis. Second, using read-depth analysis, we extended our characterization of *Sdic* CNV to a panel that includes strains from populations derived from five different locations around the globe in order to estimate population parameters that can help uncover *Sdic*'s evolutionary mode of structural remodeling across *D. melanogaster*'s entire range.

Individual *D. melanogaster* Populations Consist of Various Numbers of *Sdic* Copies

We annotated the *Sdic* region in 14 de novo, reference-quality genome assemblies scaffolded with SMRT sequencing reads (Chakraborty et al. 2018, 2019). Thirteen of them correspond to strains from the *Drosophila* Synthetic Population Resource (DSPR) and are virtually isogenic (King, Merkes, et al. 2012); the 14th strain is the commonly used laboratory, wild-type stock OR-R. The structural and sequence features of the region were compared across assemblies against its updated reconstruction in the ISO-1 reference strain, which is based on the sequence of the GCA_000778455 assembly (Berlin et al. 2015) as opposed to that of the Release 6 (dos Santos et al. 2015), as the former more accurately recapitulates the *Sdic* region (Clifton et al. 2017). This prevents inaccurate inferences about the type and magnitude of genetic differences across the strains considered (Supplementary Text).

Upon annotating the *Sdic* region in these 14 assemblies (fig. 2.1 and supplementary fig. S2.1, Supplementary Material online), we found that all assemblies but three (A2, A6, and B4; supplementary text and supplementary fig. S2.2, Supplementary Material online) show the *Sdic* region unfragmented and flanked by the same genes as in the reference strain, that is, *sw* upstream and *AnxB10* downstream, occupying a proximal position relative to the centromere. All copies of the *Sdic* repeat examined were essentially the same length within and across assemblies. Excluding two unreliable assemblies (A2 and A6) for the *Sdic* region, only those from Cape Town (B2) and Riverside (B4) harbor six copies as in the reference strain (Berlin et al. 2015; Clifton et al. 2017). Overall, we observed a noticeable breadth in CN with a coefficient of variation of 26.8% (n = 12; 4.25 ± 1.14, avg ± SD; 4, median). This CNV contributed to size differences in the *Sdic* region, which ranges from ~34 kb (Canton-S, A1) to ~57 kb (Cape Town, B2) (supplementary table S2.2, Supplementary Material online).

CN Estimates from Gene Annotation Are Only Partially Validated

We attempted to validate the CN estimates obtained from annotating the *Sdic* region in reference-quality assemblies both computationally and experimentally. In the first case, we performed read-depth analyses using CNVnator (Abyzov et al. 2011), which was optimized for the special features of the *Sdic* region (fig. 2.2A; Materials and Methods and Supplementary Text). The final analyses were done using synthetic reference genomes derived from A4 and ISO-1 separately, showing a high degree of agreement between the average read-depth estimates from both analyses (fig. 2.2B). These synthetic genomes contain only one single repeat of *Sdic* and lack the parental genes, removing redundancy across the *Sdic* region. Overall, we found a 50% (i.e., seven out of 14 strains) discrepancy rate between the estimates obtained with CNVnator and those from genome annotation (fig. 2.2C and supplementary tables S2.2 and S2.3, Supplementary Material online).

We additionally estimated *Sdic* CN using qPCR. Given the structural relationship between *Sdic* and its parental gene sw, we estimated *Sdic* CN as the difference between the CN inferred from an amplicon associated with both *sw* and *Sdic*, and another amplicon specific to *sw* (fig. 2.3A; Materials and Methods and supplementary table S2.4, Supplementary Material online). We first calibrated our ability to discern CN differences across a set of genotypes that correspond with particular strains and their progenies with known CNs for *Sdic* and *sw*. Specifically, we used w^{1118}, an isogenic strain used to engineer structural variants (Parks et al. 2004), a set of derivative engineered genotypes carrying either the full deletion (Yeh et al. 2012; Clifton et al. 2017) or the duplication in tandem (this work; supplementary fig. S2.3, Supplementary Material online) of the *Sdic* region, and the progeny from reciprocal matings involving some of these strains (fig. 2.3B). The results strongly supported our ability to correctly infer the number of *Sdic* copies using qPCR assays (Supplementary Text), which were extended to 12 strains belonging to the DSPR panel and OR-R (AB8 was unavailable). In total, 24 genotypes were examined (supplementary table S2.2, Supplementary Material online and fig. 2.3C and D). The comparison of the qPCR and gene annotation estimates showed that they were coincidental for only ~50% (7/13) of the strains.

Conversely, the comparison of the rounded-off CN values obtained by read-depth analysis estimates and qPCR assays showed a perfect agreement (fig. 2.3E and supplementary table S2.5, Supplementary Material online). Using the CNVnator estimates, as they include one more strain than those from qPCR, we noticed that the discrepancies did not follow a consistent trend, that is, CNVnator estimates were in five cases higher and in two cases lower than those from the genome annotation analysis. The three approaches show complete agreement for only seven out of 13 strains investigated (A4, A5, A7, B1, B2, B3, and B6). This, combined with the findings noted above for several assemblies, points to the estimates from the genome annotation analysis as the least reliable. This could presumably result from artifactually collapsing or adding copies while assembling the *Sdic* region, offering a cautionary note to solely depending on reference-quality assemblies when characterizing structural variation in complex regions. Overall, the CNVnator and qPCR estimates confirm that the *Sdic* region has undergone extensive structural remodeling (for CNVnator, n = 14 strains; copies = 4.86 ± 0.95, avg ± SD; CV = 19.54%), harboring four structurally distinct alleles based on CN alone, and showing similar copy range (3–6) across different continental regions (supplementary table S2.2, Supplementary Material online).

SMRT-Based Assembly Properties Affect Accurate Region Recapitulation

To determine what factors affect the inaccurate recapitulation of the *Sdic* region in some assemblies scaffolded with SMRT sequencing reads, we performed a multiple logistic regression to precisely evaluate the predictive power of different assembly metrics when used genome-wide, including sequence coverage, assembly N50 (Earl et al. 2011), and NR50—the median read length above which half of the total coverage is contained (Chakraborty et al. 2018). None of the assembly metrics evaluated turned out to be a good predictor of a faithful recapitulation of the *Sdic* region (supplementary table S2.6, Supplementary Material online). Subsequently, as assembly metrics fluctuate locally, we focused on the individual reads related to the *Sdic* region, recalculating both coverage and NR50 and adding a few other metrics such as the interpolated size of the region based on CN as estimated with CNVnator. Across strains, the number of reads related to the *Sdic* region was 134 ± 56.8 (avg ± SD), with the maximum and minimum number of reads being 275 (A4) and 53 (A6), respectively (supplementary table S2.7, Supplementary Material online). We found no strain for which there was at least one sequencing read spanning from *sw* to *AnxB10*. The

A4 strain stood out showing the second-highest local NR50 (17.9 kb) and the highest local coverage (~93x), confirming not only that it is arguably the best assembly of the euchromatin of *D. melanogaster* (Chakraborty et al. 2018) but also in relation to complex regions like *Sdic*. When the metrics were restricted to the *Sdic* region, the multiple logistic regression analysis found that the local coverage has a significant predictive power (*P* = 0.0057), with a higher local coverage increasing the likelihood of faithfully recapitulating a complex region like *Sdic*. For the seven reliable assemblies within the DSPR panel, the minimum local coverage was ~29x (B3), with their average coverage being significantly higher than that of the unreliable assembly (~39x vs. ~27x, respectively; Kruskal–Wallis, *P* = 0.015).

Global Molecular Diversity Patterns in the *Sdic* Region

The *Sdic* Region Is Polymorphic for Structurally Distinct Alleles around the World

Each population included in the DSPR panel and OR-R is derived from a single individual, which prevents an accurate inference of the level of polymorphism and population differentiation, if any, for the *Sdic* region at the structural level. To circumvent this limitation, we used CNVnator on a second panel of isogenic lines, the Global Diversity Lines, derived from five collection sites: Beijing, Ithaca, the Netherlands, Tasmania, and Zimbabwe (Grenier et al. 2015). None of the 70 individuals ultimately considered lacked *Sdic* and 39% featured CNs outside the range seen in the DSPR panel. More importantly, we found up to seven structurally distinct alleles based on variable CN (4–10 copies), with no more than five of these alleles in any given population (minimum = 3; Beijing; maximum = 5, Ithaca, the Netherlands, and Zimbabwe) (fig. 2.2D and supplementary table S2.8, Supplementary Material online). In all populations, there are at least three structurally distinct alleles at a frequency >5%.

Using the V_{ST} statistic (Redon et al. 2006), we found that population differentiation in the *Sdic* region is greater than expected by chance alone (V_{ST} = 0.1714, P = 0.0023; 10,000 Monte Carlo simulations). Subsequent global and pairwise nonparametric tests showed that the Beijing population features significantly lower CNs than the Zimbabwe and Ithaca populations (supplementary table S2.9, Supplementary Material online). In fact, the two latter populations exhibit the highest frequencies of structurally distinct alleles carrying the maximum CNs documented (9 and 10). An additional analysis of a third panel of strains from Zambia, each strain corresponding to a different haploid embryo genome, allowed us to zoom in on a different location of *D. melanogaster*'s ancestral distribution range (Lack et al. 2016), extending the detection of additional structural distinct alleles beyond those present in DSPR and GDL individuals; two embryos were found to carry two copies and one with 12 (supplementary fig. S2.4 and table S2.8, Supplementary Material online).

TE Insertions Contribute to *Sdic* Structural Variation

We looked for additional structural variants in the assembly of the seven most reliable strains of the DSPR panel for the *Sdic* region. In all strains, the copies are tandemly oriented head-to-tail, consistent with the absence of inversions. Nevertheless, we found three population-specific TE insertions (fig. 2.1), none of them presumably compromising the protein-coding potential of the copies (supplementary table S2.10, Supplementary Material online). Considering differences in CN and TE insertions, we find that each population in this subset of strains harbors a structurally distinct version of the *Sdic* region.

Sdic Copy Differentiation Affects the Carboxyl End of *Sdic* Protein Variants

The most reliable subset of strains harbors 31 *Sdic* copies. Consistent with the age of the region and the occurrence of NAHR and gene conversion events (see below), the level of nucleotide differentiation is very limited among copies both within and across strains (supplementary table S2.11, Supplementary Material online). This observation holds not only for the *Sdic* transcriptional unit but also for the upstream noncoding interval present at each repeat, including the presumed pseudogene *AnxB10*-like for which we did not find evidence of expression (Materials and Methods; supplementary text and supplementary table S2.12, Supplementary Material online). Importantly, a given *Sdic* allele can occupy different physical locations within the tandem array across strains and be present as several copies in the same strain. We refer to these *Sdic* alleles as paratypes (Fiddes et al. 2018). Based on particular combinations of diagnostic amino acid motifs spanning _5 residues in the presumably encoded products, the copies were categorized into one out of 13 paratypes (a–m; fig. 2.1), adding eight new distinct protein variants to the pool of five previously identified paratypes (Clifton et al. 2017). Like in the ISO-1 strain, the new paratypes show notable differences at the level of length and actual amino acid sequence of the carboxyl-terminus (supplementary table S2.13, Supplementary Material online), which is due to the preferential location of nucleotide differences in the two exons most proximal to the STOP codon (Clifton et al. 2017). Despite length differences, all copies considered presumably encode proteins with 4–7 WD40 motifs, as seen in the ISO-1 strain (Ma et al. 2019). Further, only one paratype, *e*, is found in all strains, and always present as a single copy and adjacent to the parental gene *AnxB10* (fig. 2.1). The global paratype diversity generated within the *Sdic* region is reflected in the presence of six paratypes as a single copy in the one strain in which they reside (fig. 2.4A), in the fact that each strain harbors 3–5 paratypes (3.86 ± 0.90; mean ± SD; fig. 2.4B), and in that three strains (A5, A7, and B6) carry each resident paratype as a single copy. Overall, the similarity between populations based on CN and paratype composition reflects neither phylogenetic relationship nor geographic proximity (supplementary fig. S2.5, Supplementary Material online).

A Common Landscape of Gene Conversion across Strains

To assess the role of gene conversion in shaping the region's sequence evolution, and whether its mode of action and magnitude differed among strains, we identified tracts of gene conversion (Sawyer 1989). Gene conversion is rampant across strains, with paratype *e* and *sw* dominating the landscape of events as they contribute to 61% of all detected ones (fig. 2.4C and supplementary fig. S2.6; table S2.14, Supplementary Material online). In addition, gene conversion events exhibit common topological patterns along the *Sdic* repeat in all strains, showing a good agreement between boundaries of gene conversion tracts predicted by GeneConv and recombination breakpoints inferred with ACG (O'Fallon 2013) (supplementary figs. S2.7 and S2.8, Supplementary Material online).

This gene conversion landscape supports a different chronology for the formation of the *Sdic* multigene family from that proposed based on the ISO1 strain alone (Clifton et al. 2017). In an ancestor of the strains examined, an early *Sdic* copy would have engaged in gene conversion events with the most proximal third of the length of *sw* to its 3' end. At some point, this early copy duplicated. The paralog adjacent to *sw* continued exchanging DNA tracts with sw, whereas the paralog adjacent to *AnxB10* gave rise to paratype *e*. This new cluster configuration likely favored gene conversion between both *Sdic* paralogs, at their 2.3–7.2 kb interval. This, however, limited exchange between *sw* and paratype *e*, possibly owing to their more distant positioning, separated by an intervening copy. Escaping gene conversion events with *sw* permitted paratype *e* to accumulate sequence differences at its 3' end, a region that evolves under positive selection (Clifton et al. [2017] and below). This scenario is compatible with alternative phylogenetic reconstructions in which all paratype *e* copies from the different strains always conform to a well-supported monophyletic clade, basal to the remaining paratypes (supplementary fig. S2.9, Supplementary Material online). The branch leading to this clade is comparatively long, despite rampant levels of gene conversion involving paratype *e*, in line with fixed differences at its 3' region. Additional paratypes would have been formed and eliminated afterward, resulting into a floating set of additional *Sdic* copies, whose divergence would have been confined to sections of the most 3' third of the *Sdic* transcriptional unit. These additional copies might still be engaged in gene conversion events with the central sequence interval of paratype *e*, limiting further differentiation for that part of the repeat.

Positive Selection in Coding and Noncoding Sequences of the *Sdic* Repeat

The common positional patterns among predicted gene conversion boundaries and recombination breakpoints across the length of the *Sdic* repeat and strains prompted us to assess the impact of positive selection separately for each partition. Overall, we find strong evidence for the action of purifying selection but for the coding fraction of the *Sdic* transcriptional unit, we detect an unequivocal signal of positive selection in subpartition P6.1 (supplementary fig. S2.8, Supplementary Material online), which encodes part of the carboxyl-termini of the *Sdic* protein (P_{adj} = 0.012). In this region, the basal lineage leading to the ancestor of eight nearly identical copies (one copy per strain, corresponding to paratype *e*), accumulates nonsynonymous changes faster than expected under neutrality. We also identified various lineages in the *Sdic* family tree showing statistical evidence for positive selection in multiple partitions (P1, P3, P5, P6), many of them encompassing noncoding sites (in both internal branches and tips; supplementary table S2.15, Supplementary Material online). These results are consistent with positive selection playing a major role in driving not only the evolution of the 3'-UTR of the ancestral *Sdic* copy and of the copies that form the diverged clade that corresponds to paratype *e* but also of a fraction of the noncoding sequence elsewhere in the *Sdic* repeat. The 3’-UTRs of the *Sdic* copies in the ISO-1, particularly that of *Sdic1* (paratype *e*), were previously shown to have been extensively remodeled in their miRNA binding site composition relative to *sw* (Clifton et al. 2017).

Sdic Global Expression Level Does Not Correlate with CNV

Complete gene duplications, that is, those including regulatory sequences, are thought to result in additive changes in transcript abundance that have the potential of affecting organismal fitness (Kondrashov et al. 2002; Kondrashov 2010). To test whether a higher *Sdic* CN actually results in a higher expression level, we estimated the aggregate expression from all *Sdic* copies in males, the sex in which *Sdic* exhibits preferential expression (Clifton et al. 2017). Using qRT–PCR, and with ISO-1 as a reference, we surveyed *Sdic* expression levels across the five strains from the DSPR panel for which there was no discrepancy across methodologies to estimate CN (supplementary table S2.2, Supplementary Material online) and OR-R, spanning the observed CN range, that is, 3–6 (fig. 2.5A and B). Although we found global differences in expression levels (one-way ANOVA, $F = 9.99$, df = 6, $P < 0.0001$; supplementary table S2.16, Supplementary Material online), there is limited evidence of significantly different expression across pairwise comparisons mirroring the direction of the differences in CN between strains. Seven of the 21 pairwise comparisons entail a statistically significant alteration in expression ($P < 0.05$, Tukey–Kramer HSD post hoc test; supplementary table S2.16, Supplementary Material online), with only four of those comparisons agreeing with the CN differences. For example, strain A7, which harbors four *Sdic* copies, exhibits the lowest *Sdic* expression, being significantly different from B3 (also harboring four copies), A4 (five copies), and OR-R (six copies), but not from B2 (six copies) and B6 (three copies). Relative to the reference strain ISO-1, only three of the six strains surveyed showed significantly different expression (A7, four copies; B2, six copies; and B3, four copies), being lower in all cases. The largest difference in transcript abundance is found between strains with identical CN, B3, and A7 (~97% more transcript in the former). Overall, we found no evidence of a positive association between transcript abundance and CN in natural populations ($r^2 = 0.06$, $P > 0.05$; fig. 2.5C).

This substantial decoupling between CN and transcript level could result from buffering mechanisms acting in the face of excessive CN, such as negative feedback loops and access limitations to transcriptional factories in the nucleus (Harewood et al. 2012; Rogers et al. 2017), and from differential composition of expression modifiers acting in *cis*- and *trans*- across populations. To help clarify this extent, we surveyed *Sdic* expression levels in w^{1118} and its two derivative engineered genotypes carrying a duplication of the *Sdic* region, thus evaluating the impact on gene expression solely resulting from CN differences, without any confounding effect arising from differences in genomic background. Reminiscent of findings with tandemly arranged duplicate pairs of the *D. melanogaster* gene *Adh* (Loehlin and Carroll 2016), we found that duplicating the *Sdic* region in the same genetic background results in statistically significant increases in expression beyond a mere 2-fold change, that is, 100% more: 2T, 158% more; 4M, 209% more (one-way ANOVA, $F = 61.73$, df = 3, $P < 0.0001$; fig. 2.5C and supplementary table S2.16, Supplementary Material online). This result suggests that within-strain buffering mechanisms have very little effect on aggregate *Sdic* male expression, and therefore the interplay between *Sdic* CN and expression level in natural populations is primarily shaped by regulatory variants.

More Functional *Sdic* Copies Do Not Result in Increased Sperm Competitive Ability

When considering the 146 individuals or haploid embryos genotyped for CN using CNVnator, ~91% of them show within three and seven copies, with decreasing frequencies for CN values outside this range (fig. 2.2E). Given the advantageous effect that *Sdic* confers to males in sperm competition (Yeh et al. 2012), it is not apparent why there are not more individuals carrying higher CNs. Accordingly, we tested whether a substantial increase in CN enhances sperm competitive ability by testing differences for this trait among males carrying the wild-type-like version of this region, its deletion, or its duplication, in all cases in w^{1118} background.

In phenotypic tests performed to detect differences in sperm competitive ability between competing males by tracking the fraction of the progeny fathered by different males that have mated with the same female, males carrying the duplication of the *Sdic* region did not exhibit a significantly higher sperm competitive ability (fig. 2.6). Although there is no perfect consistency in the performance shown by the males of the two duplication-bearing strains, having twice as many copies of *Sdic* as in w^{1118} decreases sperm competitive ability to the same extent as if no *Sdic* copy is present in the genome (4M vs. E^{-}) or does not differ from carrying the default CN in the w^{1118} background (2T vs. B^{+} and w^{1118}) (supplementary table S2.17, Supplementary Material online).

DISCUSSION

We have generated a detailed portrait of the organization and patterns of intraspecific genetic and functional variation of arguably one of the most recently formed and structurally complex regions in the *D. melanogaster* euchromatin. We find compelling evidence that the *Sdic* region has undergone extensive structural remodeling in natural populations from very diverse geographical origins. Its inherent properties, that is, multiple copies of high sequence identity in the same orientation, and other genomic features can explain the susceptibility of this region to remodeling. For example, close proximity to replication origins has been shown to be related to CNV (Lee et al. 2007; Langley et al. 2012). Interestingly, two origins of replication have been annotated at the 5’ end of *AnxB10* and *sw*, respectively (Eaton et al. 2011). Further, *Sdic* adds to the limited list of NAHR hotspots whose evolutionary dynamics is likely to be influenced by sexual selection, although in this case at the post- rather than premating level (Karn and Laukaitis 2009; Pezer et al. 2015; Pezer et al. 2017).

For a subset of seven cosmopolitan populations from one of the panels analyzed, for which genetic changes could be tracked both at the sequence and structural levels, we found one structurally distinct version of the region per population. This level of variation results from both changes in CN and recent TE insertions. Further, the breadth of CNV was evaluated in six populations from different continents, two of them corresponding to different locations within the presumed ancestral range of *D. melanogaster* (Begun and Aquadro 1993). The extensive degree of CN polymorphism found in these two populations is compatible with a scenario in which the ancestral population that migrated into Eurasia from Africa _10,000 years ago (Li and Stephan 2006; Stephan and Li 2007) was polymorphic for *Sdic* CN. Additionally, we observed that many of the structurally distinct alleles based on CN are shared across the populations from the GDL panel, although there is evidence of statistically significant population differentiation involving the Zimbabwe and Beijing populations. This last pattern mirrors previous inferences based on genome-wide SNP data analysis (Grenier et al. 2015).

The frequency distribution for *Sdic* CN in natural populations is far from that expected under a runaway amplification process in which additional functional copies would be correlated with higher expression, ultimately having a directional effect on the phenotype (Brown et al. 1998; Schmidt et al. 2010; Soh et al. 2014). In contrast, we found that intermediate CN values are prevalent, that differences in the aggregate transcript abundance are not correlated with CNV in a geographically diverse set of strains, and that significantly increased *Sdic* expression as a result of artificially doubling CN does not result in enhanced sperm competitive ability based on progeny contribution in double-mating assays. The prevalence of individuals bearing intermediate CN values could result from a scenario of stabilizing selection, or from a mutation-drift equilibrium coupled with the action of purifying selection sculpting the range boundaries as proposed for some multigene families in mammals (Hollox 2008; Teitz et al. 2018).

In relation to *Sdic* expression levels, the lack of correlation between CN and transcript abundance is in line with previous reports in other *Drosophila* species, rat, and in peach-potato aphids (Field et al. 1999; Guryev et al. 2008; Rogers et al. 2017), but it is at odds with a general trend previously reported in *D. melanogaster* (Cardoso-Moreira et al. 2016). At least in relation to the upper end of transcription, buffering mechanisms do not seem to be a good explanation as shown by the enhanced expression documented in our engineered duplications of the *Sdic* region. Alternatively, expression modifiers present in different genomic backgrounds could explain the lack of correlation documented. Such modifiers include regulatory variants in *cis* and *trans* (Lemos et al. 2008; Catalan et al. 2016), as well as alterations of copy functionality by TE insertions or premature termination codons that activate the nonsense-mediated decay pathway (Hug et al. 2016; Scott et al. 2016). Based on sequence analyses in the strains examined, we do not observe overt mutations that could damage promoter activity nor evidence of disruptive mutations that could compromise transcript stability in the reliably annotated *Sdic* copies. Overall, our results suggest that the across population variation in aggregate male gene expression level for the *Sdic* multigene family is not as much influenced by CN as by population differences in regulatory input, possibly in *trans*.

As for the lack of association between enhanced *Sdic* expression through increased CN and sperm competitive ability, it is not immediately apparent what is the cause. First, the boosting effect of *Sdic* on sperm competitive ability (Yeh et al. 2012) might plateau beyond an unknown threshold expression level. Second, an increased CN might result in enhanced sperm competitive ability, but this beneficial effect is offset by detrimental effects that reduce the viability of the progeny carrying the duplication of *Sdic*. This second scenario is feasible as in the double-mating assays performed, differential sperm competitive ability is inferred through differential progeny contribution between competing males carrying different CN when they are second to mate (P2) rather than by a more reliable method based on the direct observation of the sperm from those genotypically different males in the female reproductive tract (Jayaswal et al. 2018). This would result in no significantly different P2 values between males carrying 6 and 12 *Sdic* copies even though there were true differences in sperm displacement (Civetta and Ranz 2019). Further, reduced progeny viability can be related to increased expression above a threshold, which is conceivable in the case of *Sdic* as it is expressed in somatic tissues of both genders, having the potential to affect other traits beyond sperm competition (Clifton et al. 2017). The nature of this detrimental effect could take place directly by triggering molecular imbalance, energetic waste, or titrating out limiting factors such as RNA polymerases and ribosomes (Rice and McLysaght 2017), or indirectly through an excessive downregulation of the parental and dosage-dependent gene sw, as *Sdic* can presumably compete with it in the context of the interactions that *sw* establishes with several protein complexes (Boylan et al. 2000; Boylan and Hays 2002). Alternatively, a putatively reduced progeny viability might be unrelated to an increased expression and instead be linked to an enhanced genome instability with higher CN (Didion et al. 2015; Fouche et al. 2018). More refined assays and functional tests should help support or refute these possibilities. At this point, we are only certain of a boosting effect on sperm competitive ability when *Sdic* is expressed in males with six copies relative to males lacking *Sdic* (Yeh et al. 2012), an effect that is not detectable when this CN doubles. Only by testing additional intermediate CN values it will be clearer the fitness-dosage interplay in the case of *Sdic* (Kondrashov 2010).

In contrast to the relatively constrained range of CN and lack of correlation between transcript abundance and CN in natural populations, the *Sdic* region shows a remarkable capability to generate protein diversity in each strain that could be reliably analyzed. We found extensive paratype breadth primarily associated with distinct 3' carboxyl ends, no evidence of a particular paratype being preeminent in CN within any given strain, and only one of the 13 paratypes—paratype *e*—being present in all strains. This paratype shows strong evidence of having evolved under positive selection both at coding and noncoding levels. Further, this paratype diversity has accumulated despite profuse gene conversion events. The topology of the gene conversion landscape shows extensive commonalities across strains, with the fixed paratype *e* and the parental gene *sw* being major mutually exclusive contributors along the *Sdic* repeat. As these patterns have been documented in cosmopolitan strains, it will be interesting to determine whether they hold in strains from the ancestral range of *D. melanogaster*.

Collectively, our results suggest that *Sdic* CNV in contemporary populations of *D. melanogaster* secures a minimal necessary expression level across different genomic backgrounds and sexual selection regimes, serving also as a substrate to prevent nucleotide change via gene conversion and NAHR events for essentially all the *Sdic* repeat but the two most 3' exons and the 3'-UTR of *Sdic* copies (Rozen et al. 2003; Teitz et al. 2018). Equally important, maintaining multiple copies that encode different and possibly fully functional paratypes is compatible with a mechanism that safeguards functional diversity at the protein level (Traherne et al. 2010) while enabling expression profile diversification. *Sdic* copies in conventional laboratory strains show evidence of expression divergence across life stages and anatomical parts of the adult (Clifton et al. 2017), which is concurrent with profound 3'-UTR remodeling. At least for the copies associated with paratype *e*, we find evidence of positive selection acting on this portion of the *Sdic* repeat. An equivalent pattern could be taking place for copies of the same paratype but in different populations. Functional characterization of a set of strains with different CN and paratype composition can be highly informative relative to the extent of evolutionary tinkering, that is, the magnitude and mode of diversification of expression attributes, as well as to precisely evaluate the role of putative disruptive mutational events such as TEs during the early stages of formation and consolidation of *Sdic* and similar tandemly repeated multigene families in eukaryotic genomes.

MATERIALS AND METHODS

Fly Husbandry

A combination of strains, including some with wild-type genotypes of diverse geographical origin (King, Macdonald, et al. 2012) and others carrying synthetic genotypes, was used (supplementary table S2.1, Supplementary Material online). Flies were reared on dextrose–cornmeal–yeast medium in a 25 C chamber under constant lighting conditions.

Engineering the Duplication of the *Sdic* Region

Engineered duplications of the *Sdic* region were generated using TE-bearing strains with w^{1118} genomic background (supplementary table S2.1, Supplementary Material online) (Parks et al. 2004), and following the same mating scheme used previously for deleting the region (supplementary fig. S2.3A, Supplementary Material online) (Yeh et al. 2012). Validation of the engineered duplications was done by inspecting eye color of particular male progeny and by performing a set of diagnostic PCR controls (supplementary fig. S2.3B, Supplementary Material online). See supplementary table S2.4, Supplementary Material online, for the primers utilized.

Sperm Competition Assays

Offense double-mating experiments for duplication-bearing males were performed as reported (Yeh et al. 2013), and concomitantly with those for other male genotypes whose results were already published (Yeh et al. 2012). Briefly, sperm competitive ability for any given male genotype was calculated with the P2 metric, which measures the relative contribution of the second male to mate to the total progeny of doubly mated females. The angular transformation was applied to the P2 values (Sokal and Rohlf 1994). Transformed P2 values were stored at Dryad repository (https://doi.org/10.7280/D1RH56).

In Situ Hybridization

To further assure that the engineered duplication of the *Sdic* region was generated in tandem, in situ hybridization on polytene chromosomes of the strains 2T and 4M was performed as described (Ranz et al. 1997). Probe and signal detection are as reported (Yeh et al. 2012). Further, in order to test the recapitulation of the *Sdic* region in the assembly of the strain A2, in situ hybridization on mitotic chromosomes from larval brains was executed as reported (Pimpinelli et al. 2000). The probe used spans a common region between *Sdic* and *sw*. See supplementary table S2.4, Supplementary Material online, for the primers utilized to generate the probes.

Genome Assemblies

Assemblies corresponding to the 13 strains from the *Drosophila* Synthetic Population Resources (King, Merkes, et al. 2012) plus OR-R were obtained from the NCBI bioproject PRJNA418342. These assemblies were scaffolded with SMRT sequencing reads and polished with Paired End 100 Illumina reads, and are characterized by N50 values $\geq$ 18.5Mb (average ~ 21.2Mb), coverages for the euchromatic fraction $\geq$ 36x (average ~70x), and complete BUSCO values $\geq$ 99.9% (Chakraborty et al. 2018, 2019). The Oxford_Nanopore- and Bionano-based assemblies (Solares et al. 2018) were obtained from https://github.com/danrdanny/Nanopore_ISO1 (last accessed February 1, 2019) and the Nanopore sequencing reads retrieved from the NCBI bioproject PRJNA433573.

Sdic Region Annotation

We used BlastN (Altschul et al. 1990) to locate the 5' section of *sw* and the 3' section of *AnxB10* to identify the boundaries of the *Sdic* region in each genome assembly. To extract the region from these assemblies, we used SAMtools/1.3 (Li et al. 2009) using the coordinates from BlastN plus 10 kb added to each side. Annotation of the *Sdic* region was done by searching for sequence motifs corresponding to exon 1 as in the ISO-1 assembly (Clifton et al. 2017). *Sdic* copies were numbered sequentially from *sw* to *AnxB10*. Raw reads associated with the *Sdic* region in each assembly were retrieved for detailed analyses upon identification using BlastN and mapped against the corresponding assembly using minimap2 (Li 2018). Additional features, essentially TE insertions, were characterized by BlastN through FlyBase (dos Santos et al. 2015), and their junctions confirmed by PCR; see supplementary table S2.4, Supplementary Material online, for the primers utilized. Open reading frames were inspected in MEGA X (Kumar et al. 2018), and the number of WD40 motifs associated with each putatively encoded *Sdic* protein determined according to a specialized database for WD40-repeat proteins (Ma et al. 2019).

Read-Depth Analysis

CNVnator (Abyzov et al. 2011) was used to survey CNV in the *Sdic* region using the "-genome" option and a bin size of 100 nt. Illumina sequencing outputs for the DSPR panel (King, Merkes, et al. 2012) and the ISO-1 strain (Langley et al. 2012) were retrieved from GenBank and mapped against a collection of synthetic reference genomes. These synthetic genomes were derived from the assemblies of the A4 and ISO-1 strains. Each synthetic genome contains a different single *Sdic* copy of those present in the mentioned assemblies and lacks the parental flanking genes *sw* and *AnxB10* (Supplementary Text). For any given strain surveyed, the average among all the read-depth estimates obtained from the different reference assemblies was calculated and then rounded off to its closest integer. From this value, 1 was subsequently subtracted because of the contribution of reads from the flanking genes *sw* and *AnxB10* to the read-depth estimates as, combined, they behave essentially as an additional *Sdic* copy. Given the overall high agreement between the average read-depth values obtained using the reference genomes derived from A4 and ISO-1 (Supplementary Text), only those from A4 were used in subsequent surveys of CNV across two additional panels of strains: PRJNA268111 (Grenier et al. 2015); and SRP006733 (Lack et al. 2016). As for these two additional panels of strains no qPCR estimates were available, we adopted the conservative criterion of considering read-depth average values from those strains showing CNV target sizes within reasonable boundaries, that is, 7.2–8.0 kb; in A4, *Sdic* copies range in size from 7.4 to 7.75 kb. Read-depth estimates associated with reference genomes for which the CNV target size was outside of the indicated range were omitted. Only strains for which the number of reliable read-depth estimates were 4–5 were considered in downstream analyses.

Population Differentiation

The V_{ST} statistic (Redon et al. 2006) was calculated for the CNVnator estimates as $V_{\mathrm{ST}} = (V_{\mathrm{T}} – V_{\mathrm{S}})/V_{\mathrm{T}}$, where V_{T} is the total variance in CN among all the considered individuals and V_{S} is the average of the variance within each single population, weighted for size. The calculation of the V_{ST} statistic was done for the rounded-off CN values, the uncorrected average read-depth values, and their log2, finding no difference. The probability of finding V_{ST} values equal or higher than that observed given the data was assessed by performing 10,000 simulations of bootstrap resampling.

qPCR CNV Assays

For each interrogated genotype, three genomic DNA extractions, that is, biological replicates, were performed. In each extraction, 20 entire whole bodies from<10-day post-eclosion individuals were homogenized with motorized pestles in 1.5ml tubes. Genomic DNA was extracted using the Qiagen's Puregene Core Kit B, and further purified using Zymo Research's Genomic DNA Clean & Concentrator-10 kit following manufacturer's instructions. DNA purity was confirmed with a NanoDrop 8000 spectrophotometer (Thermo Fisher), and the specificity of expected amplicons by agarose gel electrophoresis of the qPCR products and the analysis of the melting curves from the qPCR instrument. DNA concentrations were measured using a Qubit 2.0 fluorometer with either Qubit dsDNA BR Assay Kit or Qubit dsDNA HS Assay Kit reagents when appropriate. Real-time qPCR CNV assays were performed accommodating *Sdic*'s chimeric nature, which prevents designing reliable *Sdic*-specific primers. Thus, the number of *Sdic* copies was inferred by performing two sets of qPCR assays in which the first set was specific to *sw* whereas the second annealed with both *sw* and all *Sdic* copies (*Sdic/sw*). Accordingly, the number of *Sdic* copies in any given genotype was inferred by subtracting the number of *sw* copies from the number of *Sdic/sw* copies. Raw CNs estimates were obtained accounting for variable primer efficiencies for the gene of interest and the reference gene (Pfaffl 2001). A randomly chosen single copy autosomal gene *Triose phosphate isomerase* (*Tpi*) was used as a reference. Real-time PCR experiments were performed in 20 ml reactions using PowerUP SYBR Green Master Mix (Applied Biosystems), 5 mM of each primer, and ~30 ng of purified genomic DNA in 96-well plates on a Bio-Rad CFX-96 1000 touch real-time PCR instrument. Primer sets are listed in supplementary table S2.4, Supplementary Material online. The average raw gene CN across genotypes was calculated relative to ISO-1 females. Calling CN was done by rounding average raw CN estimates to the nearest integer. Original Ct values were stored at Dryad repository (https://doi.org/10.7280/D11091).

qRT–PCR Expression Assays

Experiments were done using four replicates of total RNA extractions from whole-body males with a CFX-96 1000 touch real-time instrument (BioRad) using the PowerUP SYBR Green Master Mix (Applied Biosystems) with 1 ml cDNA in a 20ml reaction. Total RNA was extracted from ten strains (fig. 2.5) using TRIzol reagent (Thermo Fisher) following manufacturer instructions. Fifty naive males per replicate per strain were systematically sacrificed at 3 pm to control for circadian rhythms and extracted on separate days to avoid strain cross-contamination. DNA traces were subsequently eliminated using the RNeasy mini kit with DNase I (Qiagen). RNA integrity, purity, and concentration were assessed using gel electrophoresis, Nanodrop, and a Qubit RNA BR assay kit, respectively. Each sample was converted to cDNA using 1.5 mg total RNA and the SuperScript IV first-strand synthesis system with an RNase inhibitor (Invitrogen). Effective reverse transcriptase reactions were confirmed through successful RT–PCR of the gene Gapdh2. The gene *clot* was used as the reference gene and males from ISO-1 were used for calibration. Expression estimates were obtained accounting for variable primer efficiencies for the gene of interest (*Sdic*) and the reference gene (Pfaffl 2001). Primers used are provided in supplementary table S2.4, Supplementary Material online. Primer design for *Sdic* took into consideration sequence differences with *sw* and *AnxB10* to confidently survey solely *Sdic* expression, as well as perfect sequence conservation across copies and strains to prevent any copy or population bias. Original Ct values were stored at Dryad repository (https://doi.org/10.7280/D1W98H).

Expression Profiling of *AnxB10*-Like

Thirty-eight libraries representing 29 biological conditions throughout the *D. melanogaster* life cycle (Graveley et al. 2011) were downloaded from the NCBI FTP site (supplementary table S2.12, Supplementary Material online). Reads with remaining adapters or with a quality value $Q \leq 20$ were discarded. All remaining reads were then examined for >70-nt alignments with a 130-nt sequence that includes a core motif distinctive of three of the *AnxB10*-like copies (ATAGGTCAGTATATACATATTTAACTGTTCCGTT; underlined, insertion absent in *AnxB10*) using an in-home script that incorporated the local alignment function from the Biopython package (Cock et al. 2009). The whole core motif was required to be part of the alignment with no mismatch or gap allowed; the extension of the alignment upstream or downstream could contain a single-nucleotide mismatch or indel. An in-house Python script was used to ultimately determine the number of sequencing reads fulfilling the above conditions.

Gene Conversion Analysis

Multiple sequence alignments (MSA) for the *Sdic* repeats in each strain and for all strains for which their genome assemblies were dubbed as reliable were generated and aligned with MUSCLE within MEGA X (Kumar et al. 2018). Each MSA included a synthetic composite sequence consisted of *Sdic*'s equivalent regions in *sw* and *AnxB10*. Levels of nucleotide differentiation were calculated under a Jukes–Cantor substitution model in MEGA X. All positions containing gaps and missing data were eliminated (completed deletion option). Gene conversion tracts were inferred using the GeneConv software (Sawyer 1989) under the assumption that no nucleotide mismatch occurred among the tracts, thus limiting the number of false positives. In addition, only gene conversion tracts with an associated probability < 0.05 after correcting for multiple tests were considered. Inference of recombination breakpoints was done with the ACG software (O'Fallon 2013) under 20,000,000 iterations and a burn-in period of 5,000,000. Circular layouts showing the topology of gene conversion events in each strain were generated with the Circos software (Krzywinski et al. 2009).

Phylogenetic Analysis of the DSPR Strains

Contigs containing the mitochondrial genome of each DSPR strain and OR-R were identified via BlastN and extracted from genome assemblies using SAMtools/1.3 (Li et al. 2009). The mitochondrial genome sequence from the reference ISO-1 strain was retrieved from GenBank (accession number: KJ947872) and included in the analysis. Sequence alignment was generated using MUSCLE and subsequently minimally curated by visual inspection. The best model of nucleotide evolution was found to be the Hasegawa–Kishino–Yano model (Hasegawa et al. 1985). The evolutionary history was inferred by using the Maximum Likelihood method. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Joining (NJ) and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories; +G, parameter = 0.0500). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 49.13% sites). All positions containing gaps and missing data were eliminated (complete deletion option). The final data set included 17,964 nucleotide sites. Bootstrapping (1,000 replicates) was performed to determine the confidence of the branches (Felsenstein 1985). Evolutionary analyses were conducted in MEGA X (Kumar et al. 2018).

Phylogenetic Analysis of Annotated *Sdic* Copies

The phylogenetic relationship among the *Sdic* copies from a subset of strains from the DSPR panel was inferred using a MSA including all *Sdic* copies and composites, and RAxML 8.1.2 (Stamatakis 2014), under a GTRGamma model of sequence evolution. The resulting topology was evaluated through 1,000 bootstrap replicates. This topology is very similar to an alternative one as inferred with PhyML 3.0 (Guindon et al. 2010), which is based on the best-fit substitution model HKY85 + G + I with four gamma categories according to SMS (http://www.atgc-montpellier.fr/sms/; last accessed October 21, 2019).

Positive Selection Analysis

The software package HyPhy (Kosakovsky Pond et al. 2020) was used to test for positive selection acting on coding and noncoding *Sdic* sequences. The adaptive branch-site random effects model (aBSREL; Smith et al. [2015]) and the batch script written by Oliver Fredigo (Haygood et al. [2007]; upgraded to run on Hyphy version 2.5, https://github.com/spond/TestForPositiveSelection/nonCodingSelection.bf; last accessed October 21, 2019) were applied to the coding and noncoding regions, respectively, of the MSA of the *Sdic* repeat in all strains, including the synthetic composite sequences from different strains, and the composite sequence consisted of their corresponding orthologous stretches to *sw* and *AnxB10* in *D. simulans*, which was used as a more external outgroup. See Supplementary Materials for further details. To accommodate for the different gene tree topologies and total branch lengths of sampled genealogies for each partition (or subpartitions) along the MSA identified by the ACG recombination breakpoints, we conducted the test separately for each of these partitions using their respective gene tree (one per partition).

Statistical Analyses

One-way ANOVA and post hoc Tukey's HSD tests for detecting differences in mRNA levels across genotypes were done in JMP 12.2.0 (SAS Institute Inc.). Nonparametric Kruskal–

Wallis H and pairwise Stell–Dwass tests, which corrects for multiple testing, for detecting differences in sperm competitive ability among genotypes as well as for assessing differences in CN among populations from the GDL panel were done also with the same statistical package. Bootstrap resampling, hierarchical clustering, and logistic regression analyses were done in R (R Development Core Team 2016).

ACKNOWLEDGEMENTS

We are in debt to Mahul Chakraborty, J.J. Emerson, Brandon Gaut, Anthony Long, John Parsch, and Andrei Tatarenkov for feedback on analytical procedures and early versions of this article, and for providing us early access to genome assemblies. We also thank Ali Mortazavi for granting us access to his qRT–PCR instrument and to the University of California, Irvine High-Performance Computing cluster for facilitating our analyses. This work was supported by a National Science Foundation Grant (MCB-1157876) and by University of California, Irvine intramural funds (to J.M.R.).

REFERENCES

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*. 21(6):974–984.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215(3):403–410.

Begun DJ, Aquadro CF. 1993. African and North American populations of *Drosophila melanogaster* are very different at the DNA level. *Nature* 365(6446):548–550.

Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol*. 33(6):623–630.

Bingham PM. 1980. The regulation of white locus expression: a dominant mutant allele at the white locus of *Drosophila melanogaster*. *Genetics* 95(2):341–353.

Birkhead TR. 1998. Sperm competition in birds. *Rev Reprod*. 3(2):123–129.

Boylan K, Serr M, Hays T. 2000. A molecular genetic analysis of the interaction between the cytoplasmic dynein intermediate chain and the glued (dynactin) complex. *Mol Biol Cell*. 11(11):3791–3803.

Boylan KL, Hays TS. 2002. The gene for the intermediate chain subunit of cytoplasmic dynein is essential in *Drosophila*. *Genetics* 162(3):1211–1220.

Brown CJ, Todd KM, Rosenzweig RF. 1998.Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Mol Biol Evol*. 15(8):931–942.

Cardoso-Moreira M, Arguello JR, Gottipati S, Harshman LG, Grenier JK, Clark AG. 2016. Evidence for the fixation of gene duplications by positive selection in *Drosophila*. *Genome Res*. 26(6):787–798.

Carpenter D, Dhar S, Mitchell LM, Fu B, Tyson J, Shwan NA, Yang F, Thomas MG, Armour JA. 2015. Obesity, starch digestion and amylase: association between copy number variants at human salivary (*AMY1*) and pancreatic (*AMY2*) amylase genes. *Hum Mol Genet*. 24(12):3472–3480.

Catalan A, Glaser-Schmitt A, Argyridou E, Duchen P, Parsch J. 2016. An indel polymorphism in the *MtnA* 3' untranslated region is associated with gene expression variation and local adaptation in *Drosophila melanogaster*. *PLoS Genet*. 12(4):e1005987.

Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. 2019. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat Commun*. 10(1):4872.

Chakraborty M, VanKuren NW, Zhao R, Zhang X, Kalsow S, Emerson JJ. 2018. Hidden genetic variation shapes the structure of functional elements in *Drosophila*. *Nat Genet*. 50(1):20–25.

Civetta A, Ranz JM. 2019. Genetic factors influencing sperm competition. *Front Genet*. 10:820.

Clifton BD, Librado P, Yeh SD, Solares ES, Real DA, Jayasekera SU, Zhang W, Shi M, Park RV, Magie RD, et al. 2017. Rapid functional and sequence differentiation of a tandemly repeated species-specific multigene family in *Drosophila*. *Mol Biol Evol*. 34(1):51–65.

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11):1422–1423.

Darwin C. 1871. The descent of man and selection in relation to sex. London: John Murray.

Dennis MY, Eichler EE. 2016. Human adaptation and evolution by segmental duplication. *Curr Opin Genet Dev*. 41:44–52.

Dennis MY, Harshman L, Nelson BJ, Penn O, Cantsilieris S, Huddleston J, Antonacci F, Penewit K, Denman L, Raja A, et al. 2017. The evolution and population diversity of human-specific segmental duplications. *Nat Ecol Evol*. 1(3):69.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, et al. 2012. Evolution of human-specific neural *SRGAP2* genes by incomplete segmental duplication. Cell 149(4):912–922.

Didion JP, Morgan AP, Clayshulte AM, McMullan RC, Yadgary L, Petkov PM, Bell TA, Gatti DM, Crowley JJ, Hua K, et al. 2015. A multi-megabase copy number gain causes maternal transmission ratio distortion on mouse chromosome 2. *PLoS Genet*. 11(2):e1004850.

dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, Emmert DB, Gelbart WM, FlyBase C, the FlyBase Consortium. 2015. FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res*. 43(D1):D690–D697.

Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HO, Buffalo V, Zerbino DR, Diekhans M, et al. 2011. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res*. 21(12):2224–2241.

Eaton ML, Prinz JA, MacAlpine HK, Tretyakov G, Kharchenko PV, MacAlpine DM. 2011. Chromatin signatures of the *Drosophila* replication program. *Genome Res*. 21(2):164–174.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39(4):783–791.

Feyereisen R, Dermauw W, Van Leeuwen T. 2015. Genotype to phenotype, the molecular and physiological dimensions of resistance in arthropods. *Pestic Biochem Physiol*. 121:61–77.

Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, van den Bout A, Bishara A, Rosenkrantz JL, et al. 2018. Human-specific *NOTCH2NL* genes affect notch signaling and cortical neurogenesis. *Cell* 173(6):1356–1369.

Field LM, Blackman RL, Tyler-Smith C, Devonshire AL. 1999. Relationship between amount of esterase and gene copy number in insecticide-resistant *Myzus persicae* (Sulzer). *Biochem J*. 339(3):737–742.

Fouche S, Plissonneau C, McDonald BA, Croll D. 2018. Meiosis leads to pervasive copy-number variation and distorted inheritance of accessory chromosomes of the wheat pathogen *Zymoseptoria tritici*. *Genome Biol Evol*. 10(6):1416–1429.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al. 2011. The developmental transcriptome of *Drosophila melanogaster*. *Nature* 471(7339):473–479.

Grenier JK, Arguello JR, Moreira MC, Gottipati S, Mohammed J, Hackett SR, Boughton R, Greenberg AJ, Clark AG. 2015. Global diversity lines – a five-continent reference panel of sequenced *Drosophila melanogaster* strains. *G3* (Bethesda) 5:593–603.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 59(3):307–321.

Guryev V, Saar K, Adamovic T, Verheul M, van Heesch SA, Cook S, Pravenec M, Aitman T, Jacob H, Shull JD, et al. 2008. Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet*. 40(5):538–545.

Harewood L, Chaignat E, Reymond A. 2012. Structural variation and its effects on expression. *Methods Mol Biol*. 838:173–186.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*. 22(2):160–174.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009.Mechanisms of change in gene copy number. *Nat Rev Genet*. 10(8):551–564.

Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet*. 39(9):1140–1144.

Hollox EJ. 2008. Copy number variation of beta-defensins and relevance to disease. *Cytogenet Genome Res*. 123(1–4):148–155.

Hollox EJ. 2012. The challenges of studying complex and dynamic regions of the human genome. *Methods Mol Biol*. 838:187–207.

Huddleston J, Eichler EE. 2016. An incomplete understanding of human genetic variation. *Genetics* 202(4):1251–1254.

Hug N, Longman D, Caceres JF. 2016. Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res*. 44(4):1483–1495.

Hurles M. 2004. Gene duplication: the genomic trade in spare parts. *PLoS Biol*. 2(7):E206.

Jayaswal V, Jimenez J, Magie R, Nguyen K, Clifton B, Yeh S, Ranz JM. 2018. A species-specific multigene family mediates differential sperm displacement in *Drosophila melanogaster*. *Evolution* 72(2):399–403.

Jiang W, Johnson C, Jayaraman J, Simecek N, Noble J, Moffatt MF, Cookson WO, Trowsdale J, Traherne JA. 2012. Copy number variation leads to considerable diversity for B but not A haplotypes of the human KIR genes encoding NK cell receptors. *Genome Res*. 22(10):1845–1854.

Jugulam M, Niehues K, Godar AS, Koo DH, Danilova T, Friebe B, Sehgal S, Varanasi VK, Wiersma A, Westra P, et al. 2014. Tandem amplification of a chromosomal segment harboring *5-enolpyruvylshikimate-3-phosphate synthase* locus confers glyphosate resistance in *Kochia scoparia*. *Plant Physiol*. 166(3):1200–1207.

Karn RC, Laukaitis CM. 2009. The mechanism of expansion and the volatility it created in three pheromone gene clusters in the mouse (*Mus musculus*) genome. *Genome Biol Evol*. 1:494–503.

Katju V, Bergthorsson U. 2013. Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Front Genet* . 4:273.

King EG, Macdonald SJ, Long AD. 2012a. Properties and power of the *Drosophila* Synthetic Population Resource for the routine dissection of complex traits. *Genetics* 191(3):935–949.

King EG, Merkes CM, McNeil CL, Hoofer SR, Sen S, Broman KW, Long AD, Macdonald SJ. 2012b. Genetic dissection of a model complex trait using the *Drosophila* Synthetic Population Resource. *Genome Res*. 22(8):1558–1566.

Kondrashov FA. 2010. Gene dosage and duplication. In: Dittmar K, Liberles D, editors. Evolution after gene duplication. Hoboken (NJ): Wiley-Blackwell. p. 57–76.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. *Genome Biol*. 3(2):research0008.1. RESEARCH0008.

Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, Shank SD, Magalis BR, Bouvier D, Nekrutenko A, et al. 2020. HyPhy 2.5 – a customizable platform for evolutionary hypothesis testing using phylogenies. *Mol Biol Evol*. 37(1):295–299.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res*. 19(9):1639–1645.

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

Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE. 2016. A thousand fly genomes: an expanded *Drosophila* genome nexus. *Mol Biol Evol*. 33(12):3308–3313.

Langley CH, Stevens K, Cardeno C, Lee YC, Schrider DR, Pool JE, Langley SA, Suarez C, Corbett-Detig RB, Kolaczkowski B, et al. 2012.Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics* 192(2):533–598.

Lee JA, Carvalho CM, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* 131(7):1235–1247.

Lemos B, Araripe LO, Fontanillas P, Hartl DL. 2008. Dominance and the evolutionary accumulation of *cis*- and *trans*-effects on gene expression. *Proc Natl Acad Sci U S A*. 105(38):14471–14476.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34(18): 3094–3100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet*. 2(10):e166.

Loehlin DW, Carroll SB. 2016. Expression of tandem gene duplicates is often greater than twofold. *Proc Natl Acad Sci USA*. 113(21):5988–5992.

Long M, VanKuren NW, Chen S, Vibranovski MD. 2013. New gene evolution: little did we know. *Annu Rev Genet*. 47(1):307–333.

Ma J, An K, Zhou JB, Wu NS, Wang Y, Ye ZQ, Wu YD. 2019.WDSPdb: an updated resource for WD40 proteins. *Bioinformatics* 35(22):4824–4826.

Naseeb S, Ames RM, Delneri D, Lovell SC. 2017. Rapid functional and evolutionary changes follow gene duplication in yeast. *Proc R Soc B*. 284(1861):20171393.

Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL. 1998. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature* 396(6711):572–575.

Nuttle X, Giannuzzi G, Duyzend MH, Schraiber JG, Narvaiza I, Sudmant PH, Penn O, Chiatante G, Malig M, Huddleston J, et al. 2016. Emergence of a *Homo sapiens*-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* 536(7615):205–209.

O'Fallon BD. 2013. ACG: rapid inference of population history from recombining nucleotide sequences. *BMC Bioinformatics* 14:40.

Obbard DJ, Maclennan J, Kim KW, Rambaut A, O'Grady PM, Jiggins FM. 2012. Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. *Mol Biol Evol*. 29(11):3459–3473.

Parker GA. 1970. Sperm competition and its evolutionary consequences in the insects. *Biol Rev*. 45(4):525–567.

Parks AL, Cook KR, Belvin M, Dompe NA, Fawcett R, Huppert K, Tan LR, Winter CG, Bogart KP, Deal JE, et al. 2004. Systematic generation of high-resolution deletion coverage of the *Drosophila melanogaster* genome. *Nat Genet*. 36(3):288–292.

Pezer Z, Chung AG, Karn RC, Laukaitis CM. 2017. Analysis of copy number variation in the *Abp* gene regions of two house mouse subspecies suggests divergence during the gene family expansions. *Genome Biol Evol*. 9(6):1393–1405.

Pezer Z, Harr B, Teschke M, Babiker H, Tautz D. 2015. Divergence patterns of genic copy number variation in natural populations of the house mouse (*Mus musculus domesticus*) reveal three conserved genes with major population-specific expansions. *Genome Res*. 25(8):1114–1124.

Pfaffl MW. 2001. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res*. 29:e45.

Pimpinelli S, Bonaccorsi S, Fanti L, Gatti M. 2000. Preparation and analysis of *Drosophila* mitotic chromosomes. In: Sullivan W, Ashburner M, Hawley RS, editors. *Drosophila* protocols. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press. p. 3–23.

R Development Core Team. 2016. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Ranz J, Clifton B. 2019. Characterization and evolutionary dynamics of complex regions in eukaryotic genomes. *Sci China Life Sci*. 62(4):467–488.

Ranz JM, Segarra C, Ruiz A. 1997. Chromosomal homology and molecular organization of Muller’s elements D and E in the *Drosophila repleta* species group. *Genetics* 145:281–295.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen WW, et al. 2006. Global variation in copy number in the human genome. *Nature* 444(7118):444–454.

Rice AM, McLysaght A. 2017. Dosage-sensitive genes in evolution and disease. *BMC Biol*. 15(1):78.

Rogers RL, Shao L, Thornton KR. 2017. Tandem duplications lead to novel expression patterns through exon shuffling in *Drosophila yakuba*. *PLoS Genet*. 13(5):e1006795.

Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC. 2003. Abundant gene conversion between arms of palindromes in human and ape *Y* chromosomes. *Nature* 423(6942):873–876.

Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol Biol Evol*. 6(5):526–538.

Schmidt JM, Good RT, Appleton B, Sherrard J, Raymant GC, Bogwitz MR, Martin J, Daborn PJ, Goddard ME, Batterham P, et al. 2010. Copy number variation and transposable elements feature in recent, ongoing adaptation at the *Cyp6g1* locus. *PLoS Genet*. 6(6):e1000998.

Scott EC, Gardner EJ, Masood A, Chuang NT, Vertino PM, Devine SE. 2016. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res*. 26(6):745–755.

Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. 2015. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol Biol Evol*. 32(5):1342–1353.

Soh YQS, Alfoldi J, Pyntikova T, Brown LG, Graves T, Minx PJ, Fulton RS, Kremitzki C, Koutseva N, Mueller JL, et al. 2014. Sequencing the mouse *Y* chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell* 159(4):800–813.

Sokal RR, Rohlf FJ. 1994. Biometry: the principles and practice of statistics in biological research. San Francisco: W. H. Freeman.

Solares EA, Chakraborty M, Miller DE, Kalsow S, Hall K, Perera AG, Emerson JJ, Hawley RS. 2018. Rapid low-cost assembly of the *Drosophila melanogaster* reference genome using low-coverage, long-read sequencing. *G3* (Bethesda) 8:3143–3154.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

Stephan W, Li H. 2007. The recent demographic and adaptive history of *Drosophila melanogaster*. *Heredity* 98(2):65–68.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Eichler EE, 1000 Genomes Project. 2010. Diversity of human copy number variation and multicopy genes. *Science* 330(6004):641–646.

Teitz LS, Pyntikova T, Skaletsky H, Page DC. 2018. Selection has countered high mutability to preserve the ancestral copy number of *Y* chromosome amplicons in diverse human lineages. *Am J Hum Genet*. 103(2):261–275.

Traherne JA, Martin M, Ward R, Ohashi M, Pellett F, Gladman D, Middleton D, Carrington M, Trowsdale J. 2010. Mechanisms of copy number variation and hybrid gene formation in the *KIR* immune gene complex. *Hum Mol Genet*. 19(5):737–751.

Yeh SD, Chan C, Ranz JM. 2013. Assessing differences in sperm competitive ability in *Drosophila*. *J Vis Exp*. 78:e50547.

Yeh SD, Do T, Chan C, Cordova A, Carranza F, Yamamoto EA, Abbassi M, Gandasetiawan KA, Librado P, Damia E, et al. 2012. Functional evidence that a recently evolved *Drosophila* sperm-specific gene boosts sperm competition. *Proc Natl Acad Sci U S A*. 109(6):2043–2048.

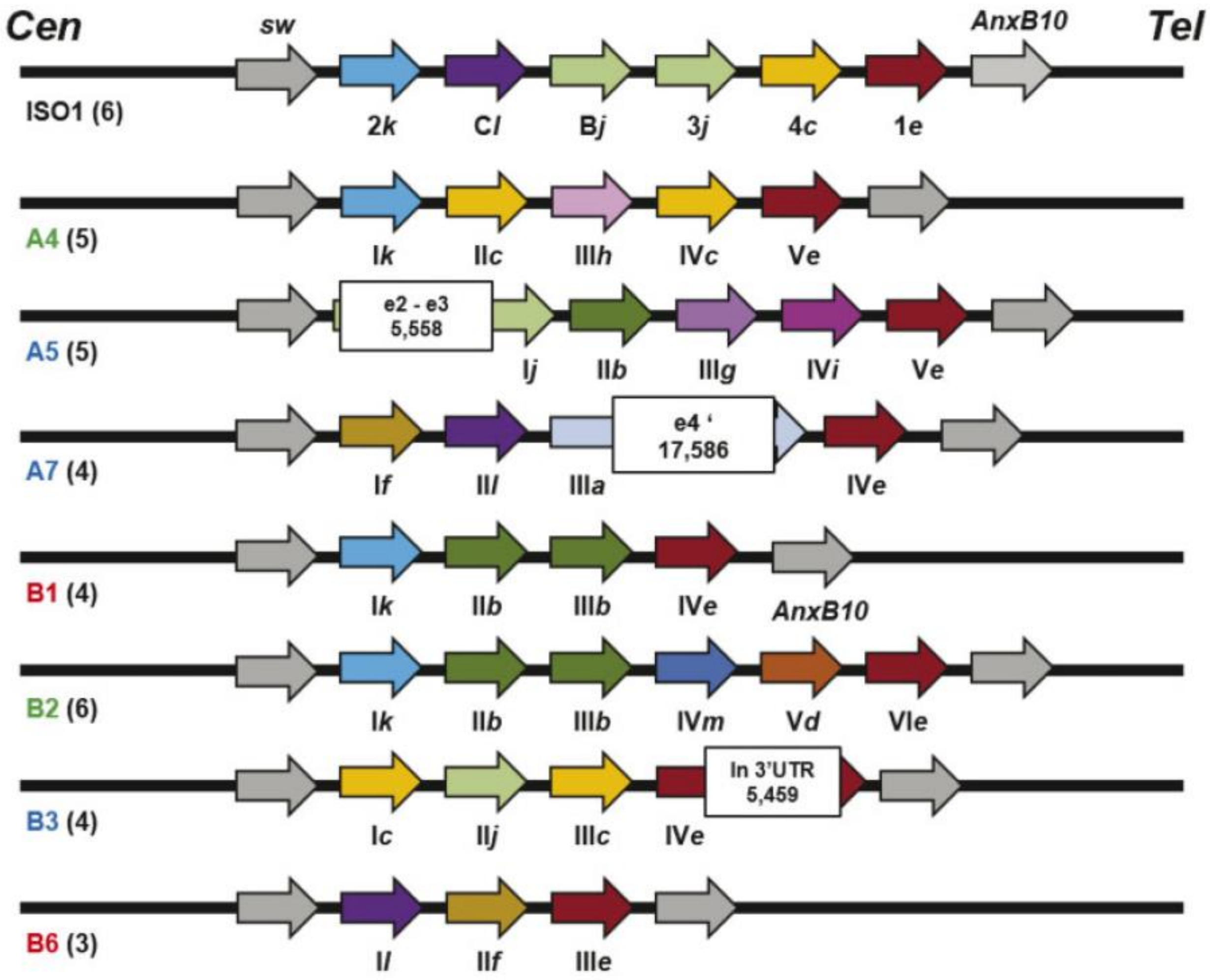


Figure 2.1. Annotation of the *Sdic* region across seven populations of the DSPR panel. The most reliable organization of the region at 19C1 on the *X* chromosome in the ISO-1 is provided as a reference (Clifton et al. 2017). The region is depicted from centromere (Cen) to telomere (Tel), including the flanking genes *sw* and *AnxB10* (gray-filled arrows). Population names are color-coded based on the broad continental region where they were collected: green, Africa; red, Americas; and blue, Eurasia. The number of annotated *Sdic* copies in reference-quality genome assemblies (Chakraborty et al. 2018, 2019) is indicated in parentheses next to the name of the population. *Sdic* copies in the ISO-1 strain are named as reported (Clifton et al. 2017). In the rest of populations, the copy identifiers are roman numerals according to their relative order from *sw* to *AnxB10*. *Sdic* copies are color-coded, and a lower character (a–m) added to their identifier, both indicating the associated paratype. Three TE insertions (solid boxes) are shown, indicating both their size in kb and the location in relation to the gene structure (e, exon). One TE insertion is located within intronic sequence (A5_I), a common occurrence (Chakraborty et al. 2018). In the other two cases, A7_III and B3_IV, the TE disrupts coding and 3’-UTR sequence, respectively. In the first case, the TE has possibly no functional consequence as a premature STOP codon resides upstream of the TE insertion; the apostrophe indicates an ancestral coding exon, which now situates outside of the predicted open reading frame.

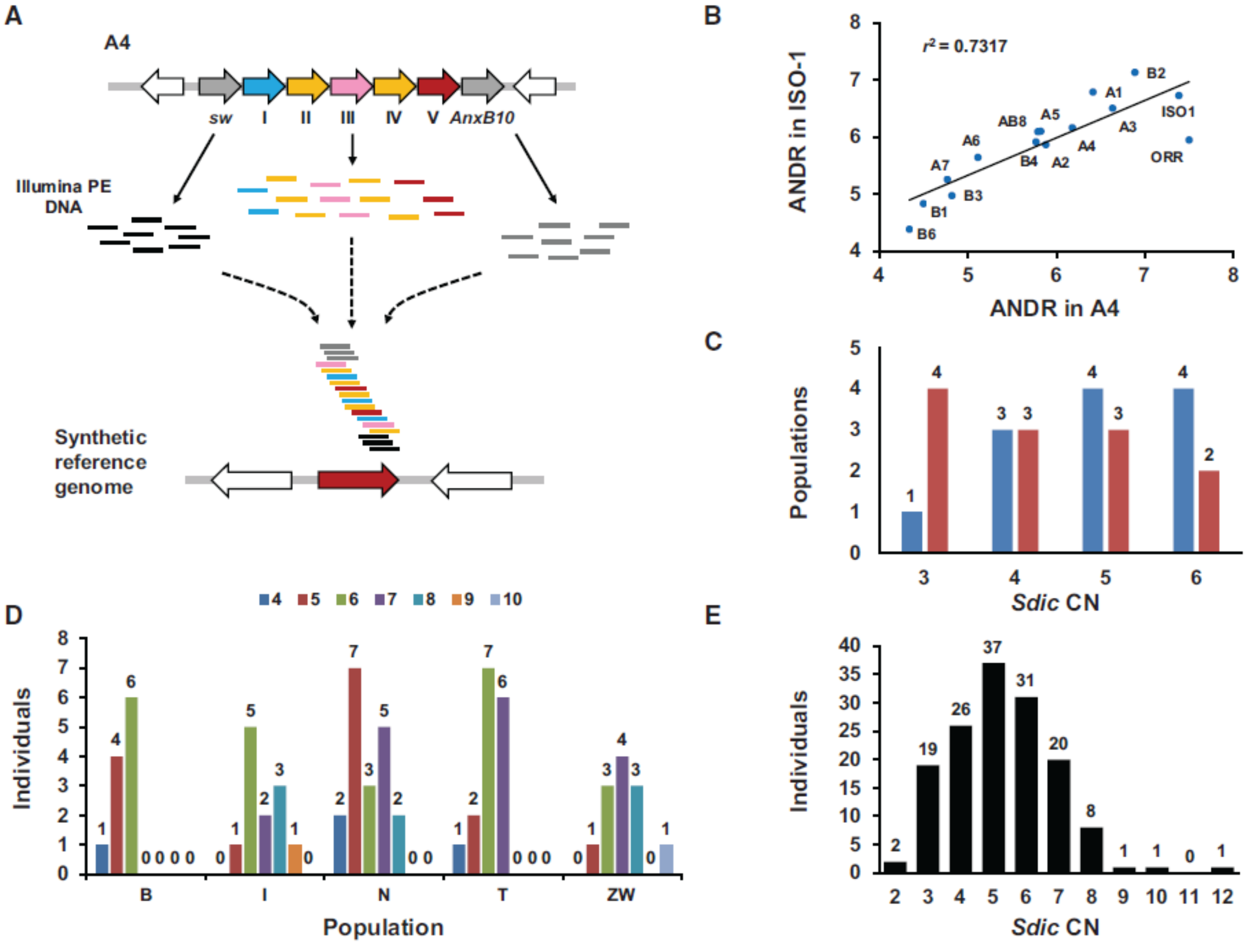


Figure 2.2. *Sdic* CNV estimation using a read-depth methodology. (A) Normalized read-depth estimates were obtained using CNVnator (Abyzov et al. 2011). To use as a reference genome, we generated a collection of synthetic *X* chromosomes carrying one *Sdic* repeat each from all the copies in the A4 and ISO-1 strains (only one of them, from the A4 strain, is shown). These synthetic *X* chromosomes also lacked the parental genes *sw* and *AnxB10* (gray-filled arrows), as advised by our benchmarking analysis. Therefore, all Illumina reads belonging to the *Sdic* copies and most of those from the parental genes should presumably map against the *Sdic* copy present in the synthetic genome. Open arrows, genes flanking the *Sdic* region. (B) Scatter plot of the averaged normalized read-depth (ANRD) estimates obtained using the synthetic genomes from ISO-1 and A4 for each of the strains assayed. Eliminating the most discordant strain, OR-R, the shown determination coefficient (r^2) becomes 0.901; r^2 is statistically significant ($P < 0.0001$) in both cases. These results show that the estimates do not depend on the reference strain used to generate the synthetic reference chromosomes. (C) Frequency distribution of populations from the DSPR panel based on the number of structurally distinct alleles in CN that they carry. A2 and A6 are omitted due to obvious errors in the assembly of the *Sdic* region. Blue, CNVnator round-off values; red, gene annotation values. (D) *Sdic* CNV across five populations of *Drosophila melanogaster*. Rounded-off average read-depth estimates obtained with CNVnator on the number of *Sdic* copies across 70 strains (each strain represents one individual) are shown (supplementary table S2.8, Supplementary Material online). The average read-depth estimate is calculated using

the values obtained from all synthetic reference *X* chromosomes. Different CNs are color-coded above. CN estimates and sequence coverage were not found to be correlated (r^2 = 0.0008; *P* = 0.8198). B, Beijing, n = 11; I, Ithaca, n = 12; N, The Netherlands, n = 19; T, Tasmania, n = 16; Z, Zimbabwe, n = 12. (E) Frequency distribution of all individuals genotyped for *Sdic* CN, that is, OR-R plus the strains from the DSPR and GDL panels, as well as those from a Zambian population.

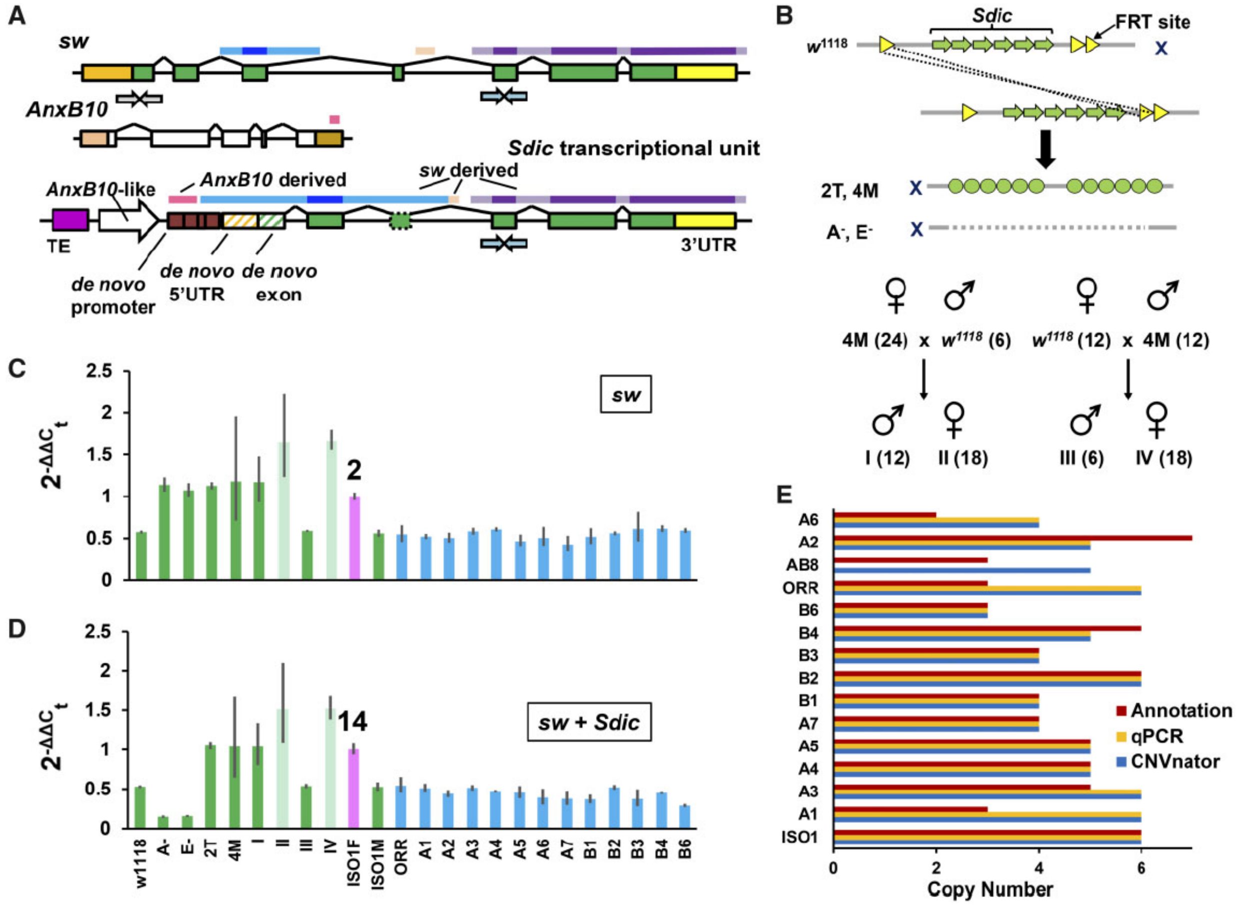


Figure 2.3. CNV estimates by qPCR. (A) Structure of *Sdic* and its parental genes *sw* and *AnxB10*. Colored horizontal bars above the gene models denote those regions donated to the chimeric gene *Sdic* from its parental genes. *Sdic* is part of a repeat also consisting of a partial fragment of the non-LTR retrotransposon Rt1c and an *AnxB10*-like entity, that is, a presumed pseudogene derived from *AnxB10*. *Sdic* exons are shown in green, with the exon one, a de novo exon not translated in *sw*, indicated with green diagonal stripes. A predicted alternatively spliced exon is indicated with a dotted box (Nurminsky et al. 1998). Two sets of primers were designed for the qPCR experiment; one exclusive of *sw* (gray-filled arrows) and the other able to amplify both *sw* and *Sdic* sequence (green-filled arrows). (B) Top, w^{1118}, a strain derived from OR-R (Bingham 1980) and used to generate FRT-bearing strains (Parks et al. 2004), which can be implemented in mating schemes to generate engineered *X* chromosomes carrying the deletion and the duplication of the *Sdic* cluster (middle). These induced chromosomal rearrangements result from FLP-mediated recombination events between FRT sites (see supplementary fig. S2.3, Supplementary Material online, for further details). Bottom, reciprocal crosses between a strain carrying the wild-type version of the cluster and another carrying its duplication in tandem to obtain progenies with a particular number of *Sdic* copies (in parenthesis). The known CN for *Sdic* and *sw* in each of the synthetic genotypes was used to calibrate our ability to discern differences in CN at the *Sdic* region. (C and D) Average fold change in CN for the gene *sw* and for *sw* jointly with *Sdic* across a set of control genotypes (green) and across a second set of geographically diverse strains (blue). The difference between the CNs associated with both amplicons corresponds to the number of *Sdic* copies for each genotype. Females from the reference strain (ISOF; pink) were used as calibrator in the estimation

of CN. Female genotypes are shown in faint colors. Error bars, SEM. ISOF and ISOM, females and males of the ISO-1 strain; A^{-} and E^{-}, deletion-bearing strains; 2T and 4M, duplication-bearing strains; I–IV, genotypes in the progeny from the reciprocal crosses outlined in (B). (E) Horizontal histogram showing the CN estimates obtained by qPCR, CNVnator, and genome annotation.

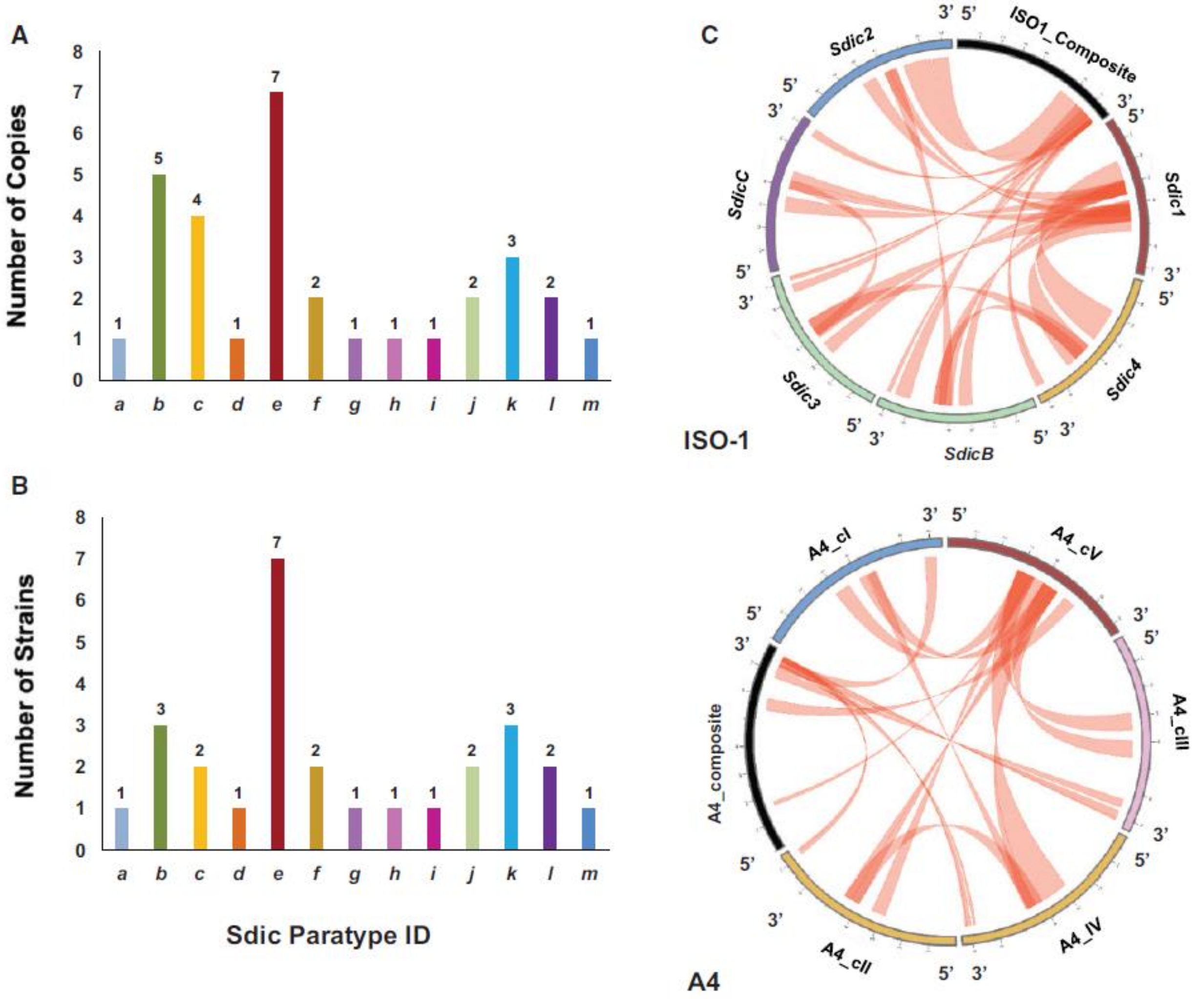


Figure 2.4. Salient patterns of molecular diversity in the *Sdic* region of seven populations of the DSPR panel. Each of these populations is represented by one isogenic strain derived from one single individual. The different paratypes are color-coded according to figure 1. (A) Number of copies in which the 13 Sdic paratypes were present across strains. Each paratype is present as 2.38 ± 1.89 copies, with six of them as a single copy (a, d, g, h, I, and m). (B) Presence of the 13 *Sdic* paratypes across strains. Each strain harbors *Sdic* copies associated with 3–5 paratypes (3.86 ± 0.90; mean ± SD), whereas each *Sdic* paratype is present in 1–7 copies across strains (2.17 ± 1.70; mean ± SD). For both (A) and (B), only data from the strains of the DSPR panel considered to be the most reliable for the *Sdic* region were examined. Two additional paratypes are not shown as they are not present in this subset of strains. (C) Gene conversion landscape in the *Sdic* region. Circular layout showing the topology of gene conversion events across *Sdic* copies and the composite (in black), that is, the fragments from *sw* plus *AnxB10* that align with *Sdic*. The results from GenConv (Sawyer 1989) are graphed for ISO-1 and A4; equivalent layouts for the other six strains are provided in supplementary fig. S2.6, Supplementary Material online. Gene conversion was found rampant across strains with an average of 5.6 events per copy and strain, showing distinctive topological patterns. Events involving paratype *e* primarily occur within the interval

2.3–7.2 kb from the start of the repeat, that is, from slightly upstream of the 5’-UTR of the *Sdic* transcriptional unit toward an internal position within the intron between *Sdic*’s exons 2 and 3. In contrast, the events involving *sw* occur 7.2 kb downstream from the start of the repeat, that is, within the intron between *Sdic*’s exons 2 and 3 (supplementary fig. S2.7, Supplementary Material online).

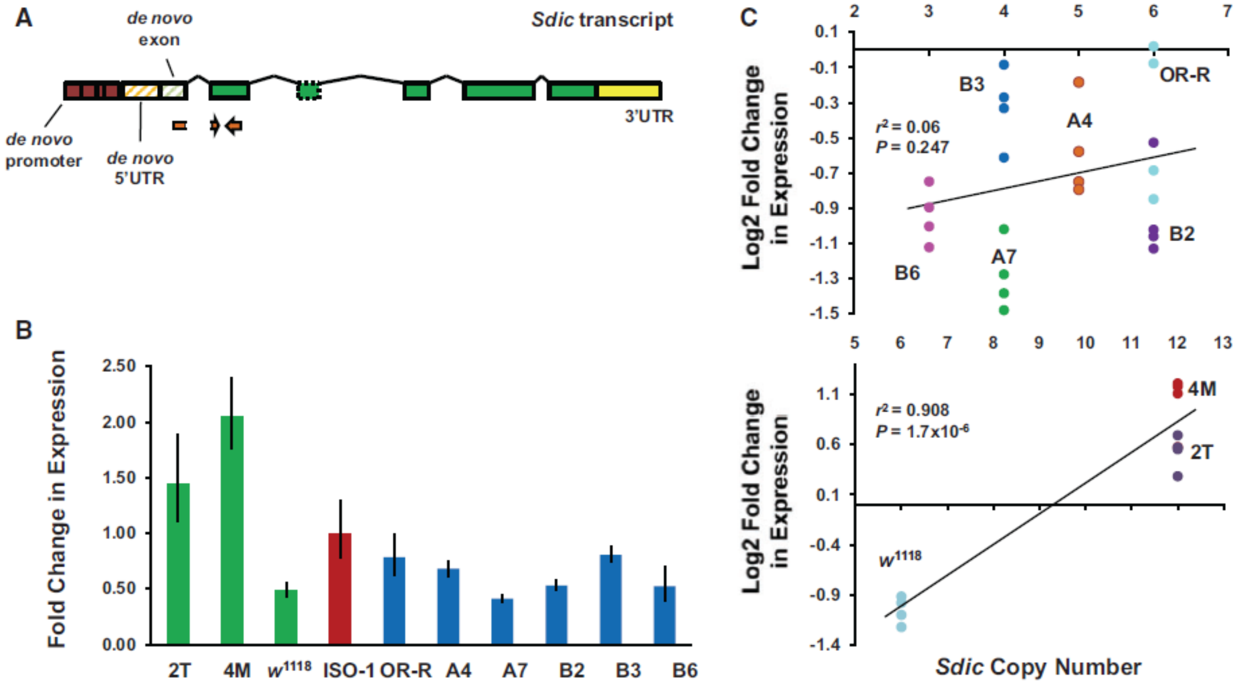


Figure 2.5. Global expression of the *Sdic* multigene family in whole-body males using qRT–PCR. (A) *Sdic* primers are shown relative to the *Sdic* transcriptional unit. See figure 3A for details about the relationship of different parts of this transcriptional unit with the structure of the parental genes. Primers were designed upon examining the sequence of all the copies across all the strains of geographically diverse origin, plus ISO-1, making sure that there was no mismatch or gap. The upstream primer was designed spanning the intron between exons 1 and 2 of *Sdic*, with only 5 nt within exon 2, to prevent amplification of *sw*. (B) Fold change in expression of ten strains, including ISO-1 (value of 1 on the y axis), which was used as calibrator. Green, w^{1118} and its synthetic derivatives carrying the duplication of the *Sdic* region (2T and 4M). Blue, strains of different geographical origin plus OR-R. Error bars, SEM. (C) Linear regression between CN and log2-fold change in expression for the two subsets of strains examined. Each dot represents the values obtained for each biological replicate included in the analysis. Determination coefficients (r^2) and their corresponding *P* values are shown.

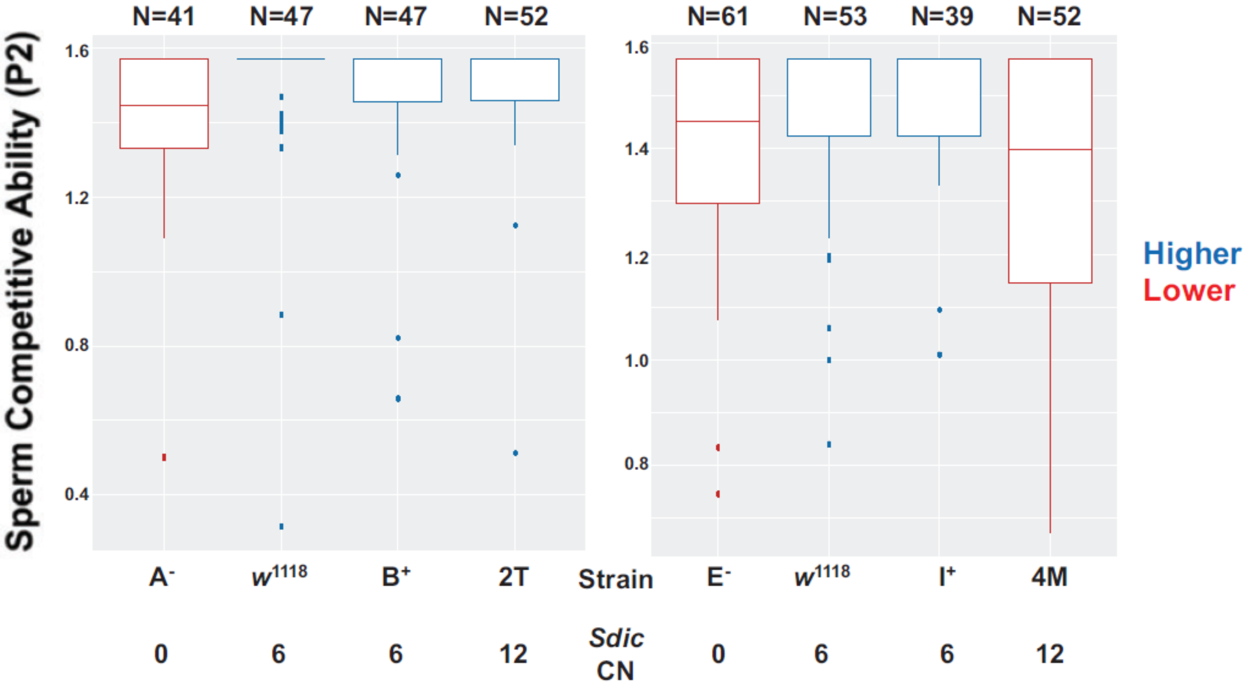


Figure 2.6. Sperm competitive ability in offense assays for males with different genotypes at the *Sdic* region. Left and right, two strain sets generated in the course of different structural modifications of the *Sdic* region, all of them derived from *w*^{1118}. Strains 2T and 4M, *Sdic* duplication-bearing males; A^{-} and E^{-}, *Sdic* deletion-bearing males; B^{+}, I^{+}, and *w*^{1118}, wild-type-like presence of the *Sdic* region. The data for 2T and 4M were obtained at the same time as for A^{-} , E^{-}, B^{+}, I^{+}, and *w*^{1118}; the data for the latter were reported (Yeh et al. 2012). Males from these strains were tested for differences in sperm competitive ability in displacing the sperm from a reference male when they were second to mate in double-mating experiments. The metric to measure sperm competitive ability in this type of experimental setting, P2, informs about the proportion of the progeny sired in double-matings. The angular transformation was applied to the P2 values, which are shown. Box plots show dispersion around the median and are color-coded indicating significantly different sperm competitive abilities (P_{adj}<0.05; supplementary table S2.17, Supplementary Material online, for the *P* adjusted values from all pairwise contrasts performed). The box plots of male genotypes showing significantly higher sperm competitive ability are shown in blue, whereas those performing poorer are in red. Genotypes with identical color denote no significant differences in the trait assayed. Males from *Sdic* duplication-bearing strains never show higher sperm competitive ability than males carrying the wild-type-like form of the *Sdic* region. In fact, these males can have even lower sperm competitive ability compared with males from *Sdic* deletion-bearing males (4M vs. E^{-}). Top, number of females for which their progeny was examined.

SUPPLEMENTARY TEXT

Organization of the *Sdic* region in the reference strain of *D. melanogaster*

The structural and sequence features of the *Sdic* region in the ISO-1 reference strain have been subject to recurrent updates in different releases (Ranz and Clifton 2019). A comparison across assemblies (Clifton, et al. 2017), including Release 6 (dos Santos, et al. 2015) and others generated with long sequencing reads, pointed to one scaffolded with single-molecule real-time (SMRT) (Kim, et al. 2014) sequencing reads as the most accurate: GCA_000778455 or *Berlin* hereafter (Berlin, et al. 2015). This reconstruction of the *Sdic* region entails discrepancies in copy number (six instead of seven) and internal positioning within the array in relation to Release 6. Further support for this different reconstruction derives from an independent assembly that used the same SMRT sequencing input, Illumina sequencing reads (Langley, et al. 2012), and a different computational pipeline (Chakraborty, et al. 2016). The nucleotide-to-nucleotide comparison of the *Sdic* region between these two assemblies uncovered no discrepancy relative to copy number, orientation, or internal positioning, displaying just 9 nt differences, 7 of them part of nucleotide runs.

To further test the reliability of our CN estimate for the *Sdic* region in the ISO-1 strain independently from SMRT-based assemblies, we adopted two strategies. First, we examined an Oxford Nanopore assembly finding five copies, and another assembly using Bionano Irys finding three copies (Solares, et al. 2018). In the case of the Nanopore assembly, up close examination of 112 sequencing reads associated with the *Sdic* region found no evidence of any of them spanning the whole region (from *AnxB10* to *sw*), providing no convincing indication that the recapitulation of the *Sdic* region was done reliably. Additionally, we performed a read-depth analysis using CNVnator (Abyzov, et al. 2011), finding a normalized read depth compatible with 6 copies (see below and Material and Methods). Collectively, we concluded that the *Berlin* assembly should be used as a reference for the *Sdic* region.

Assemblies with a fragmented *Sdic* region

In three assemblies of the DSPR panel (A2, A6, and B4), we found the *Sdic* region fragmented. Fragmentation was associated with the presence of assembly gaps, which were not supported by further scrutiny of individual SMRT reads associated with the *Sdic* region as we found reads that precisely recover the stretches that presumably correspond to the assembly gaps. In the case of A2, and in addition to examining the reads associated with this region, we performed *in situ* hybridization on mitotic chromosomes finding a single signal, which indicated that the clustering of the *Sdic* copies at two different sites of the *X* chromosome is an assembly artifact (fig. S2.2).

Benchmarking of CNVnator

First, we examined under which conditions CNVnator (Abyzov, et al. 2011) can provide reliable CN estimates given the complexity of the *Sdic* region, *i.e.* the presence of multiple copies with high sequence identity among themselves, as well as with their flanking single-copy parental genes, *AnxB10* and *sw*. To this end, we used arguably the most reliable assemblies so far generated in *D. melanogaster*: GCA_000778455 (Berlin, et al. 2015) for ISO-1; and GCA_002300595.1 for A4 (Chakraborty, et al. 2018). First, we generated a set of synthetic *X* chromosomes for the A4 and the ISO-1 strains in which different *ad hoc* modifications were implemented, *i.e.* deleting all but one *Sdic* copy and the parental genes. A separate synthetic *X* chromosome was generated for each *Sdic* copy in both strains, five from A4 and six from ISO-1, which were used as references for read-depth analysis. The average read-depth values obtained with the sequencing data of the A4 strain were 6.18 and 6.16 when using the A4 and the ISO-1 synthetic reference chromosomes, respectively. Rounding off the average between both values to the closest integer and subtracting one because of the contribution of the reads from the parental genes, the estimated number of *Sdic* copies in A4 is 5. Following the same rationale with the ISO-1 strain, the estimated number of copies was 6 (average read-depth values were 7.38 and 6.73 when using the A4 and the ISO-1 synthetic reference chromosomes, respectively). These estimated numbers are identical to the number of copies found by annotating the indicated assemblies. For 13 strains of the *Drosophila* Synthetic Population Resources (King, et al. 2012b) and OR-R, the average read depth values across the five and six reference genomes from A4 and ISO-1, respectively, were highly correlated (r^2 = 0.73, P < 0.0001; fig. 2.2B; supplementary table S2.3). Further, we also examined whether sequence coverage could be positively correlated with CN estimates, finding no evidence. Specifically, we parsed this association in two sets of strains, with the first including 70 datasets from the Global Diversity Lines (Grenier, et al. 2015), and the second including 63 datasets from a Zambian population (Lack, et al. 2016a). For the first set, r^2 = 0.0008 (P = 0.8198) and for the second r^2 = 0.0055 (P = 0.5661).

Calibration of qPCR assays

Our control experiments with *sw* confirmed our ability to discern between 1, 2, and 3 copies (supplementary table S2.2; fig. 2.3C). This variation in copy number for *sw* is associated with differences between males and females of the ISO-1 strain, males carrying 2 copies of endogenous *sw* as a result of an induced duplication of the region (2T and 4M; this work), males carrying 2 copies of a *sw* transgene on chromosome 2 upon making it homozygous (A- and E-; (Clifton, et al. 2017), and heterozygous females possessing 3 copies as a result of carrying one chromosome with the wildtype configuration for the *Sdic* region and another chromosome with its duplicated version (II and IV in fig. 2.3B). For these same genotypes, the estimates about the number of copies of *Sdic* were also identical to the expectation (fig. 2.3D): 6 and 12 copies for the males and females of the ISO-1 strain, respectively; 0 copies for the males that carry the deletion of the *Sdic* region (A^- and E^-; (Clifton, et al. 2017)); 12 copies for the males that carry the duplication of the *Sdic* region (4M; this work); and 6, 12, or 18 copies in particular progenies from controlled crosses involving w^{1118}, 2T, and 4M (I-IV in fig. 2.3B). The only exception to this good agreement was the estimate for the males from the duplication strain 2T, for which the qPCR estimate was of 12.5 copies instead of 12. Collectively, these results are consistent with a suitable ability to infer the number of *Sdic* copies through our qPCR assay at least between 0 and 18.

Patterns of nucleotide variation across the *Sdic* repeat

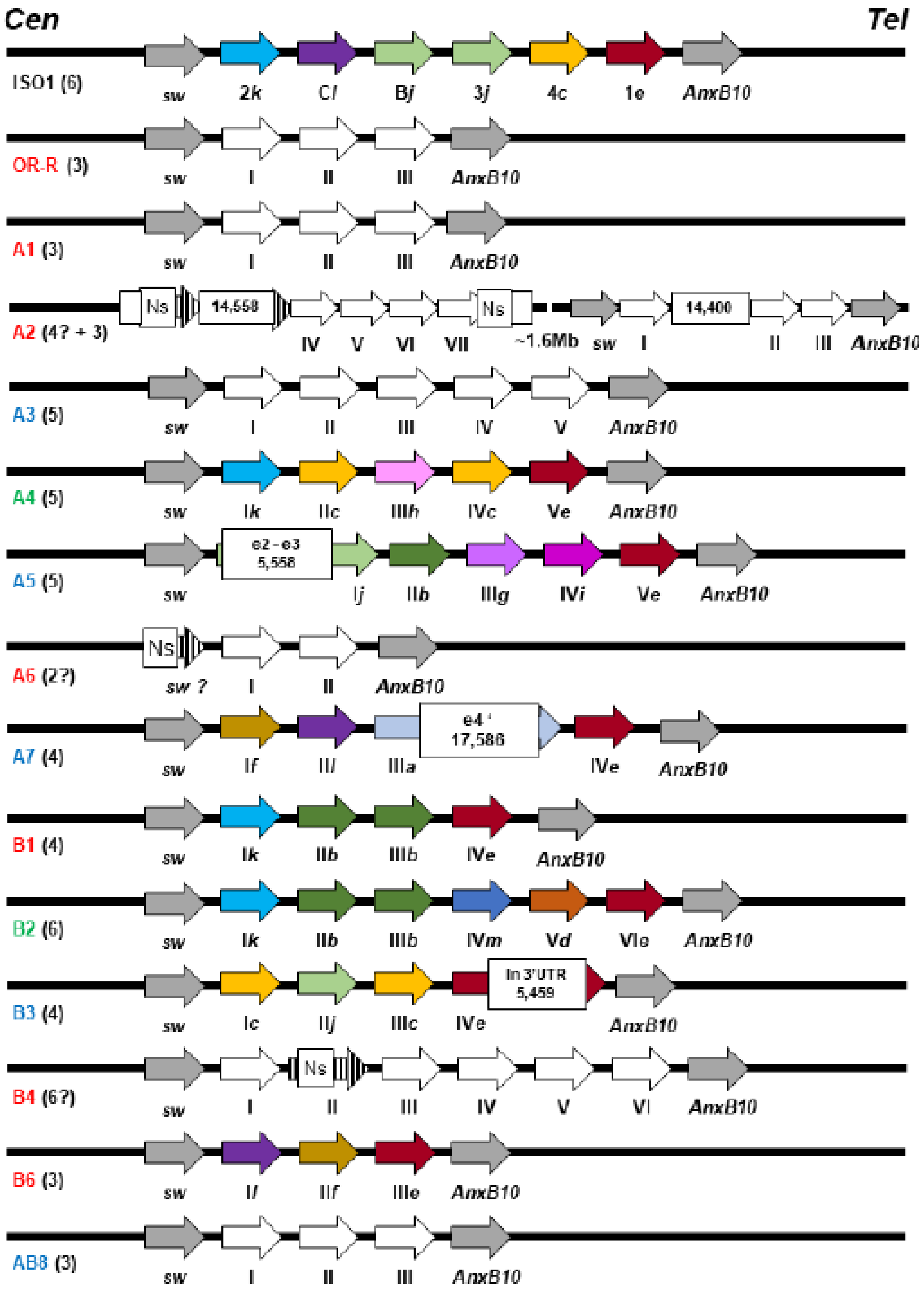
For the fraction of each *Sdic* repeat that corresponds to the *Sdic* transcriptional unit, the magnitude of within-strain pairwise sequence identity at the nucleotide level was very similar across the strains considered, with median sequence identity values ranging from 98.62% (B3) to 99.53% (B2); 99.44% when all 31 copies are considered jointly (supplementary table S2.11). Nevertheless, nucleotide differences in the two exons most proximal to the *Sdic* stop codon result in notable differences at the amino acid level, impacting the length of the putatively encoded variants as previously documented in the ISO-1 strain (Clifton, et al. 2017). These variants varied by up to 29% in length (388-544 residues; supplementary table S2.13). Further, and also within the *Sdic* transcriptional unit, there are 112 nt corresponding to the presumed *Sdic* promoter (Nurminsky, et al. 1998). We found two additional promoter sequences in relation to the two previously documented (Clifton, et al. 2017). Both additional promoters show nucleotide differences at the same two sites already known to vary among previously delineated promoter sequences of *Sdic* (Clifton, et al. 2017).

We also examined the level of nucleotide differentiation at other sequence intervals that are part of the *Sdic* repeat, *i.e.* ~1,800 nt corresponding to a combination of non-deleted intervals of the canonical sequence of the TE *Rt1c*, and a ~850 nt portion corresponding to the presumed pseudogene *AnxB10*-like (fig. 2.3A). We found a striking degree of conservation (number of base differences per site assuming a Jukes-Cantor substitution model; TE *Rt1c*, d = 0.005; *AnxB10*-like, d = 0.002; *Sdic* exonic sequence, d = 0.007). For *AnxB10*-like, we examined the possibility that this strong nucleotide conservation could actually reflect functional constraints contrary to previous reports (Yeh, et al. 2012a). By using an in-home pipeline that tracks small sequence motifs to differentiate expression between very similar duplicated sequences (Clifton, et al. 2017), we screened RNA-seq datasets corresponding to 29 biological conditions (Material and Methods; supplementary table S2.12), finding no evidence of *AnxB10*-like expression. In the absence of evidence for functionality, the high-level sequence conservation for these intervals of the *Sdic* repeat might be suggestive of structural constraints.

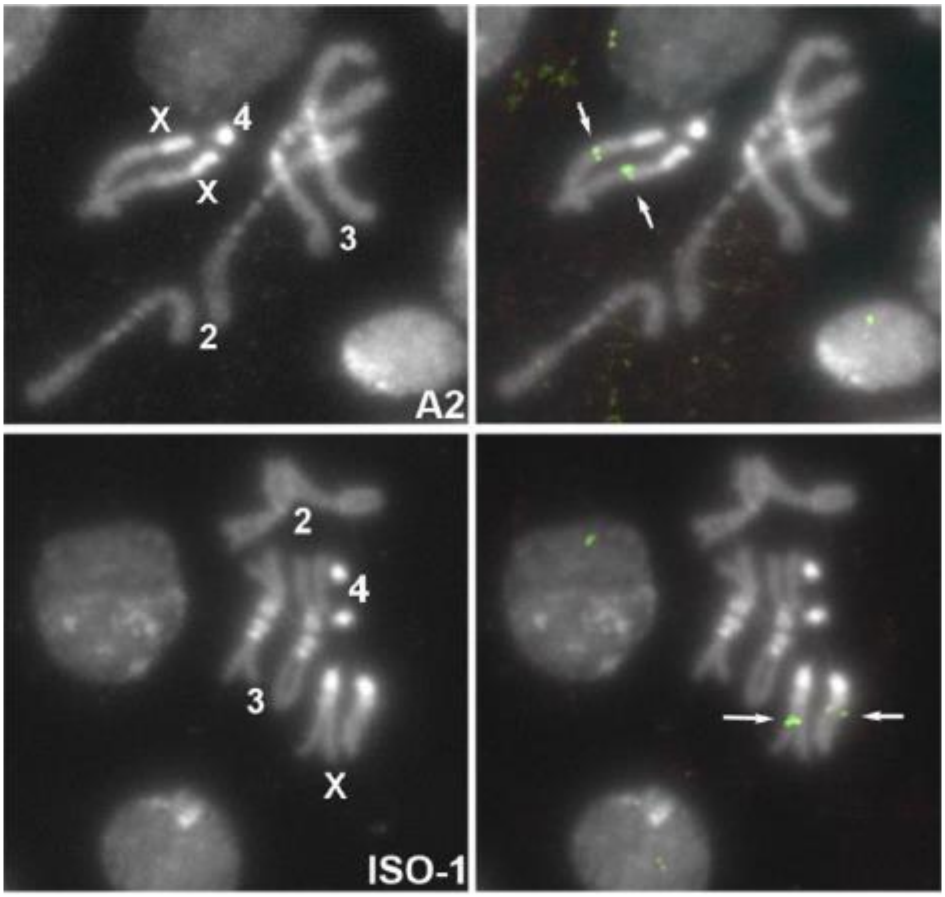
Detecting positive selection across the *Sdic* repeat

Several approaches were used to determine the pattern of sequence evolution across the *Sdic* repeat taking into account the presence of both coding and noncoding sequences. The first method was used to test if positive selection occurred on *Sdic* protein-coding sequences (*i.e*. whether there is proportion of sites with an excess of nonsynonymous substitutions in relation to the expectation under a neutral model) for each branch of the phylogeny. In this model, the number of site classes with a particular nonsynonymous to synonymous rate ratio (ω) in each branch is not fixed but estimated using a small sample AIC. Then, a likelihood-ratio test (LRT) was used to compare the positive selection to the null model (classes with $\omega > 1$ are not allowed), and the *p*-value for each branch was corrected for multiple testing using the Holm-Bonferroni correction

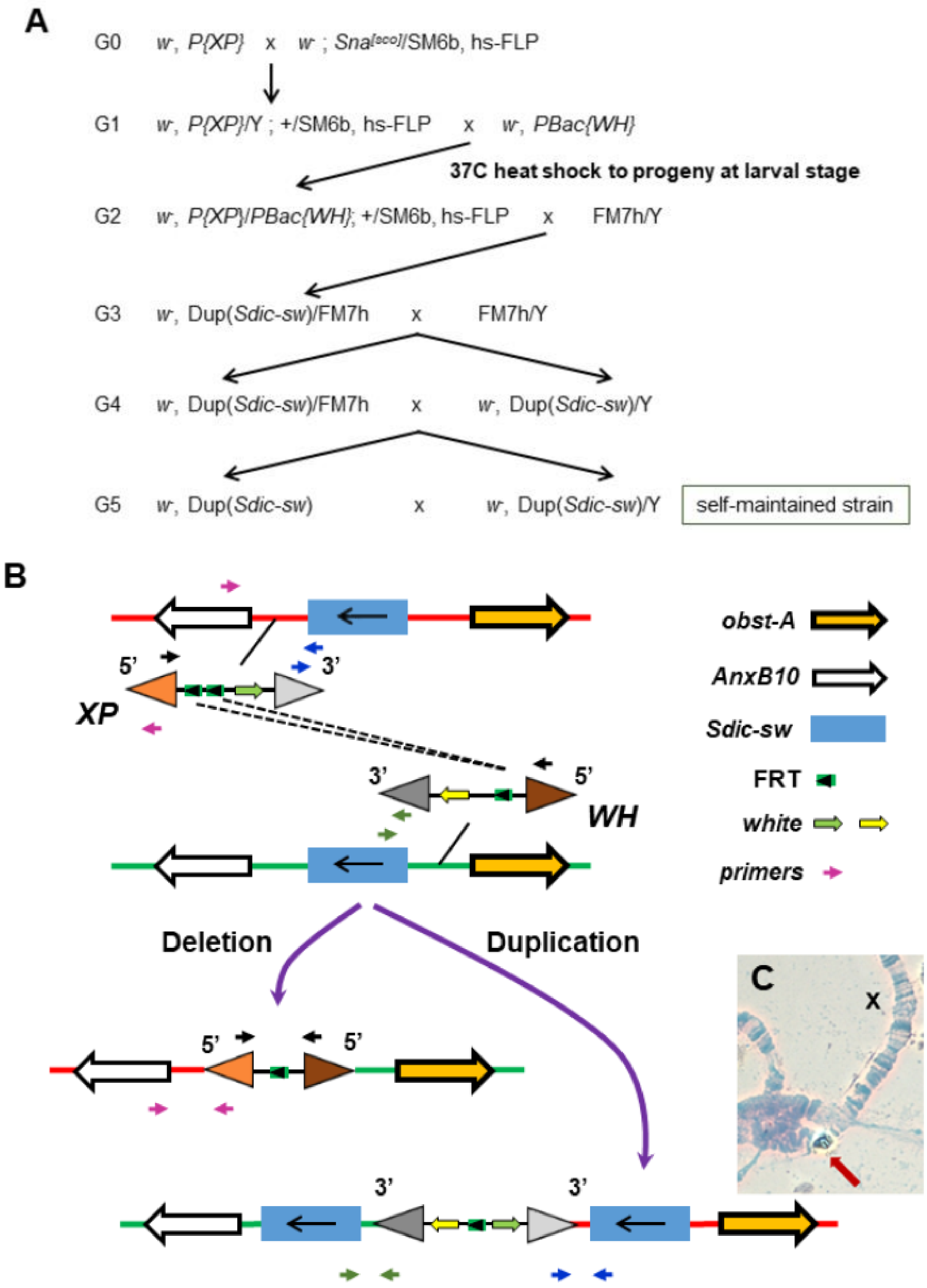
(Holm 1979). Similarly, the batch script for detecting positive selection on noncoding sites evaluates whether the substitution rate in this class of sites exceeds significantly a neutral class of sites (here represented by the synonymous sites). In this case, under the null model, the number of noncoding site classes for each branch is set to three: (i) those that are selectively neutral; (ii) those evolving under purifying selection; and (iii) those completely constrained in background lineages (BG) or neutrally evolving in foreground lineages (FG). In the alternate model, this third class of sites is forced to evolve under positive selection in the foreground lineages, and an extra class of sites, neutrally evolving in BG and positively selected in FG, is added. Thus, under this configuration, the relaxation of purifying selection at some sites is already accounted for by the null model. The LRT was used to compare these two nested models by setting each of the branches of the *Sdic* tree (reconstructed using RAxML and MSA positions) as a background lineage in an independent test. Final *p*-values were adjusted for multiple comparisons using the False Discovery Rate (FDR) correction (Benjamini and Hochberg 1995).



Supplementary Figure S2.1. Annotation of the *Sdic* region across 13 populations of the DSPR panel and the wild-type stock OR-R. The most reliable organization of the region at 19C1 on the *X* chromosome in the ISO-1 is provided as a reference (Clifton, et al. 2017). The region is depicted from centromere (Cen) to telomere (Tel), including the flanking genes *sw* and *AnxB10* (grey filled arrows). Population names are color-coded based on the broad continental region where they were collected: green, Africa; red, Americas; and blue, Eurasia. The number of annotated *Sdic* copies in reference-quality genome assemblies (Chakraborty, et al. 2019; Chakraborty, et al. 2018) is indicated in parentheses next to the name of the population. Arrows filled with vertical lines are partial copies. *Sdic* copies in the ISO-1 strain are named as reported (Clifton, et al. 2017). In the rest of populations, the copy identifiers are roman numerals according to their relative order from *sw* to *AnxB10*. In the most reliable genome assemblies, copies are color coded, and a lower character (*a-m*) added to their identifier, both indicating the associated paratype. The size of the TE insertions (solid boxes), as well as their location, are indicated. Ns, assembly gap. e, exon. The apostrophe in the case of A7_III*a* indicates a no longer coding exon, as the STOP codon is upstream of the TE insertion. The *Sdic* region was found unfragmented except for the strains from Bogota (A2) and Georgia (A6). In the case of A2, 6 full *Sdic* copies form two different clusters ~1.6 Mb apart on the *X* chromosome. The distal cluster harbors 3 of the copies, which are flanked by *sw* and *AnxB10.* In contrast, the proximal cluster is flanked by gap assemblies, which in turn are adjacent to TEs. Within this cluster, we found 3 full copies, another copy almost in its entirety, and the remnants of 2 other copies, which are separated by a cluster of TEs. In the case of A6, only two complete *Sdic* copies were found, upstream of which a 1,190 nt long fragment corresponding to the 3' end of either *sw* or an *Sdic* copy is present. This fragment is separated from other genes further upstream of the parental gene *sw,* such as *obst-A*, which is not found in the assembly due to an assembly gap.

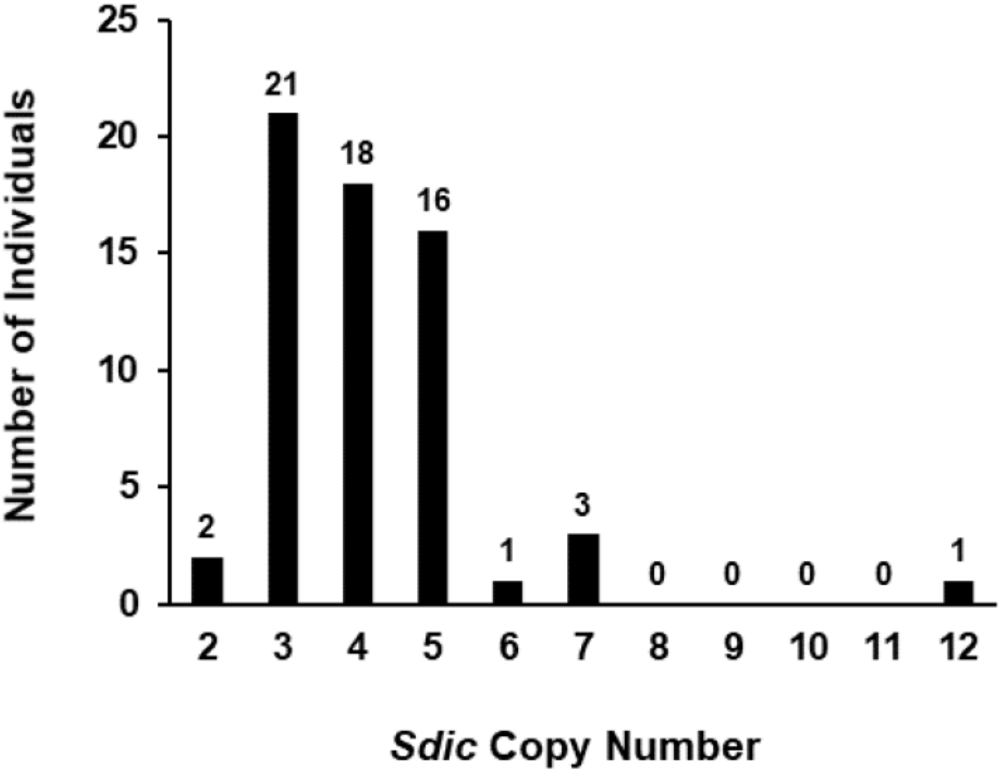


Supplementary Figure S2.2. *In situ* hybridization on mitotic chromosomes of A2. A single hybridization signal (arrow) on the *X* chromosome is observed both for A2 (top right) and ISO-1 (bottom right) strains. The squashes shown were obtained from female larvae; squashes from male larvae show the same result discarding any additional copy on the *Y* chromosome.

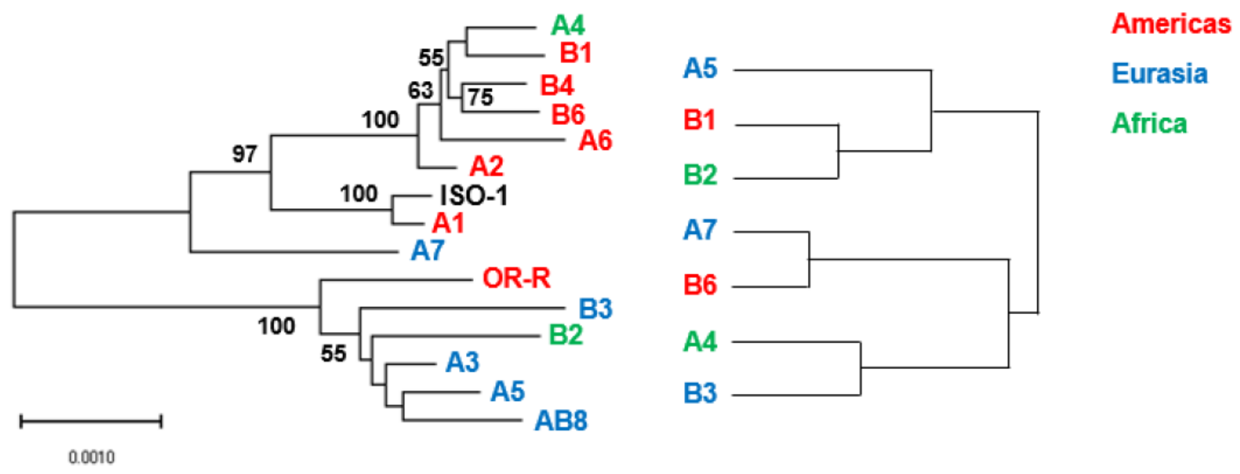

A
G0 w-, P{XP} x w- ; Sna[sco]/SM6b, hs-FLP
G1 w-, P{XP}/Y ; +/SM6b, hs-FLP x w-, PBac{WH}
37C heat shock to progeny at larval stage
G2 w-, P{XP}/PBac{WH}; +/SM6b, hs-FLP x FM7h/Y
G3 w-, Dup(Sdic-sw)/FM7h x FM7h/Y
G4 w-, Dup(Sdic-sw)/FM7h x w-, Dup(Sdic-sw)/Y
G5 w-, Dup(Sdic-sw) x w-, Dup(Sdic-sw)/Y
self-maintained strain
B
5'
3'
XP
3'
5'
WH
Deletion
Duplication
5'
5'
3'
3'
obst-A
AnxB10
Sdic-sw
FRT
white
primers
C
X

Supplementary Figure S2.3. Duplicating the *Sdic* region. (A) Mating scheme followed to duplicate the *Sdic* region through an induced FRT-FLP recombination event. The recombination event took place between the engineered TEs *P{XP}d03903* and *PBac{WH}f02348*, following the same mating scheme used to previously generate the deletion of the *Sdic* region (Yeh, et al. 2012b).

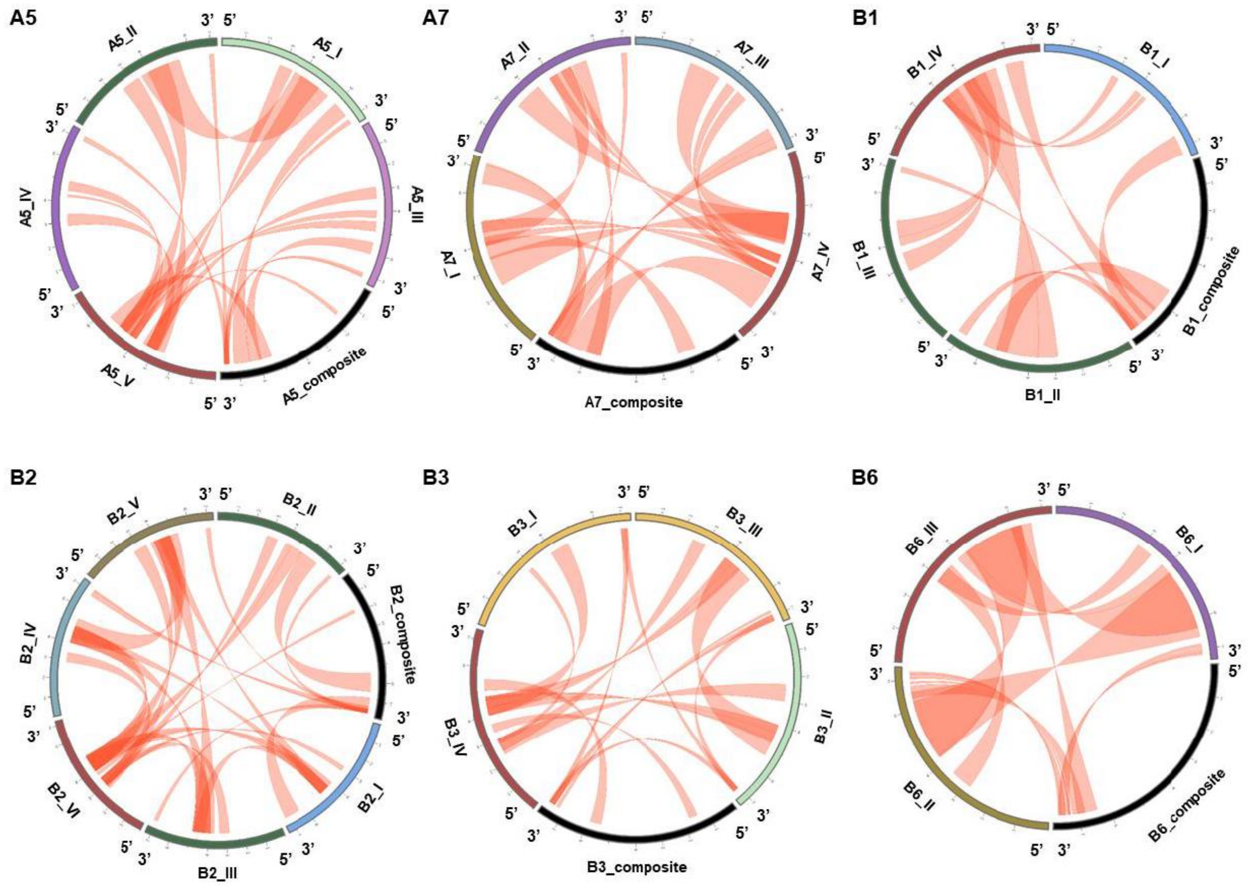
The TE *P{XP}d03903* is located in the intergenic region between the genes *AnxB10* and *Sdic1* while *PBac{WH}f02348* is between *sw* and *obst-A.* Therefore, the actual duplication spans from the *Sdic* copy adjacent to *AnxB10* to *sw*, inclusive. To discern which females, out of 174 obtained in G3, were actual carriers of the duplication of the *Sdic* region, we used two approaches. First, we visually inspected and PCR-screened the male progeny of 174 females, classifying each female as a duplication or non-duplication bearer based on the eye color of their male progeny. Male progeny was PCR-screened through four controls that provided complementary information (supplementary table S2.4). Once females carrying the duplication were identified, the mating scheme was continued to make the duplication homozygous. (B) Gene and TE molecular organization along the original TE-bearing chromosomes (top) and those resulting from an ectopic recombination event (bottom). As a duplication event results into a hybrid TE carrying only the 3' ends of the two TEs, two controls (amplicons 1 and 2 respectively in supplementary table S2.4) were designed to confirm their presence in the PCR screening. Male progeny of these females should give rise to two amplicons (one per 3' end), which were multiplexed in the same PCR reaction. After separating the females presumably carrying the duplication of the *Sdic* region from those carrying its deletion, the females were subjected to two additional PCR controls (amplicons 3 and 4 respectively in supplementary table S2.4). Amplicon 3 allows to confirm that the *X* chromosome under examination does not carry the deletion or the original chromosome carrying *P{XP}d03903*; the amplicon corresponding to the downstream end of the duplicated *Sdic* region should not be detected. Lastly, a fourth amplicon that corresponds to the hybrid TE that preserves the 5' ends of the two original chromosomes, and should only result from a deletion event, should not be observed either (Yeh, et al. 2012b). The combination from these four PCR controls designated 36 females out of the initial 174 as carriers of a duplicated *Sdic* region. Two of them (2T and 4M) were used in downstream analyses. (C) Chromosomal location of the duplicated *Sdic* region. An extremely intense, single *in situ* hybridization signal can be detected on the *X* chromosome of one of the duplication strains (4M), denoting a local duplication of the *Sdic* region, which is in good agreement with qPCR results.



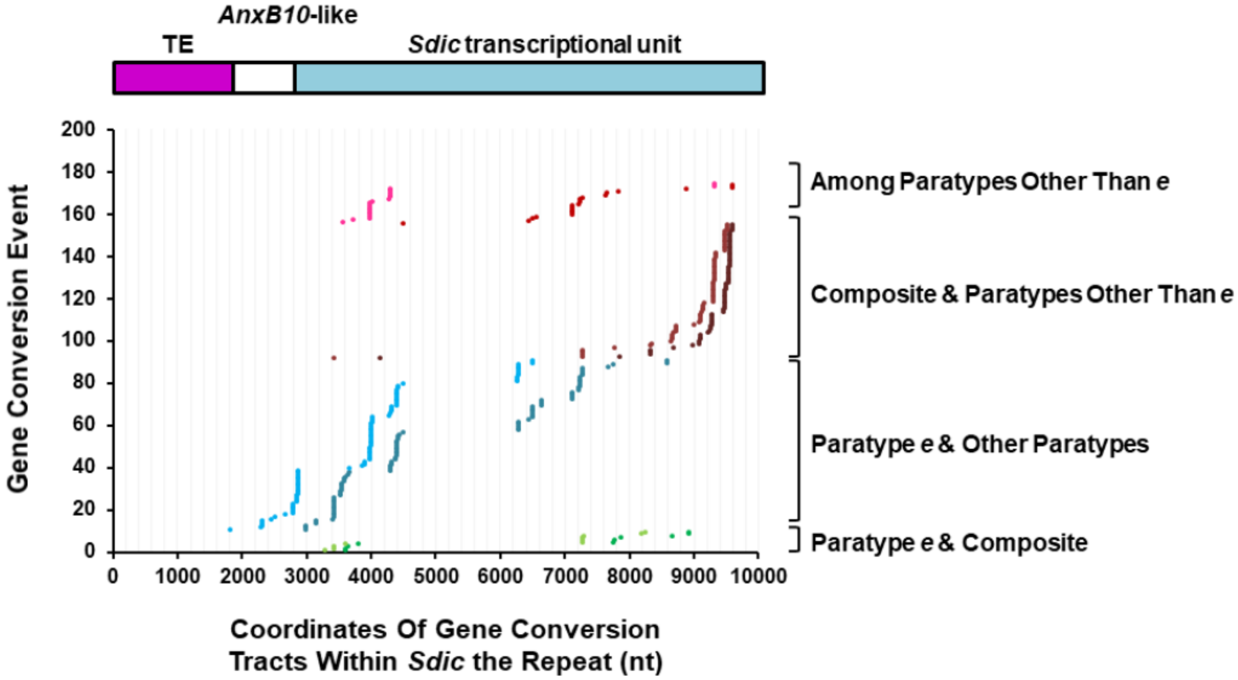
Supplementary Figure S2.4. Frequency distribution of *Sdic* CN estimation in haploid embryo genomes from a Zambian population. Each genome dataset corresponds to one female gamete from each strain. Sixty-two haploid embryos were ultimately considered (Material and Methods). The CNVnator program (Abyzov, et al. 2011) was utilized to calculate read-depth average values across a set of synthetic reference genomes derived from A4. The round-off read-depth average values are shown.



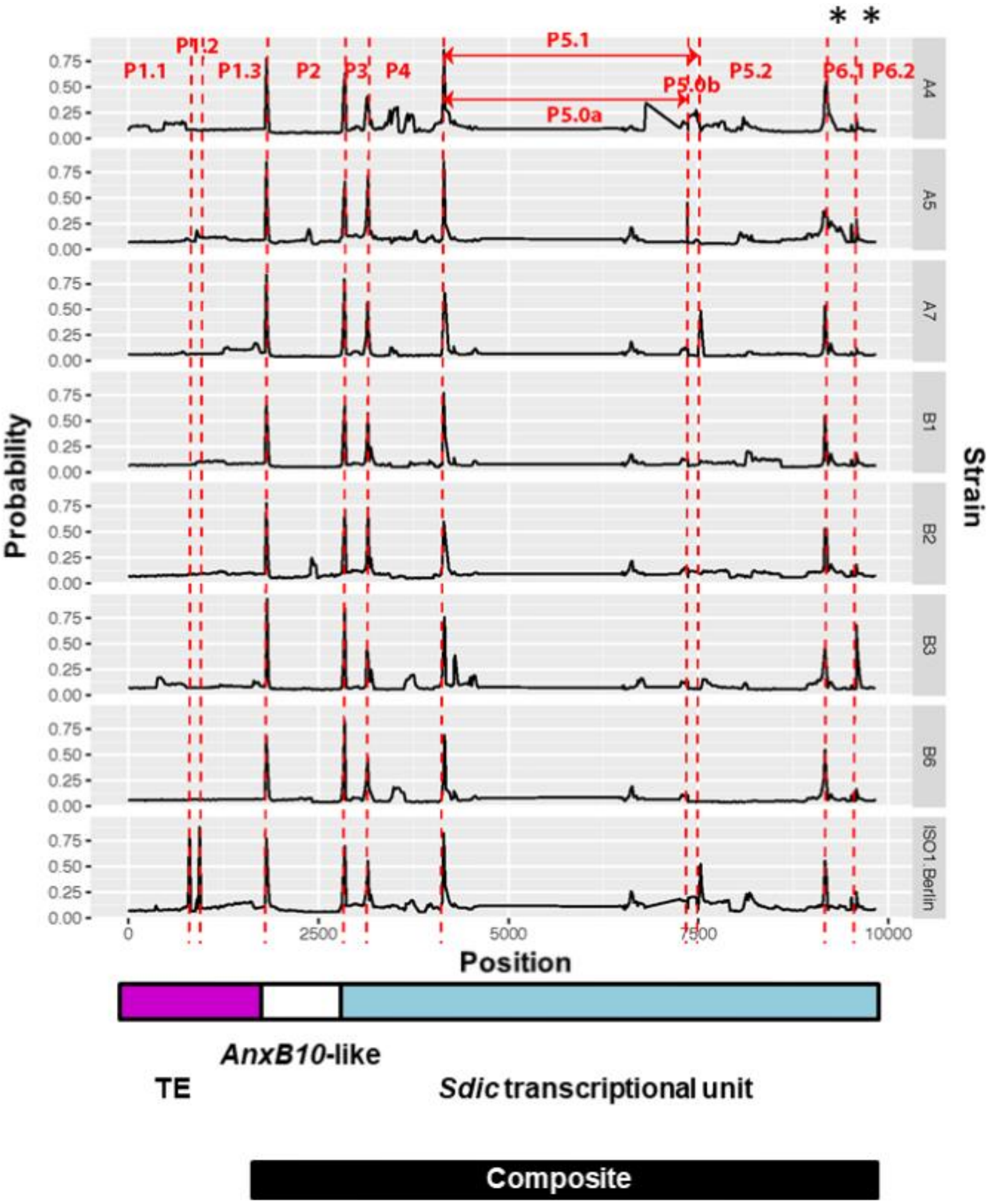
Supplementary Figure S2.5. Relationship among strains from the DSPR panel. Left, mtDNA phylogeny of the 14 strains for which their *Sdic* region was annotated in this study plus the reference strain ISO-1. The phylogenetic relationship among strains was inferred by using the Maximum Likelihood method. The tree with the highest log likelihood (-24205.28) is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1,000 replicates) is shown next to the branches when higher than the cut-off value of 50. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The continental site of collection for each strain is color-coded as indicated in the legend. The reference strain ISO-1 is shown in black. Right, hierarchical clustering of populations based on their paratype and CN composition. Three principal components that explain ~90% of the total variation were used. The observed patterns of compositional similarity for the *Sdic* region were coincidental with the sorting of the populations based on the Bray-Curtis index (Bray and Curtis 1957), a metric typically used to assess compositional similarity between, for example, two ecological communities based on count data (data not shown). The resulting clustering matches neither the geographical proximity of the collection site of the strains nor, more importantly, the phylogenetic relationship of the populations.



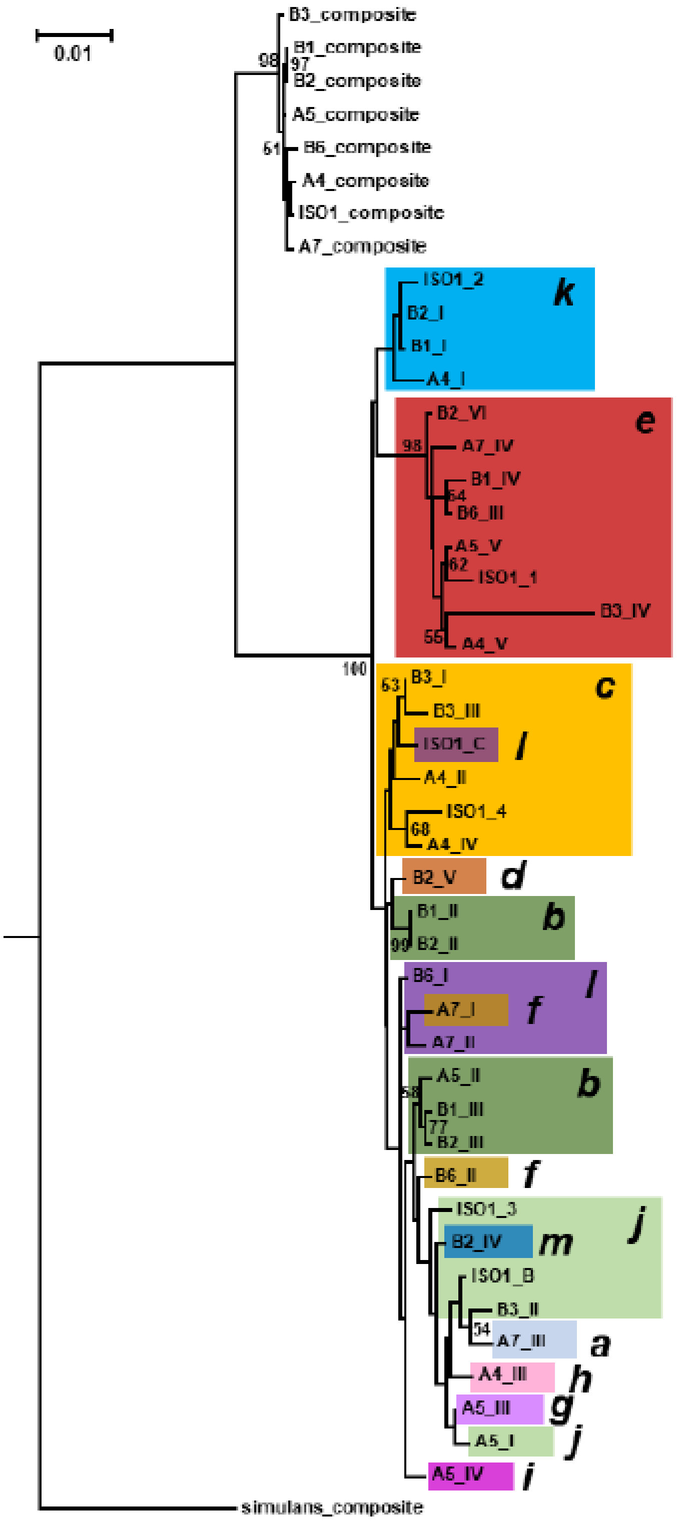
Supplementary Figure S2.6. Topology of gene conversion events that have occurred in the *Sdic* region. Circular layouts showing the patterns of gene conversion events occurred between *Sdic* copies and the composite, *i.e.* the fragments from *sw* plus *AnxB10* (black) that align with *Sdic*. The results from GenConv (Sawyer 1989) are graphed for the six strains not shown in fig. 2.4C.



Supplementary Figure S2.7. Topological occurrence of gene conversion events along *Sdic* repeats. Plot of coordinates of 174 gene conversion tracts involving different *Sdic* paratypes and the parental gene *sw* as detected with GeneConv across strains. Coordinates for different types of events are color-coded (see legend on the right). Start and end coordinates, in lighter and darker tones respectively, of the same gene conversion event project onto the same value of the *y*-axis. All tracts detected across strains are shown. Outer events (or fragments according to the nomenclature of GeneConv) are not shown.

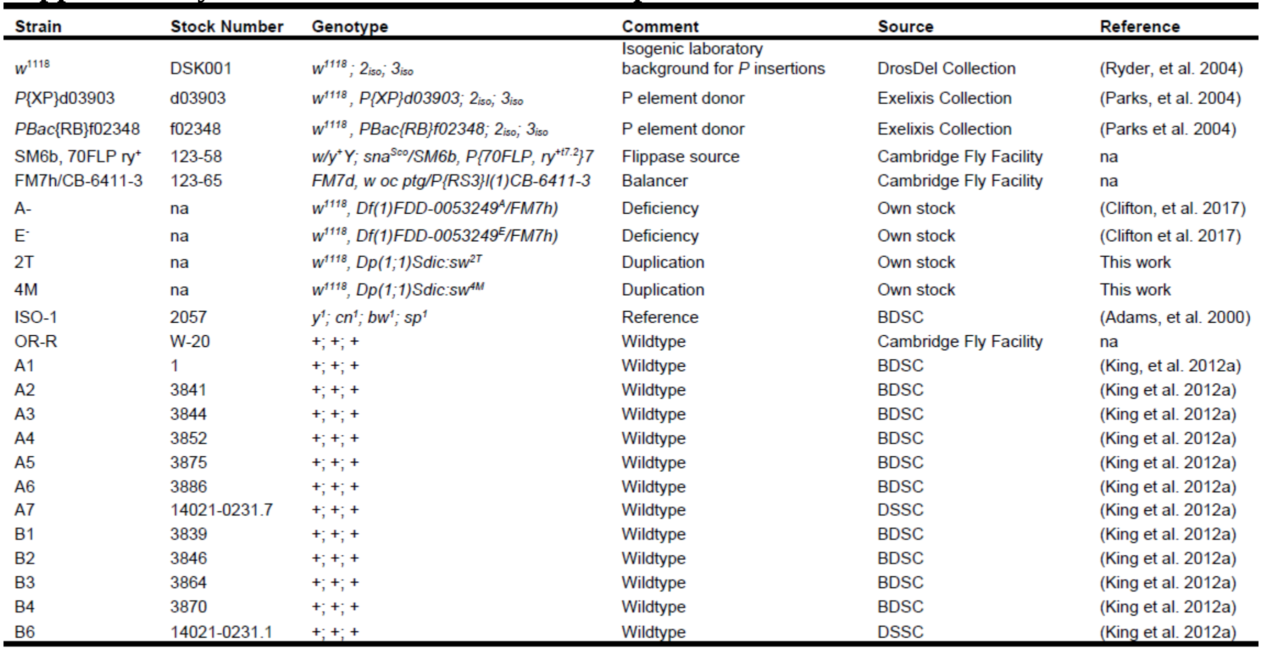


Supplementary Figure S2.8. Breakpoint distribution along the *Sdic* repeat across the strains. Breakpoint location inferred with AGC (O'Fallon 2013). The location of highly supported breakpoints is indicated with a dotted red line and the resulting partitions numbered accordingly from 5’ to 3’ (P1-P6). The partitions for which there is strong evidence of the action of positive selection are indicated with asterisks (top). Distance in nucleotides relative to the 5’ end of the *Sdic* repeat can be interpolated from the *x*-axis. The composite, *i.e.* the fragments from *sw* plus *AnxB10* (black) that align with *Sdic,* is shown at the bottom. *y*-axis, probability of breakpoint occurrence.



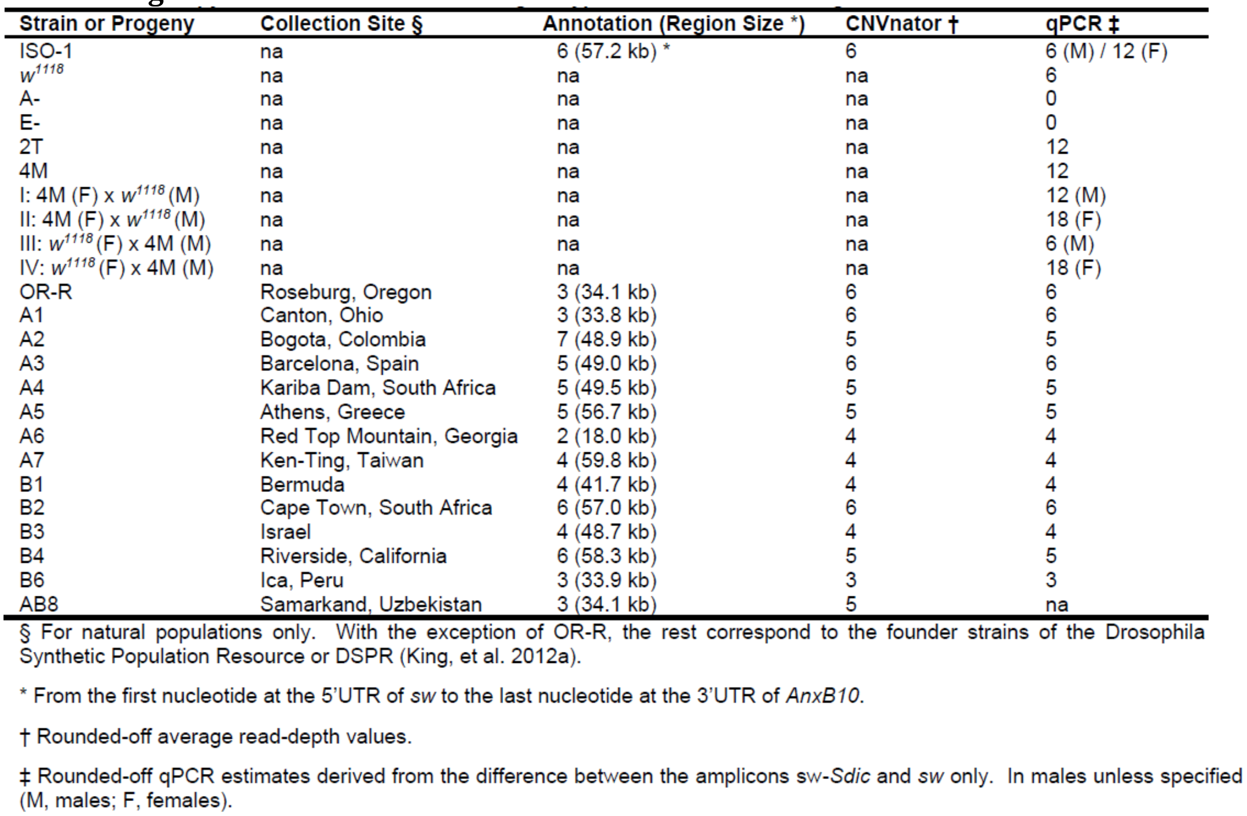
Supplementary Figure S2.9. Phylogenetic relationships among *Sdic* copies. The copies considered are the 31 from the strains A4, A5, A7, B1, B2, B3, and B6, plus the six copies from the reference strain ISO-1. Copy nomenclature is as in fig. 2.1; also copies that belong to the same paratype are shaded according to the color code in that same figure. The phylogeny shown was inferred with RAxML 8.1.2 under a GTRGamma model of sequence evolution. The composites, *i.e.* the constructs generated with the alignable stretches of DNA sequence between *Sdic* and the parental genes *sw* and *AnxB10*, from each strain were also included in the analysis. The equivalent composite was generated for *D. simulans* according to the available information in FlyBase (Hu, et al. 2013). The percentage of replicate trees in which the associated copies clustered together in the bootstrap test (1,000 replicates) is shown next to the branches when higher than the cut-off value of 50. Copies representing paratype *e*, the ones for which is found the strongest support for the most recent action of positive selection, form a very distinctive clade. The same conclusion and a very similar overall tree topology are found when inferring the phylogeny of the copies under a best-fit substitution model.

Supplementary Table S2.1. Strains used in empirical work



Strain	Stock Number	Genotype	Comment	Source	Reference
w^{1118}	DSK001	w^{1118}; 2_{iso}; 3_{iso}	Isogenic laboratory background for *P* insertions	DrosDel Collection	(Ryder, et al. 2004)
P{XP}d03903	d03903	w^{1118}, *P{XP}d03903*; 2_{iso}; 3_{iso}	P element donor	Exelixis Collection	(Parks, et al. 2004)
PBac{RB}f02348	f02348	w^{1118}, *PBac{RB}f02348*; 2_{iso}; 3_{iso}	P element donor	Exelixis Collection	(Parks et al. 2004)
SM6b, 70FLP ry^{+}	123-58	*w/*y^{+}*Y; *sna^{Sco}*/SM6b, P{70FLP, *$ry^{+t7.2}$*}7*	Flippase source	Cambridge Fly Facility	na
FM7h/CB-6411-3	123-65	*FM7d, w oc ptg/P{RS3}l(1)CB-6411-3*	Balancer	Cambridge Fly Facility	na
A-	na	w^{1118}, *Df(1)FDD-*0053249^{A}*/FM7h)*	Deficiency	Own stock	(Clifton, et al. 2017)
E^{-}	na	w^{1118}, *Df(1)FDD-*0053249^{E}*/FM7h)*	Deficiency	Own stock	(Clifton et al. 2017)
2T	na	w^{1118}, *Dp(1;1)Sdic:*sw^{2T}	Duplication	Own stock	This work
4M	na	w^{1118}, *Dp(1;1)Sdic:*sw^{4M}	Duplication	Own stock	This work
ISO-1	2057	y^{1}; cn^{1}; bw^{1}; sp^{1}	Reference	BDSC	(Adams, et al. 2000)
OR-R	W-20	+; +; +	Wildtype	Cambridge Fly Facility	na
A1	1	+; +; +	Wildtype	BDSC	(King, et al. 2012a)
A2	3841	+; +; +	Wildtype	BDSC	(King et al. 2012a)
A3	3844	+; +; +	Wildtype	BDSC	(King et al. 2012a)
A4	3852	+; +; +	Wildtype	BDSC	(King et al. 2012a)
A5	3875	+; +; +	Wildtype	BDSC	(King et al. 2012a)
A6	3886	+; +; +	Wildtype	BDSC	(King et al. 2012a)
A7	14021-0231.7	+; +; +	Wildtype	DSSC	(King et al. 2012a)
B1	3839	+; +; +	Wildtype	BDSC	(King et al. 2012a)
B2	3846	+; +; +	Wildtype	BDSC	(King et al. 2012a)
B3	3864	+; +; +	Wildtype	BDSC	(King et al. 2012a)
B4	3870	+; +; +	Wildtype	BDSC	(King et al. 2012a)
B6	14021-0231.1	+; +; +	Wildtype	DSSC	(King et al. 2012a)

Supplementary Table S2.2. ***Sdic*** **copy number estimates across 22 genotypes and three methodologies**



Strain or Progeny	Collection Site §	Annotation (Region Size *)	CNVnator †	qPCR ‡
ISO-1	na	6 (57.2 kb) *	6	6 (M) / 12 (F)
w^{1118}	na	na	na	6
A-	na	na	na	0
E-	na	na	na	0
2T	na	na	na	12
4M	na	na	na	12
I: 4M (F) x w^{1118} (M)	na	na	na	12 (M)
II: 4M (F) x w^{1118} (M)	na	na	na	18 (F)
III: w^{1118} (F) x 4M (M)	na	na	na	6 (M)
IV: w^{1118} (F) x 4M (M)	na	na	na	18 (F)
OR-R	Roseburg, Oregon	3 (34.1 kb)	6	6
A1	Canton, Ohio	3 (33.8 kb)	6	6
A2	Bogota, Colombia	7 (48.9 kb)	5	5
A3	Barcelona, Spain	5 (49.0 kb)	6	6
A4	Kariba Dam, South Africa	5 (49.5 kb)	5	5
A5	Athens, Greece	5 (56.7 kb)	5	5
A6	Red Top Mountain, Georgia	2 (18.0 kb)	4	4
A7	Ken-Ting, Taiwan	4 (59.8 kb)	4	4
B1	Bermuda	4 (41.7 kb)	4	4
B2	Cape Town, South Africa	6 (57.0 kb)	6	6
B3	Israel	4 (48.7 kb)	4	4
B4	Riverside, California	6 (58.3 kb)	5	5
B6	Ica, Peru	3 (33.9 kb)	3	3
AB8	Samarkand, Uzbekistan	3 (34.1 kb)	5	na

§ For natural populations only. With the exception of OR-R, the rest correspond to the founder strains of the Drosophila Synthetic Population Resource or DSPR (King, et al. 2012a).

* From the first nucleotide at the 5'UTR of *sw* to the last nucleotide at the 3'UTR of *AnxB10*.

† Rounded-off average read-depth values.

‡ Rounded-off qPCR estimates derived from the difference between the amplicons sw-*Sdic* and *sw* only. In males unless specified (M, males; F, females).

Supplementary Table S2.3. Normalized read-depth values obtained with CNVnator for the strains of the DSPR panel

	Reference Genome *											Rounded Off	CN
Strain	**ISO1_1**	**ISO1_2**	**ISO1_3**	**ISO1_4**	**ISO1_5**	**ISO1_6**	**A4_1**	**A4_2**	**A4_3**	**A4_4**	**A4_5**	**Average**	**Estimate †**
ISO1	7.401	3.173	7.447	7.429	7.523	7.385	7.543	6.675	7.613	7.555	7.527	7	6
ORR	7.678	7.831	7.784	7.725	2.359	2.302	7.485	7.498	7.505	7.490	7.513	7	6
A1	6.794	6.879	6.864	6.801	6.805	6.582	6.476	6.504	6.470	6.467	6.143	7	6
A2	6.053	6.115	6.098	6.047	6.121	4.748	5.905	5.930	5.930	5.894	5.743	6	5
A3	6.769	6.798	6.724	6.705	6.797	5.224	6.707	6.703	6.705	6.706	6.345	7	6
A4	6.430	6.533	5.639	6.461	5.665	6.243	6.207	6.248	6.259	6.200	5.984	6	5
A5	6.064	6.177	6.141	6.110	6.175	5.927	5.859	5.893	5.868	5.851	5.640	6	5
A6	3.744	6.361	5.575	6.243	6.317	5.616	6.108	3.779	6.062	3.747	5.872	5	4
A7	5.249	5.303	5.288	5.251	5.296	5.145	5.145	4.756	4.828	4.703	4.405	5	4
B1	4.833	4.890	4.867	4.845	4.887	4.664	4.459	4.611	4.589	4.563	4.271	5	4
B2	7.069	7.253	7.224	7.168	7.259	6.821	6.903	7.004	6.941	6.948	6.630	7	6
B3	5.020	5.103	5.084	5.049	5.104	4.468	4.851	4.876	4.861	4.838	4.670	5	4
B4	5.983	6.044	5.934	5.989	6.047	5.487	5.852	5.879	5.674	5.847	5.592	6	5
B6	4.339	4.427	4.395	4.379	4.416	4.333	4.344	4.382	4.372	4.331	4.273	4	3
AB8	6.073	6.144	6.133	6.097	6.163	5.941	5.772	5.803	5.874	5.765	5.734	6	5

DSPR, Drosophila Synthetic Population Resource.

* Each reference genome used carries one single *Sdic* copy (either from the ISO-1 or A4 strain) and lacks the two parental flanking genes *sw* and *AnxB10*. The number in the ID of the reference genome denotes the order of the *Sdic* copy in the original strain between the flanking genes, specifically from *sw* to *AnxB10*. For example, A4_5 refers to the fifth copy starting from *sw*, *i.e.* the copy adjacent to *AnxB10* in the A4 strain.

† The CN estimate is calculated as the rounded off average minus one due to the contribution of reads from *sw* and *AnxB10* to the read-depth estimates obtained with CNVnator.

Supplementary Table S2.4. Primers used

Amplicon # (Description)	Forward Primer (5’-3’)	Reverse Primer (5’-3’)	Ta (C)	Size (nt)	Experiment
1 (Unaltered *WH*3’ end)	sw-WH_*F* TGTTTGATTAAAATGCTGAGTGTG	WH3’+ CCTCGATATACAGACCGATAAAAC †	59	621	Dup.
2 (Unaltered *XP*3’ end)	XP3’- TACTATTCCTTTCACTCGCACTTATTG †	XP-Sdic1_*R* TAGAACTACCCGCATATTTGATTG	59	300	Dup.
3 (Unaltered distal junction)	AnxB10-intron1_*F* TCTCTAGCCTGGCAATCCAATC	XP5’-*R* AGCCTTCCACTGCGAATCATT §	58	900	Dup.
4 (Hybrid TE deletion)	XP5’+ AATGATTCGCAGTGGAAGGCT *	WH5’- GACGCATGATTATCTTTTACGTGAC *	55	1,400	Dup.
5 (*Sdic/sw*)	AACGGATTCACCTCCAAGC	GATCTCGAGTGGTGTGATGG	60	93	qPCR
6 (*sw*)	GCGAGAAGGAGATCAAGGAC	CTGATCCTTGTCGATGCCTG	60	74	qPCR
7 (*TPI*)	AGGCAACTGGAAGATGAACG	GATGACCACCTCCGTGTTG	60	97	qPCR
8 (*Gapdh2*)	CAAGCAAGCCGATAGATAAAC *	GTCAAATCGACCACGGAAA *	52	762	qRT-PCR
9 (*Sdic*)	CGTATTCTACTTTGAGCGGCG	GGAATGTTCGTAGCCTGCAC	60	76	qRT-PCR
10 (*clot*)	GAGCGGGCATACTGGAAG	GCAACAGAGTGGGCAAGAAG	60	82	qRT-PCR
11 (*Sdic/sw*)	TGCAGTTTCCCCTGATTTCTT *	AGACGAAGAAGAACGCGTAATG *	54.3	2,253	In situ
12 (*Sdic/sw*)	CATTTGATGCCCAAGGAGAC	AGGAAGAGGTGGCCAAAGTC	60	1,434	FISH
13 (TE_A2_uj)	CAAGATGAACCAGAGCGATG	GCACTTGGCTGTCACAAGAG	60	684	PCR
14 (TE_A2_dj)	CACAAGCGGTTTCCTTTAGC	TTGGGCTCTTTCAGTTGAGG	60	716	PCR
15 (TE_A7_uj)	TCAATCCCAACCTGATCCTC	CACAAGCGGTTTCCTTTAGC	60	886	PCR
16 (TE_A7_dj)	CGCGTCAGCATTGTTCATAC	ACCTCCGTGTCTTGGTTGAG	60	610	PCR
17 (TE_A5_uj)	CAATCTGTCCATCCACATGC	ATTGCATTTGGCTAGCTTGG	60	404	PCR
18 (TE_A5_dj)	AGTCCAAGCTAGCCAAATGC	GGAGAGAAGGAGCATTGCAG	60	623	PCR
19 (TE_B3_uj)	AGCCGCTGTACTCCTTTGAG	CTGCCCTCTTTCAACGCTAC	60	748	PCR
20 (TE_B3_dj)	TGACTAAGGACAACGCCAAG	GCTTTATGCCGAAAGAGTCG	60	659	PCR

Dup., engineered duplication experiment. In situ, *in situ* hybridization on polytene chromosomes. FISH, *in situ* hybridization on mitotic chromosomes. Ta, annealing temperature. uj, upstream junction; dj, downstream junction. Unless indicated, primer design was performed in this study.

† (Parks, et al. 2004).

§ Reverse complementary of the XP5’- primer when combined with the primer WH5’+ (Parks, et al. 2004).

* (Yeh, et al. 2012b).

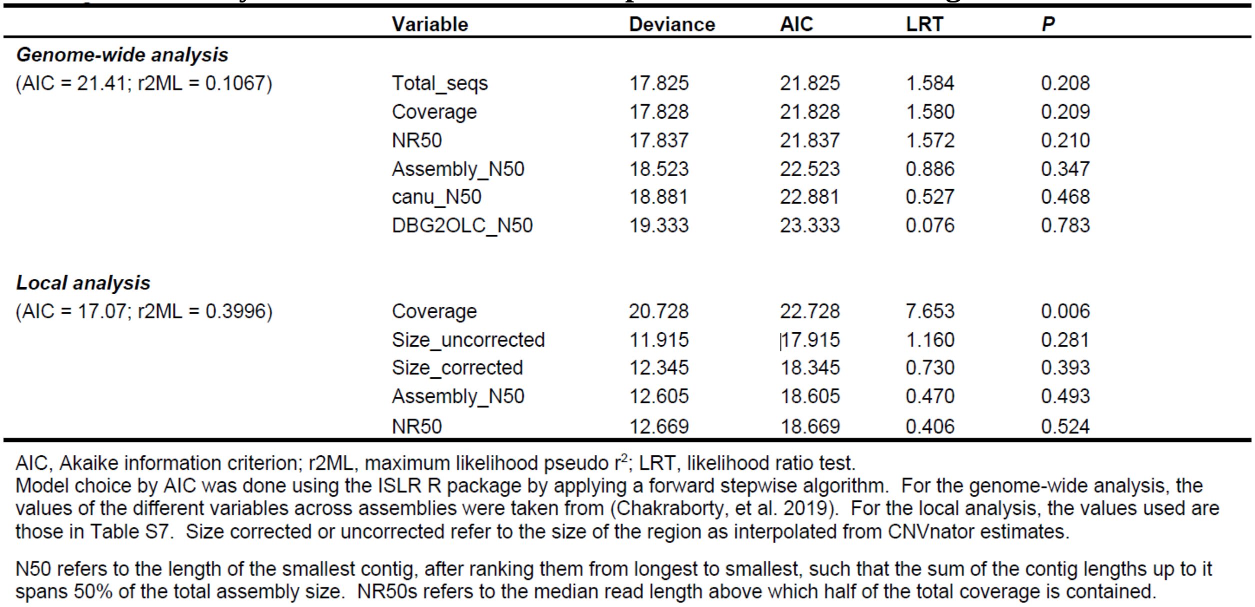
Supplementary Table S2.5. Pearson's correlation coefficient among CN estimates obtained with different methodologies

	Genome Annotation	qPCR	CNVnator
Genome Annotation	-	0.3369 (*P* = 0.2603)	0.3141 (*P* = 0.2741) *
qPCR		-	1 (*P* < 0.0001) * ‡
CNVnator			-

* In comparisons involving qPCR values, AB8 was omitted.

‡ Pearson's correlation coefficient prior to rounding-off CNVnator and qPCR original values was 0.9720.

Supplementary Table S2.6. Logistic regression analysis to evaluate the relevance of different assembly metrics in the faithful recapitulation of the *Sdic* region

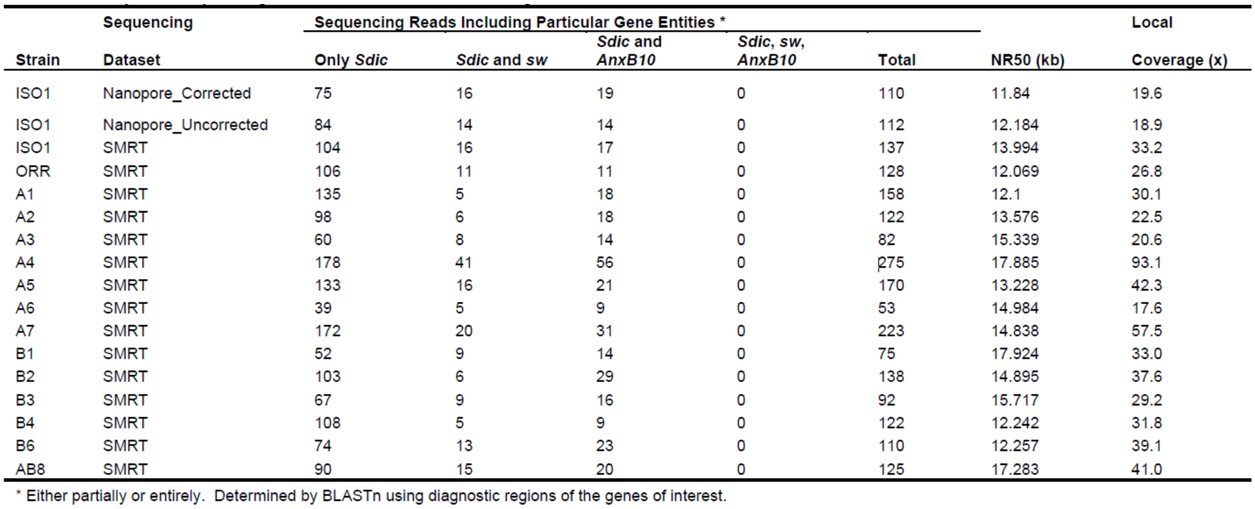


	Variable	Deviance	AIC	LRT	*P*
Genome-wide analysis					
(AIC = 21.41; r2ML = 0.1067)	Total_seqs	17.825	21.825	1.584	0.208
	Coverage	17.828	21.828	1.580	0.209
	NR50	17.837	21.837	1.572	0.210
	Assembly_N50	18.523	22.523	0.886	0.347
	canu_N50	18.881	22.881	0.527	0.468
	DBG2OLC_N50	19.333	23.333	0.076	0.783
Local analysis					
(AIC = 17.07; r2ML = 0.3996)	Coverage	20.728	22.728	7.653	0.006
	Size_uncorrected	11.915	17.915	1.160	0.281
	Size_corrected	12.345	18.345	0.730	0.393
	Assembly_N50	12.605	18.605	0.470	0.493
	NR50	12.669	18.669	0.406	0.524

AIC, Akaike information criterion; r2ML, maximum likelihood pseudo r^2; LRT, likelihood ratio test.
Model choice by AIC was done using the ISLR R package by applying a forward stepwise algorithm. For the genome-wide analysis, the values of the different variables across assemblies were taken from (Chakraborty, et al. 2019). For the local analysis, the values used are those in Table S7. Size corrected or uncorrected refer to the size of the region as interpolated from CNVnator estimates.

N50 refers to the length of the smallest contig, after ranking them from longest to smallest, such that the sum of the contig lengths up to it spans 50% of the total assembly size. NR50s refers to the median read length above which half of the total coverage is contained.

Supplementary Table S2.7. Analysis of sequencing reads associated with the *Sdic* region across datasets



	Sequencing	Sequencing Reads Including Particular Gene Entities *						Local
Strain	**Dataset**	**Only *Sdic***	***Sdic* and *sw***	***Sdic* and *AnxB10***	***Sdic, sw, AnxB10***	**Total**	**NR50 (kb)**	**Coverage (x)**
ISO1	Nanopore_Corrected	75	16	19	0	110	11.84	19.6
ISO1	Nanopore_Uncorrected	84	14	14	0	112	12.184	18.9
ISO1	SMRT	104	16	17	0	137	13.994	33.2
ORR	SMRT	106	11	11	0	128	12.069	26.8
A1	SMRT	135	5	18	0	158	12.1	30.1
A2	SMRT	98	6	18	0	122	13.576	22.5
A3	SMRT	60	8	14	0	82	15.339	20.6
A4	SMRT	178	41	56	0	275	17.885	93.1
A5	SMRT	133	16	21	0	170	13.228	42.3
A6	SMRT	39	5	9	0	53	14.984	17.6
A7	SMRT	172	20	31	0	223	14.838	57.5
B1	SMRT	52	9	14	0	75	17.924	33.0
B2	SMRT	103	6	29	0	138	14.895	37.6
B3	SMRT	67	9	16	0	92	15.717	29.2
B4	SMRT	108	5	9	0	122	12.242	31.8
B6	SMRT	74	13	23	0	110	12.257	39.1
AB8	SMRT	90	15	20	0	125	17.283	41.0

* Either partially or entirely. Determined by BLASTn using diagnostic regions of the genes of interest.

Supplementary Table S2.8. CNVnator results for strains from six different populations

		Illumina	Reference Genome †						Rounded-Off	CN
Population *	Strain	Library ID	A4_1	A4_2	A4_3	A4_4	A4_5	Average	Average	Estimate ‡
ZW	ZH23	SRX765992	6.547	6.562	6.579	6.572	6.484	6.549	7	6
ZW	ZH26	SRX766078	8.583	8.617	8.622	8.585	8.412	8.564	9	8
ZW	ZH33	SRX766073	8.437	8.473	8.516	8.500	8.456	8.476	8	7
ZW	ZH42	SRX766069	7.184	7.289	7.328	7.303	7.168	7.254	7	6
ZW	ZS10	SRX765993	10.614	10.616	10.628	10.633	10.597	10.618	11	10
ZW	ZW09	SRX766079	7.864	8.020	8.023	8.000	7.779	7.937	8	7
ZW	ZW139	SRX766075	7.380	7.404	7.416	7.388	7.424	7.402	7	6
ZW	ZW140	SRX766074	7.837	7.842		7.846	7.677	7.800	8	7
ZW	ZW142	SRX766072	8.082	8.120	8.189	8.201	8.017	8.122	8	7
ZW	ZW155	SRX766071	9.282	9.295	9.345	9.301	9.188	9.282	9	8
ZW	ZW177	SRX765990	6.049	6.058	6.085	6.058	6.042	6.058	6	5
ZW	ZW185	SRX766096	8.944	9.062	9.114	9.096	9.033	9.050	9	8
T	T05	SRX766109	6.448	6.494	6.497	6.475	6.449	6.473	6	5
T	T07	SRX766106	6.436	6.462	6.447	6.447	6.390	6.436	6	5
T	T09	SRX766105	7.929	8.042	8.027	8.022	7.882	7.980	8	7
T	T10	SRX766104	7.669	7.778	7.778	7.766	7.677	7.734	8	7
T	T14A	SRX766102	7.416	7.409	7.557	7.527	7.421	7.466	7	6
T	T22A	SRX766114	6.816	6.846	6.866	6.820	6.721	6.814	7	6
T	T23	SRX766101	7.421	7.448	7.495	7.656	7.334	7.471	7	6
T	T24	SRX766115	6.613	6.703	6.701	6.665	6.453	6.627	7	6
T	T25A	SRX766112	7.276	7.385	7.374	7.291	7.282	7.322	7	6
T	T29A	SRX766107	7.750	7.774	7.768	7.770	7.733	7.759	8	7
T	T30	SRX766111	6.937	6.965	7.002	6.853	6.837	6.919	7	6
T	T35	SRX766127	7.505	7.551	7.534	7.544	7.486	7.524	8	7
T	T36B	SRX766122	6.633	6.632	6.674	6.635	6.580	6.631	7	6
T	T39	SRX766137	7.559	7.570	7.611	7.574	7.471	7.557	8	7
T	T43A	SRX766134	4.682	4.695	4.710		4.652	4.685	5	4
T	T45B	SRX766129	7.598	7.637	7.641	7.614	7.593	7.617	8	7
N	N01	SRX766128	5.851	5.865	5.871	5.858	5.844	5.858	6	5
N	N02	SRX766120	5.755	5.774	5.848	5.799		5.794	6	5
N	N03	SRX766118	6.334	6.496	6.478	6.493		6.450	6	5
N	N04	SRX766117	7.445	7.610	7.628	7.617		7.575	8	7
N	N07	SRX766132	5.336	5.372	5.322	5.244	5.253	5.305	5	4
N	N10	SRX766131	8.437	8.611	8.663	8.620	8.483	8.563	9	8
N	N11	SRX766126	8.149	8.156	8.309	8.273	8.174	8.212	8	7
N	N13	SRX766133	7.798	7.828	7.795	7.838	7.580	7.768	8	7
N	N14	SRX766130	6.510	6.544	6.508	6.514	6.446	6.505	7	6
N	N15	SRX766125	5.535	5.454	5.571	5.427		5.497	5	4
N	N16	SRX766124	5.804	5.894	5.875	5.876	5.834	5.857	6	5
N	N17	SRX766121	5.757	5.826	5.827	5.850	5.807	5.813	6	5
N	N18	SRX766136	7.600	7.603	7.482	7.465	7.555	7.541	8	7
N	N19	SRX766123	6.793	6.884	6.861	6.745		6.821	7	6
N	N22	SRX766191	6.889	6.914	6.906	6.890		6.900	7	6
N	N23	SRX766178	7.671	7.776	7.776	7.751		7.744	8	7
N	N25	SRX766177	6.233	6.233	6.266	6.201	6.151	6.217	6	5
N	N29	SRX766189	9.283	9.316	9.370	9.311	9.040	9.264	9	8
N	N30	SRX766187	6.472	6.494	6.472	6.479		6.479	6	5
I	I03	SRX766185	5.698	5.723	5.746	5.743	5.706	5.723	6	5
I	I07	SRX766184	7.811	7.836	7.852	7.828	7.836	7.833	8	7
I	I13	SRX766183	6.796	6.831	6.743	6.731	6.693	6.759	7	6
I	I17	SRX766182	7.321	7.430	7.406	7.403		7.390	7	6
I	I22	SRX766181	7.380	7.486	7.470	7.462	7.383	7.436	7	6
I	I23	SRX766192	6.893	6.934	6.937	6.904		6.917	7	6
I	I24	SRX766190	8.957	8.997	9.029	8.973	8.947	8.981	9	8
I	I29	SRX766188	6.641		6.671	6.657	6.622	6.648	7	6
I	I33	SRX766186	9.616	9.623	9.612	9.667	9.631	9.630	10	9
I	I34	SRX766180	9.167	9.193	9.239	9.195	9.170	9.193	9	8
I	I35	SRX766179	7.737	7.752	7.799	7.782	7.710	7.756	8	7

Supplementary Table S2.8. CNVnator results for strains from six different populations

		Illumina	Reference Genome †						Rounded-Off	CN
Population *	Strain	Library ID	A4_1	A4_2	A4_3	A4_4	A4_5	Average	Average	Estimate ‡
I	I38	SRX766204	8.942	8.952	8.929	8.963	8.881	8.933	9	8
B	B04	SRX766199	6.949	6.965	6.968	6.993	6.955	6.966	7	6
B	B10	SRX766196	5.271	5.306	5.334	5.287	5.285	5.297	5	4
B	B11	SRX766193	6.227	6.288	6.320	6.294	6.287	6.283	6	5
B	B12	SRX766201	6.073	6.099	6.128	6.105	6.033	6.088	6	5
B	B23	SRX766200	7.166	7.117	7.182	7.104	7.155	7.145	7	6
B	B28	SRX766198	6.907	6.965	6.992	6.962	6.853	6.936	7	6
B	B38	SRX766195	6.762	6.767	6.812	6.800	6.826	6.793	7	6
B	B42	SRX766203	5.669	5.691	5.696	5.695	5.684	5.687	6	5
B	B43	SRX766206	6.620	6.638	6.656	6.641	6.580	6.627	7	6
B	B54	SRX766194	6.253	6.259	6.266	6.261	6.244	6.257	6	5
B	B59	SRX766197	6.646	6.716	6.719	6.664		6.686	7	6
ZM	ZI10_1-HE	SRR203502	8.068	8.138		8.167	8.319	8.173	8	7
ZM	ZI152_1-HE	SRR326790	4.160	4.018	4.057	4.157	4.223	4.123	4	3
ZM	ZI173_1-HE	SRR203330	5.170	5.185	5.227	5.216	5.279	5.215	5	4
ZM	ZI177_1-HE	SRR326796	4.015	4.057	4.024	4.114	4.140	4.070	4	3
ZM	ZI181_1-HE	SRR203069	4.475	4.444		4.506	4.550	4.494	4	3
ZM	ZI184_1-HE	SRR203068	4.176	4.166	4.235	4.231	4.296	4.221	4	3
ZM	ZI188_1-HE	SRR202123	5.918	5.918	5.948	5.959	6.044	5.957	6	5
ZM	ZI194_1-HE	SRR203319	5.861	5.929	5.938	5.967	6.021	5.943	6	5
ZM	ZI196_1-HE	SRR203467	5.681		5.845	5.725	5.834	5.771	6	5
ZM	ZI197N_1-HE	SRR342395	4.605	4.566	4.623	4.628	4.684	4.621	5	4
ZM	ZI199_1-HE	SRR203468	6.066	6.092	6.135	6.172	6.094	6.112	6	5
ZM	ZI207_1-HE	SRR202075	4.321	4.298	4.404	4.544		4.392	4	3
ZM	ZI212_1-HE	SRR204012	5.991	5.993	6.045	6.085	6.059	6.035	6	5
ZM	ZI216N_1-HE	SRR203328	4.927	4.945	4.971	4.977	4.945	4.953	5	4
ZM	ZI226_1-HE	SRR203348	5.145	5.130	5.197	5.222	5.285	5.196	5	4
ZM	ZI227_1-HE	SRR202126	5.522	5.552	5.594	5.647	5.670	5.597	6	5
ZM	ZI228_1-HE	SRR203064	3.310	3.364	3.354	3.400	3.436	3.373	3	2
ZM	ZI232_1-HE	SRR202076	3.588	3.584	3.626	3.647	3.643	3.618	4	3
ZM	ZI241_1-HE	SRR326798	5.108	5.126	5.151	5.179	5.285	5.170	5	4
ZM	ZI252_1-HE	SRR203349	6.676	6.762	6.759	6.831	6.798	6.765	7	6
ZM	ZI253_1-HE	SRR203350	6.120	6.076		6.207	6.234	6.159	6	5
ZM	ZI281_1-HE	SRR342393	5.529	5.524	5.564	5.567	5.603	5.558	6	5
ZM	ZI284_1-HE	SRR654554	5.423	5.501	5.513	5.540	5.568	5.509	6	5
ZM	ZI295_1-HE	SRR202099	6.681	6.675		6.664	6.736	6.689	7	6
ZM	ZI311N_1-HE	SRR326797	6.104	6.159	6.159	6.145	6.265	6.167	6	5
ZM	ZI317_1-HE	SRR204011	4.867	4.858	4.913	4.900	4.927	4.893	5	4
ZM	ZI319_2-HE	SRR203461	4.611	4.573	4.622	4.625	4.687	4.624	5	4
ZM	ZI320_1-HE	SRR326793	3.931	3.958	3.902	4.064		3.964	4	3
ZM	ZI324_1-HE	SRR204014	5.986	5.969		6.072	6.074	6.025	6	5
ZM	ZI335_1-HE	SRR203471	5.990	5.959	6.049	6.081	6.062	6.028	6	5
ZM	ZI342_1-HE	SRR094875	6.204	6.148	6.243	6.265	6.395	6.251	6	5
ZM	ZI348_1-HE	SRR203475		5.341	5.359	5.461	5.429	5.398	5	4
ZM	ZI353_1-HE	SRR342394	4.387	4.399	4.391	4.411	4.513	4.420	4	3
ZM	ZI358_1-HE	SRR346928	5.868	5.862	5.911	5.930	6.013	5.917	6	5
ZM	ZI362_1-HE_2.2	SRR654685	13.087	12.941	13.123	13.163	13.345	13.132	13	12
ZM	ZI362_1-HE_3.4	SRR346932	12.482	12.762	12.702	12.847	13.192	12.797	13	12
ZI	ZI364_1-HE	SRR204013	5.066		5.111	5.081	5.052	5.078	5	4
ZI	ZI368_1-HE	SRR203462	5.280	5.274	5.289	5.226	5.349	5.284	5	4
ZI	ZI373_1-HE	SRR210782	4.486	4.506	4.595	4.588	4.615	4.558	5	4
ZI	ZI374_1-HE	SRR204008	6.133	6.169	6.234	6.201	6.182	6.184	6	5
ZI	ZI378_1-HE	SRR203464	5.358	5.357		5.384	5.390	5.372	5	4
ZI	ZI381_1-HE	SRR204010		3.665	3.709	3.710	3.736	3.705	4	3
ZI	ZI384_1-HE	SRR354004	4.321	4.332	4.360	4.367	4.421	4.360	4	3
ZI	ZI395_1-HE	SRR326802	4.929	4.936	4.949	4.998	5.084	4.979	5	4
ZI	ZI396_1-HE	SRR353757	4.815	4.867		4.910	4.970	4.890	5	4
ZI	ZI397N_2-HE	SRR654677	7.689	7.690	7.700	7.731	7.841	7.730	8	7

Supplementary Table S2.8. CNVnator results for strains from six different populations

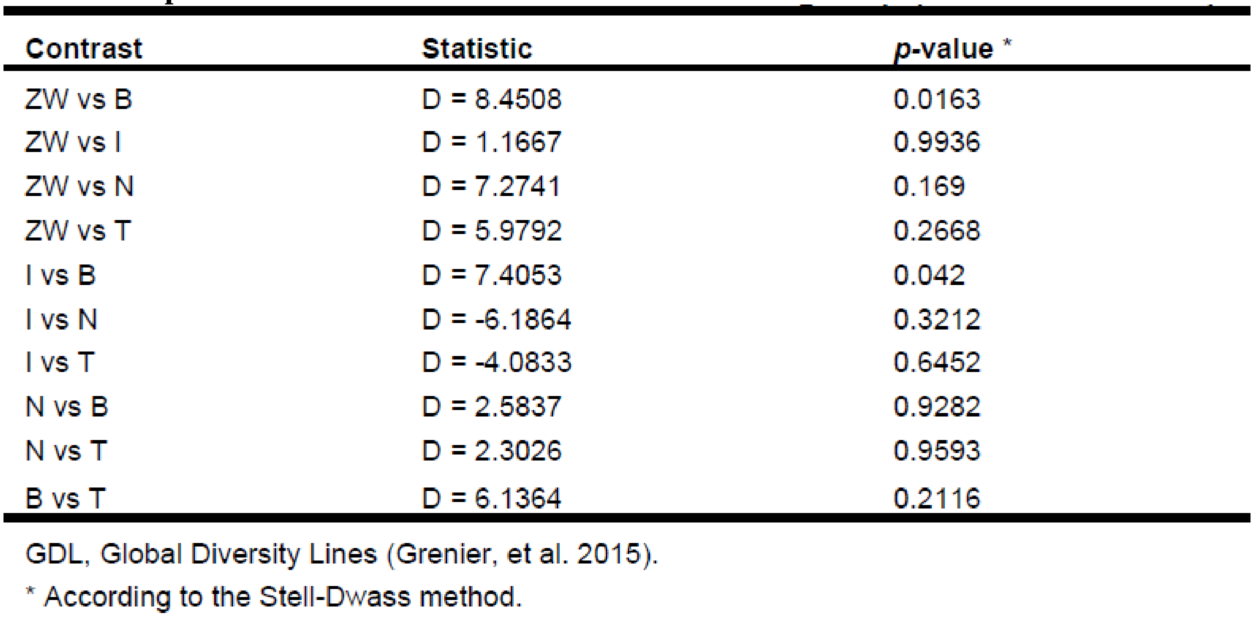
		Illumina	Reference Genome †						Rounded-Off	CN
Population *	Strain	Library ID	A4_1	A4_2	A4_3	A4_4	A4_5	Average	Average	Estimate ‡
ZI	ZI398_1-HE	SRR346930	4.365	4.362	4.369		4.523	4.405	4	3
ZI	ZI405_2-HE	SRR354003	4.229	4.232	4.233	4.260	4.287	4.248	4	3
ZI	ZI418N_1-HE	SRR202100	5.214	5.235	5.264	5.299	5.305	5.263	5	4
ZI	ZI431_1-HE	SRR654556	4.113	4.105	4.131	4.116	4.254	4.144	4	3
ZI	ZI444_1-HE	SRR203463	4.877	4.846	4.876	4.888	4.957	4.889	5	4
ZI	ZI456_1-HE	SRR203466	5.200	5.189	5.225	5.210	5.263	5.217	5	4
ZI	ZI472_1-HE	SRR203465	7.833	7.890	7.912	7.920	7.968	7.904	8	7
ZI	ZI477_1-HE	SRR353760	4.395	4.385	4.406	4.430	4.527	4.429	4	3
ZI	ZI508_1-HE	SRR346929	3.739		3.794	3.819	3.875	3.807	4	3
ZI	ZI50N_1-HE	SRR203334	6.353	6.372	6.404	6.417	6.517	6.412	6	5
ZI	ZI514N_1-HE	SRR654679		6.137	6.195	6.218	6.301	6.213	6	5
ZI	ZI523_1-HE	SRR342396	3.863	3.868		3.907	3.991	3.907	4	3
ZI	ZI530_1-HE	SRR204009	4.443	4.475	4.517	4.465	4.535	4.487	4	3
ZI	ZI59_1-HE	SRR202112	5.421	5.469	5.469	5.511	5.503	5.475	5	4

The number in the ID of the reference genome denotes the order of the *Sdic* copy in the original strain between the flanking genes, specifically from sw to AnxB10. For example, A4_5 refers to the fifth copy starting from *sw*, *i.e.* the copy adjacent to *AnxB10* in the A4 strain.

Missing values denote not considered values because their associated target size felt outside the expected range, *i.e.* 7.2-8.0 kb.

‡ The CN estimate is calculated as the rounded off average minus one due to the contribution of reads from *sw* and *AnxB10* to the read-depth estimates obtained with CNVnator.

Supplementary Table S2.9. Statistical evidence of differences in CN among five populations of the GDL panel

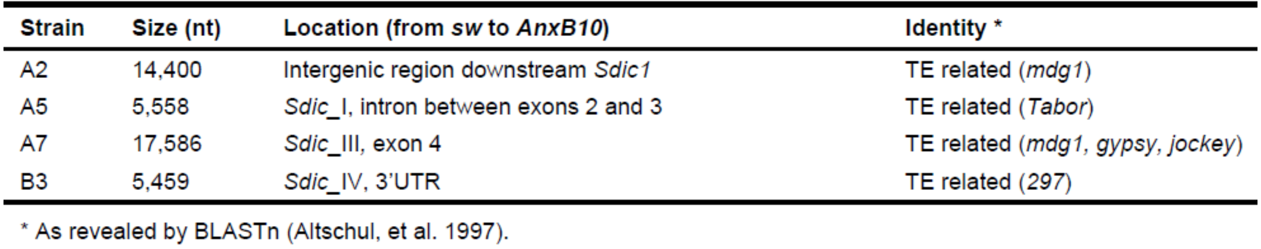


Contrast	Statistic	*p*-value *
ZW vs B	D = 8.4508	0.0163
ZW vs I	D = 1.1667	0.9936
ZW vs N	D = 7.2741	0.169
ZW vs T	D = 5.9792	0.2668
I vs B	D = 7.4053	0.042
I vs N	D = -6.1864	0.3212
I vs T	D = -4.0833	0.6452
N vs B	D = 2.5837	0.9282
N vs T	D = 2.3026	0.9593
B vs T	D = 6.1364	0.2116

GDL, Global Diversity Lines (Grenier, et al. 2015).

* According to the Stell-Dwass method.

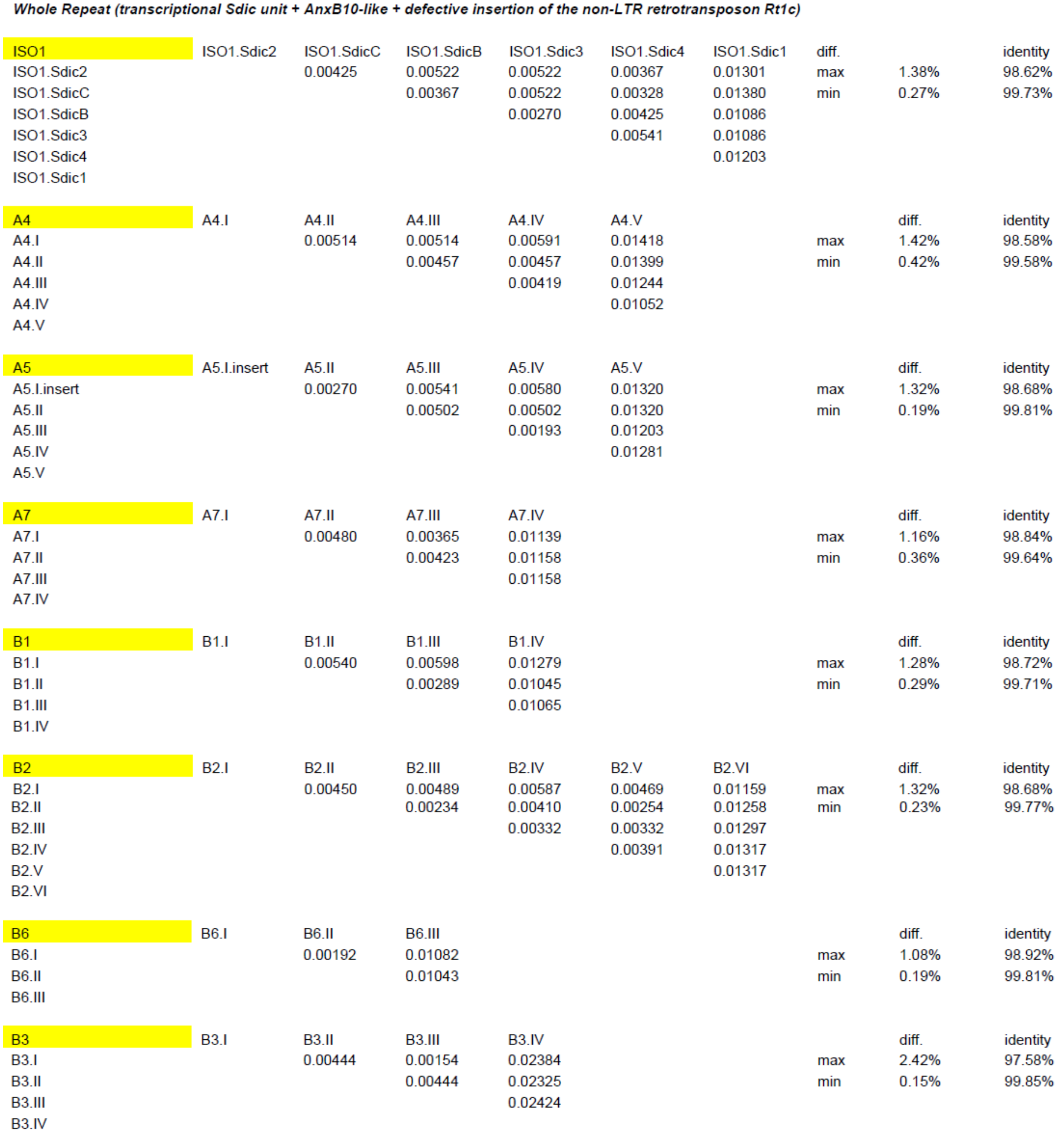
Supplementary Table S2.10. Natural population-specific transposable element (TE) insertions documented in the *Sdic* region



Strain	Size (nt)	Location (from *sw* to *AnxB10*)	Identity *
A2	14,400	Intergenic region downstream *Sdic1*	TE related (*mdg1*)
A5	5,558	*Sdic*_I, intron between exons 2 and 3	TE related (*Tabor*)
A7	17,586	*Sdic*_III, exon 4	TE related (*mdg1*, *gypsy*, *jockey*)
B3	5,459	*Sdic*_IV, 3'UTR	TE related (*297*)

* As revealed by BLASTn (Altschul, et al. 1997).

Supplementary Table S2.11. Nucleotide differentiation among *Sdic* copies in the reference strain and seven populations from diverse geographic origin of *D. melanogaster*



Whole Repeat (transcriptional Sdic unit + AnxB10-like + defective insertion of the non-LTR retrotransposon Rt1c)

ISO1	ISO1.Sdic2	ISO1.SdicC	ISO1.SdicB	ISO1.Sdic3	ISO1.Sdic4	ISO1.Sdic1	diff.		identity
ISO1.Sdic2		0.00425	0.00522	0.00522	0.00367	0.01301	max	1.38%	98.62%
ISO1.SdicC			0.00367	0.00522	0.00328	0.01380	min	0.27%	99.73%
ISO1.SdicB				0.00270	0.00425	0.01086			
ISO1.Sdic3					0.00541	0.01086			
ISO1.Sdic4						0.01203			
ISO1.Sdic1									

A4	A4.I	A4.II	A4.III	A4.IV	A4.V			diff.	identity
A4.I		0.00514	0.00514	0.00591	0.01418		max	1.42%	98.58%
A4.II			0.00457	0.00457	0.01399		min	0.42%	99.58%
A4.III				0.00419	0.01244				
A4.IV					0.01052				
A4.V									

A5	A5.I.insert	A5.II	A5.III	A5.IV	A5.V			diff.	identity
A5.I.insert		0.00270	0.00541	0.00580	0.01320		max	1.32%	98.68%
A5.II			0.00502	0.00502	0.01320		min	0.19%	99.81%
A5.III				0.00193	0.01203				
A5.IV					0.01281				
A5.V									

A7	A7.I	A7.II	A7.III	A7.IV				diff.	identity
A7.I		0.00480	0.00365	0.01139			max	1.16%	98.84%
A7.II			0.00423	0.01158			min	0.36%	99.64%
A7.III				0.01158					
A7.IV									

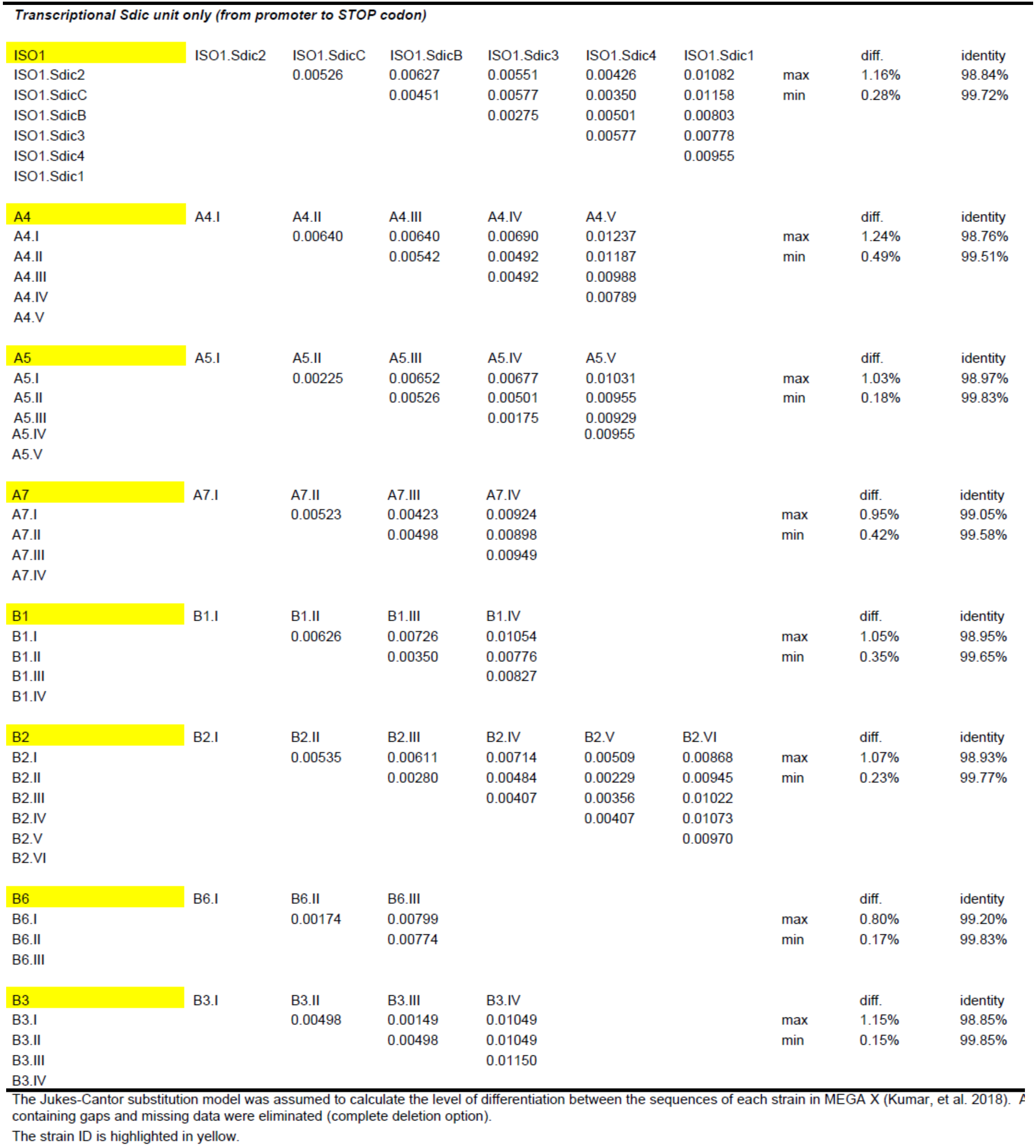
B1	B1.I	B1.II	B1.III	B1.IV				diff.	identity
B1.I		0.00540	0.00598	0.01279			max	1.28%	98.72%
B1.II			0.00289	0.01045			min	0.29%	99.71%
B1.III				0.01065					
B1.IV									

B2	B2.I	B2.II	B2.III	B2.IV	B2.V	B2.VI		diff.	identity
B2.I		0.00450	0.00489	0.00587	0.00469	0.01159	max	1.32%	98.68%
B2.II			0.00234	0.00410	0.00254	0.01258	min	0.23%	99.77%
B2.III				0.00332	0.00332	0.01297			
B2.IV					0.00391	0.01317			
B2.V						0.01317			
B2.VI									

B6	B6.I	B6.II	B6.III					diff.	identity
B6.I		0.00192	0.01082				max	1.08%	98.92%
B6.II			0.01043				min	0.19%	99.81%
B6.III									

B3	B3.I	B3.II	B3.III	B3.IV				diff.	identity
B3.I		0.00444	0.00154	0.02384			max	2.42%	97.58%
B3.II			0.00444	0.02325			min	0.15%	99.85%
B3.III				0.02424					
B3.IV									

Supplementary Table S2.11. Nucleotide differentiation among *Sdic* copies in the reference strain and seven populations from diverse geographic origin of *D. melanogaster*



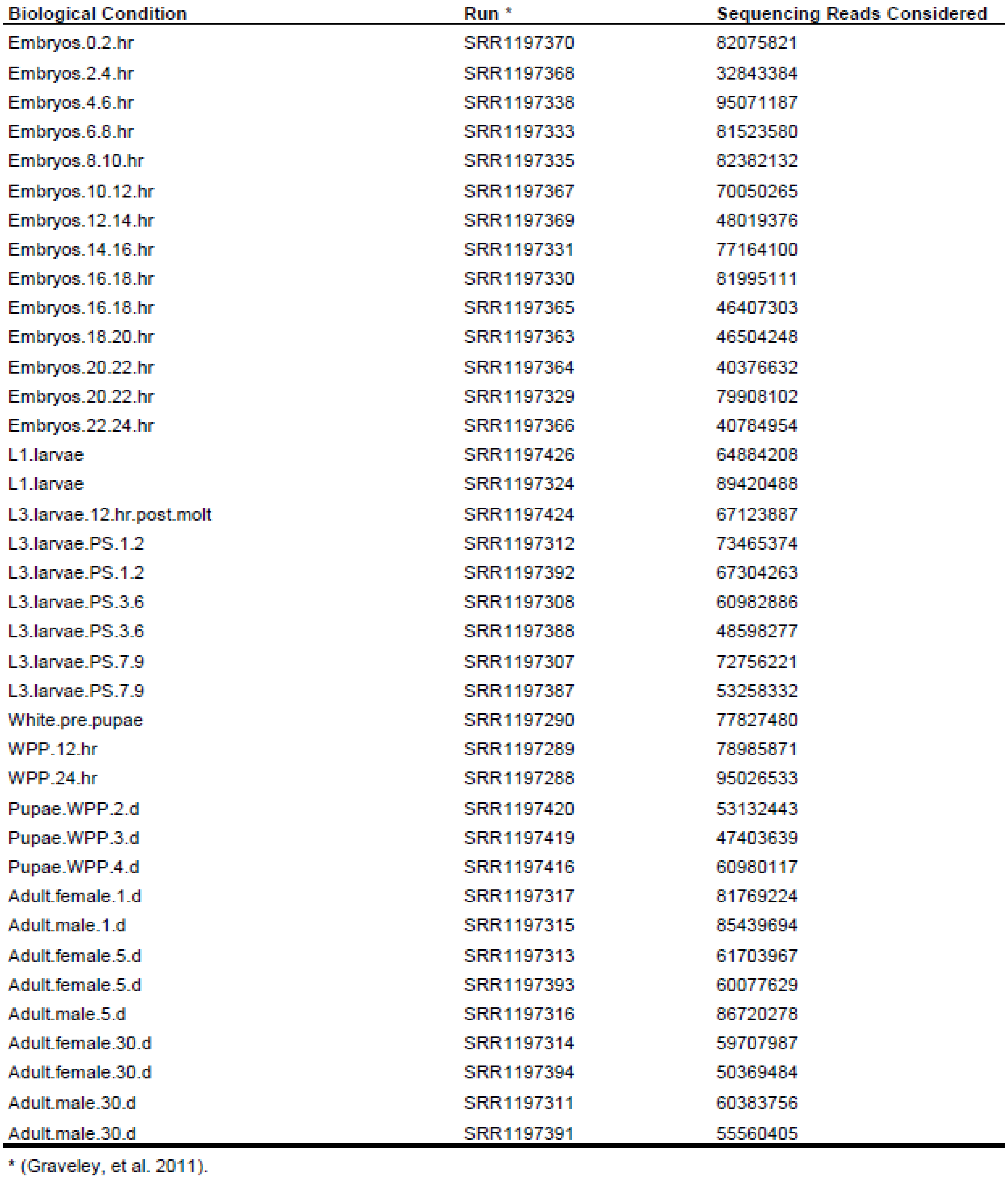
Transcriptional Sdic unit only (from promoter to STOP codon)

ISO1	ISO1.Sdic2	ISO1.SdicC	ISO1.SdicB	ISO1.Sdic3	ISO1.Sdic4	ISO1.Sdic1		diff.	identity
ISO1.Sdic2		0.00526	0.00627	0.00551	0.00426	0.01082	max	1.16%	98.84%
ISO1.SdicC			0.00451	0.00577	0.00350	0.01158	min	0.28%	99.72%
ISO1.SdicB				0.00275	0.00501	0.00803			
ISO1.Sdic3					0.00577	0.00778			
ISO1.Sdic4						0.00955			
ISO1.Sdic1									
A4	A4.I	A4.II	A4.III	A4.IV	A4.V			diff.	identity
A4.I		0.00640	0.00640	0.00690	0.01237		max	1.24%	98.76%
A4.II			0.00542	0.00492	0.01187		min	0.49%	99.51%
A4.III				0.00492	0.00988				
A4.IV					0.00789				
A4.V									
A5	A5.I	A5.II	A5.III	A5.IV	A5.V			diff.	identity
A5.I		0.00225	0.00652	0.00677	0.01031		max	1.03%	98.97%
A5.II			0.00526	0.00501	0.00955		min	0.18%	99.83%
A5.III				0.00175	0.00929				
A5.IV					0.00955				
A5.V									
A7	A7.I	A7.II	A7.III	A7.IV				diff.	identity
A7.I		0.00523	0.00423	0.00924			max	0.95%	99.05%
A7.II			0.00498	0.00898			min	0.42%	99.58%
A7.III				0.00949					
A7.IV									
B1	B1.I	B1.II	B1.III	B1.IV				diff.	identity
B1.I		0.00626	0.00726	0.01054			max	1.05%	98.95%
B1.II			0.00350	0.00776			min	0.35%	99.65%
B1.III				0.00827					
B1.IV									
B2	B2.I	B2.II	B2.III	B2.IV	B2.V	B2.VI		diff.	identity
B2.I		0.00535	0.00611	0.00714	0.00509	0.00868	max	1.07%	98.93%
B2.II			0.00280	0.00484	0.00229	0.00945	min	0.23%	99.77%
B2.III				0.00407	0.00356	0.01022			
B2.IV					0.00407	0.01073			
B2.V						0.00970			
B2.VI									
B6	B6.I	B6.II	B6.III					diff.	identity
B6.I		0.00174	0.00799				max	0.80%	99.20%
B6.II			0.00774				min	0.17%	99.83%
B6.III									
B3	B3.I	B3.II	B3.III	B3.IV				diff.	identity
B3.I		0.00498	0.00149	0.01049			max	1.15%	98.85%
B3.II			0.00498	0.01049			min	0.15%	99.85%
B3.III				0.01150					
B3.IV									

The Jukes-Cantor substitution model was assumed to calculate the level of differentiation between the sequences of each strain in MEGA X (Kumar, et al. 2018). A containing gaps and missing data were eliminated (complete deletion option).

The strain ID is highlighted in yellow.

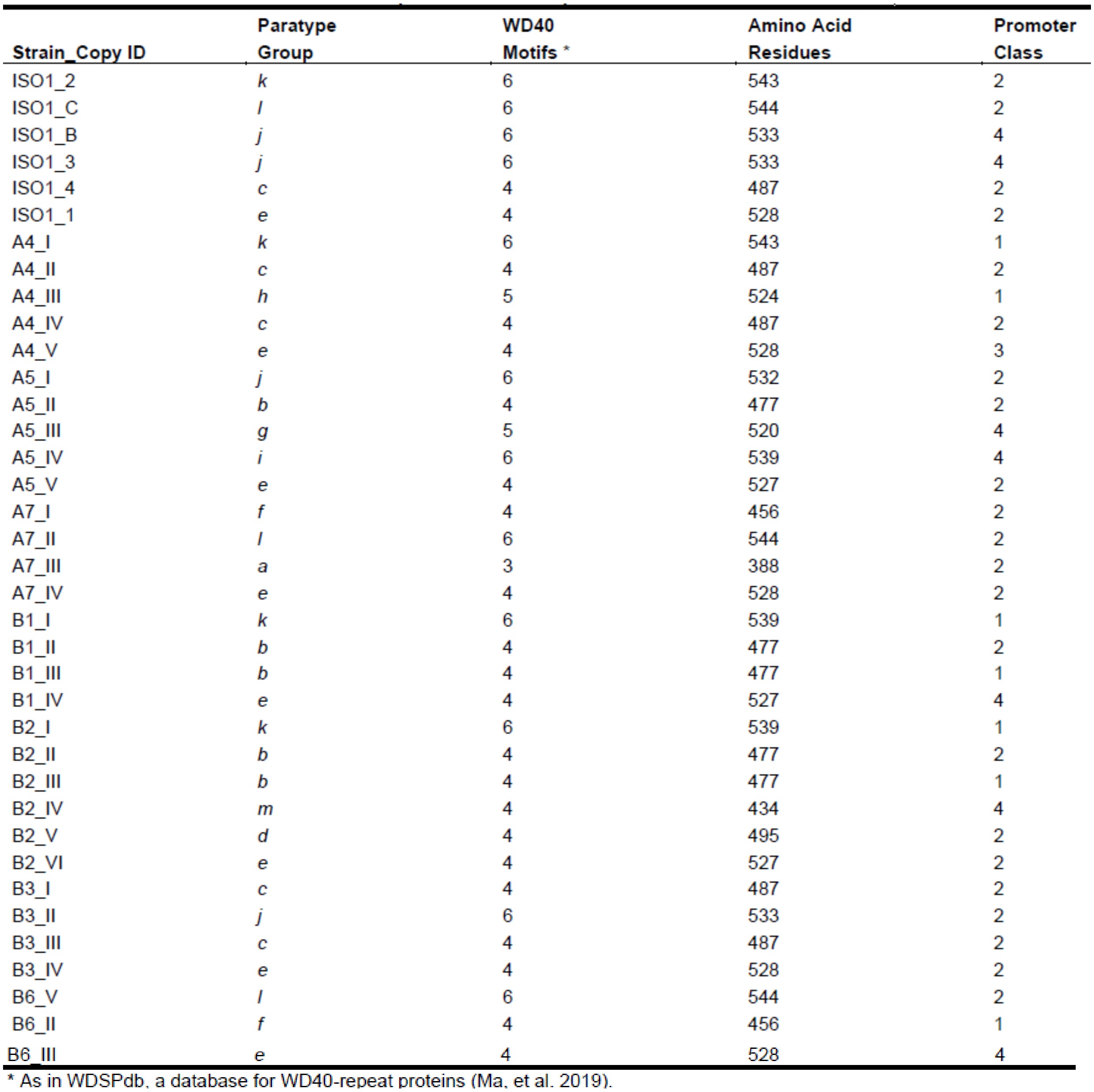
Supplementary Table S2.12. RNA-seq datasets examined for expression of *AnxB10*-like



Biological Condition	Run *	Sequencing Reads Considered
Embryos.0.2.hr	SRR1197370	82075821
Embryos.2.4.hr	SRR1197368	32843384
Embryos.4.6.hr	SRR1197338	95071187
Embryos.6.8.hr	SRR1197333	81523580
Embryos.8.10.hr	SRR1197335	82382132
Embryos.10.12.hr	SRR1197367	70050265
Embryos.12.14.hr	SRR1197369	48019376
Embryos.14.16.hr	SRR1197331	77164100
Embryos.16.18.hr	SRR1197330	81995111
Embryos.16.18.hr	SRR1197365	46407303
Embryos.18.20.hr	SRR1197363	46504248
Embryos.20.22.hr	SRR1197364	40376632
Embryos.20.22.hr	SRR1197329	79908102
Embryos.22.24.hr	SRR1197366	40784954
L1.larvae	SRR1197426	64884208
L1.larvae	SRR1197324	89420488
L3.larvae.12.hr.post.molt	SRR1197424	67123887
L3.larvae.PS.1.2	SRR1197312	73465374
L3.larvae.PS.1.2	SRR1197392	67304263
L3.larvae.PS.3.6	SRR1197308	60982886
L3.larvae.PS.3.6	SRR1197388	48598277
L3.larvae.PS.7.9	SRR1197307	72756221
L3.larvae.PS.7.9	SRR1197387	53258332
White.pre.pupae	SRR1197290	77827480
WPP.12.hr	SRR1197289	78985871
WPP.24.hr	SRR1197288	95026533
Pupae.WPP.2.d	SRR1197420	53132443
Pupae.WPP.3.d	SRR1197419	47403639
Pupae.WPP.4.d	SRR1197416	60980117
Adult.female.1.d	SRR1197317	81769224
Adult.male.1.d	SRR1197315	85439694
Adult.female.5.d	SRR1197313	61703967
Adult.female.5.d	SRR1197393	60077629
Adult.male.5.d	SRR1197316	86720278
Adult.female.30.d	SRR1197314	59707987
Adult.female.30.d	SRR1197394	50369484
Adult.male.30.d	SRR1197311	60383756
Adult.male.30.d	SRR1197391	55560405

* (Graveley, et al. 2011).

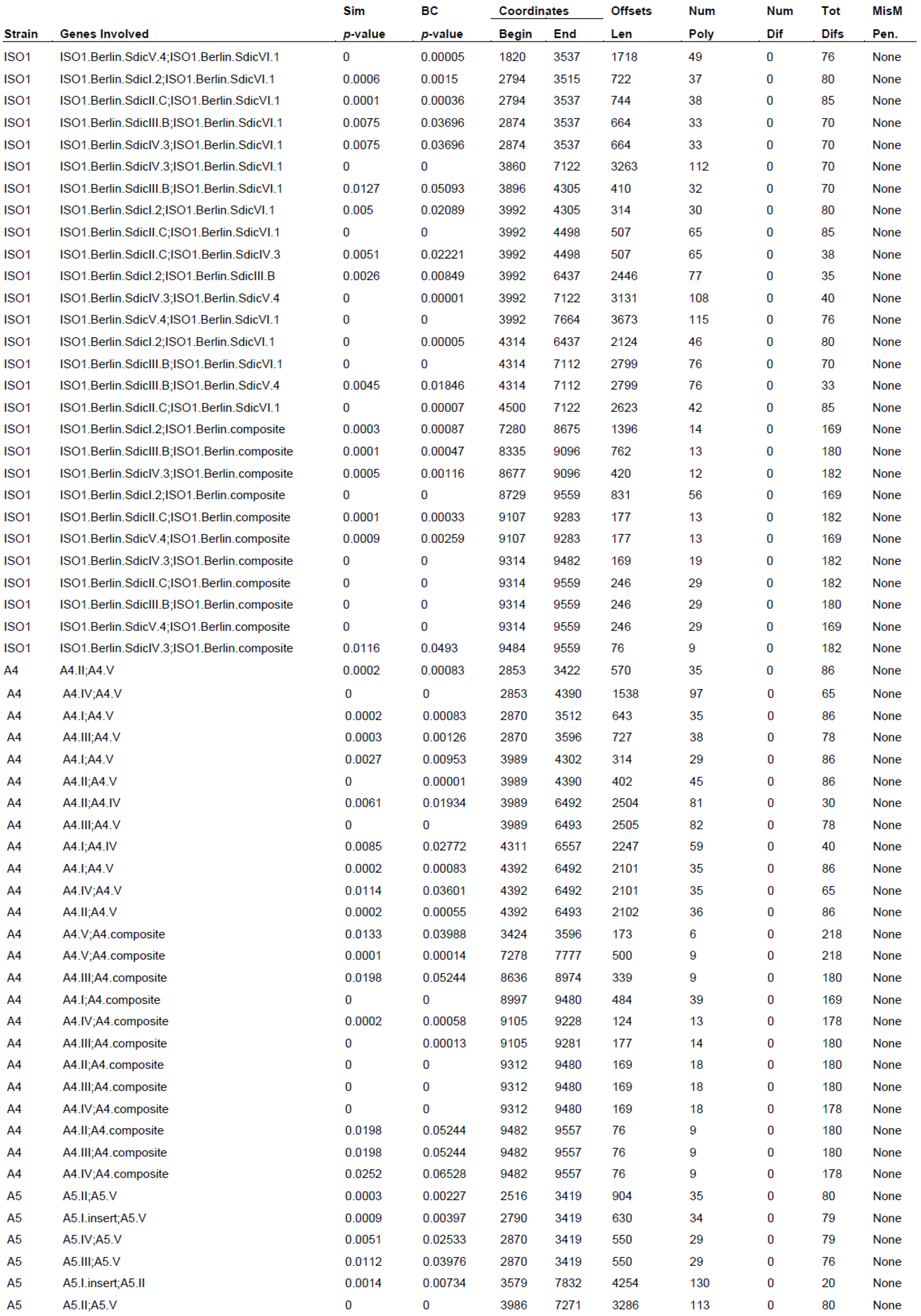
Supplementary Table S2.13. Salient features of the encoded product of *Sdic* copies annotated in reliable assemblies



Strain_Copy ID	Paratype Group	WD40 Motifs *	Amino Acid Residues	Promoter Class
ISO1_2	*k*	6	543	2
ISO1_C	*l*	6	544	2
ISO1_B	*j*	6	533	4
ISO1_3	*j*	6	533	4
ISO1_4	*c*	4	487	2
ISO1_1	*e*	4	528	2
A4_I	*k*	6	543	1
A4_II	*c*	4	487	2
A4_III	*h*	5	524	1
A4_IV	*c*	4	487	2
A4_V	*e*	4	528	3
A5_I	*j*	6	532	2
A5_II	*b*	4	477	2
A5_III	*g*	5	520	4
A5_IV	*i*	6	539	4
A5_V	*e*	4	527	2
A7_I	*f*	4	456	2
A7_II	*l*	6	544	2
A7_III	*a*	3	388	2
A7_IV	*e*	4	528	2
B1_I	*k*	6	539	1
B1_II	*b*	4	477	2
B1_III	*b*	4	477	1
B1_IV	*e*	4	527	4
B2_I	*k*	6	539	1
B2_II	*b*	4	477	2
B2_III	*b*	4	477	1
B2_IV	*m*	4	434	4
B2_V	*d*	4	495	2
B2_VI	*e*	4	527	2
B3_I	*c*	4	487	2
B3_II	*j*	6	533	2
B3_III	*c*	4	487	2
B3_IV	*e*	4	528	2
B6_V	*l*	6	544	2
B6_II	*f*	4	456	1
B6_III	*e*	4	528	4

* As in WDSPdb, a database for WD40-repeat proteins (Ma, et al. 2019).

Supplementary Table S2.14. Gene conversion events detected in the *Sdic* region for eight strains of *D. melanogaster* according to GenConv



		Sim	BC	Coordinates		Offsets	Num	Num	Tot	MisM
Strain	Genes Involved	*p*-value	*p*-value	Begin	End	Len	Poly	Dif	Difs	Pen.
ISO1	ISO1.Berlin.SdicV.4;ISO1.Berlin.SdicVI.1	0	0.00005	1820	3537	1718	49	0	76	None
ISO1	ISO1.Berlin.SdicI.2;ISO1.Berlin.SdicVI.1	0.0006	0.0015	2794	3515	722	37	0	80	None
ISO1	ISO1.Berlin.SdicII.C;ISO1.Berlin.SdicVI.1	0.0001	0.00036	2794	3537	744	38	0	85	None
ISO1	ISO1.Berlin.SdicIII.B;ISO1.Berlin.SdicVI.1	0.0075	0.03696	2874	3537	664	33	0	70	None
ISO1	ISO1.Berlin.SdicIV.3;ISO1.Berlin.SdicVI.1	0.0075	0.03696	2874	3537	664	33	0	70	None
ISO1	ISO1.Berlin.SdicIV.3;ISO1.Berlin.SdicVI.1	0	0	3860	7122	3263	112	0	70	None
ISO1	ISO1.Berlin.SdicIII.B;ISO1.Berlin.SdicVI.1	0.0127	0.05093	3896	4305	410	32	0	70	None
ISO1	ISO1.Berlin.SdicI.2;ISO1.Berlin.SdicVI.1	0.005	0.02089	3992	4305	314	30	0	80	None
ISO1	ISO1.Berlin.SdicII.C;ISO1.Berlin.SdicVI.1	0	0	3992	4498	507	65	0	85	None
ISO1	ISO1.Berlin.SdicII.C;ISO1.Berlin.SdicIV.3	0.0051	0.02221	3992	4498	507	65	0	38	None
ISO1	ISO1.Berlin.SdicI.2;ISO1.Berlin.SdicIII.B	0.0026	0.00849	3992	6437	2446	77	0	35	None
ISO1	ISO1.Berlin.SdicIV.3;ISO1.Berlin.SdicV.4	0	0.00001	3992	7122	3131	108	0	40	None
ISO1	ISO1.Berlin.SdicV.4;ISO1.Berlin.SdicVI.1	0	0	3992	7664	3673	115	0	76	None
ISO1	ISO1.Berlin.SdicI.2;ISO1.Berlin.SdicVI.1	0	0.00005	4314	6437	2124	46	0	80	None
ISO1	ISO1.Berlin.SdicIII.B;ISO1.Berlin.SdicVI.1	0	0	4314	7112	2799	76	0	70	None
ISO1	ISO1.Berlin.SdicIII.B;ISO1.Berlin.SdicV.4	0.0045	0.01846	4314	7112	2799	76	0	33	None
ISO1	ISO1.Berlin.SdicII.C;ISO1.Berlin.SdicVI.1	0	0.00007	4500	7122	2623	42	0	85	None
ISO1	ISO1.Berlin.SdicI.2;ISO1.Berlin.composite	0.0003	0.00087	7280	8675	1396	14	0	169	None
ISO1	ISO1.Berlin.SdicIII.B;ISO1.Berlin.composite	0.0001	0.00047	8335	9096	762	13	0	180	None
ISO1	ISO1.Berlin.SdicIV.3;ISO1.Berlin.composite	0.0005	0.00116	8677	9096	420	12	0	182	None
ISO1	ISO1.Berlin.SdicI.2;ISO1.Berlin.composite	0	0	8729	9559	831	56	0	169	None
ISO1	ISO1.Berlin.SdicII.C;ISO1.Berlin.composite	0.0001	0.00033	9107	9283	177	13	0	182	None
ISO1	ISO1.Berlin.SdicV.4;ISO1.Berlin.composite	0.0009	0.00259	9107	9283	177	13	0	169	None
ISO1	ISO1.Berlin.SdicIV.3;ISO1.Berlin.composite	0	0	9314	9482	169	19	0	182	None
ISO1	ISO1.Berlin.SdicII.C;ISO1.Berlin.composite	0	0	9314	9559	246	29	0	182	None
ISO1	ISO1.Berlin.SdicIII.B;ISO1.Berlin.composite	0	0	9314	9559	246	29	0	180	None
ISO1	ISO1.Berlin.SdicV.4;ISO1.Berlin.composite	0	0	9314	9559	246	29	0	169	None
ISO1	ISO1.Berlin.SdicIV.3;ISO1.Berlin.composite	0.0116	0.0493	9484	9559	76	9	0	182	None
A4	A4.II;A4.V	0.0002	0.00083	2853	3422	570	35	0	86	None
A4	A4.IV;A4.V	0	0	2853	4390	1538	97	0	65	None
A4	A4.I;A4.V	0.0002	0.00083	2870	3512	643	35	0	86	None
A4	A4.III;A4.V	0.0003	0.00126	2870	3596	727	38	0	78	None
A4	A4.I;A4.V	0.0027	0.00953	3989	4302	314	29	0	86	None
A4	A4.II;A4.V	0	0.00001	3989	4390	402	45	0	86	None
A4	A4.II;A4.IV	0.0061	0.01934	3989	6492	2504	81	0	30	None
A4	A4.III;A4.V	0	0	3989	6493	2505	82	0	78	None
A4	A4.I;A4.IV	0.0085	0.02772	4311	6557	2247	59	0	40	None
A4	A4.I;A4.V	0.0002	0.00083	4392	6492	2101	35	0	86	None
A4	A4.IV;A4.V	0.0114	0.03601	4392	6492	2101	35	0	65	None
A4	A4.II;A4.V	0.0002	0.00055	4392	6493	2102	36	0	86	None
A4	A4.V;A4.composite	0.0133	0.03988	3424	3596	173	6	0	218	None
A4	A4.V;A4.composite	0.0001	0.00014	7278	7777	500	9	0	218	None
A4	A4.III;A4.composite	0.0198	0.05244	8636	8974	339	9	0	180	None
A4	A4.I;A4.composite	0	0	8997	9480	484	39	0	169	None
A4	A4.IV;A4.composite	0.0002	0.00058	9105	9228	124	13	0	178	None
A4	A4.III;A4.composite	0	0.00013	9105	9281	177	14	0	180	None
A4	A4.II;A4.composite	0	0	9312	9480	169	18	0	180	None
A4	A4.III;A4.composite	0	0	9312	9480	169	18	0	180	None
A4	A4.IV;A4.composite	0	0	9312	9480	169	18	0	178	None
A4	A4.II;A4.composite	0.0198	0.05244	9482	9557	76	9	0	180	None
A4	A4.III;A4.composite	0.0198	0.05244	9482	9557	76	9	0	180	None
A4	A4.IV;A4.composite	0.0252	0.06528	9482	9557	76	9	0	178	None
A5	A5.II;A5.V	0.0003	0.00227	2516	3419	904	35	0	80	None
A5	A5.I.insert;A5.V	0.0009	0.00397	2790	3419	630	34	0	79	None
A5	A5.IV;A5.V	0.0051	0.02533	2870	3419	550	29	0	79	None
A5	A5.III;A5.V	0.0112	0.03976	2870	3419	550	29	0	76	None
A5	A5.I.insert;A5.II	0.0014	0.00734	3579	7832	4254	130	0	20	None
A5	A5.II;A5.V	0	0	3986	7271	3286	113	0	80	None

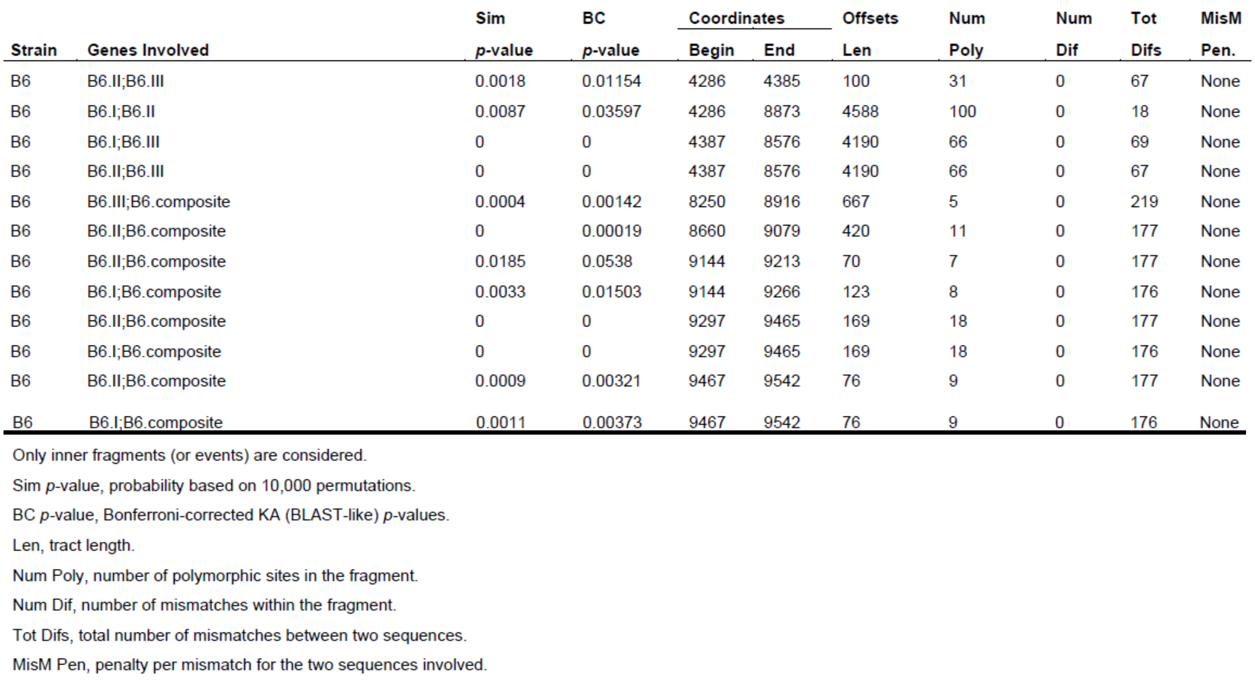
Supplementary Table S2.14. Gene conversion events detected in the *Sdic* region for eight strains of *D. melanogaster* according to GenConv

		Sim	BC	Coordinates		Offsets	Num	Num	Tot	MisM
Strain	Genes Involved	*p*-value	*p*-value	Begin	End	Len	Poly	Dif	Difs	Pen.
A5	A5.I.insert;A5.V	0	0	3986	7271	3286	113	0	79	None
A5	A5.III;A5.V	0	0.00002	3988	4389	402	50	0	76	None
A5	A5.IV;A5.V	0.0026	0.01207	4267	4389	123	31	0	79	None
A5	A5.IV;A5.V	0.0002	0.00131	6275	7236	962	37	0	79	None
A5	A5.III;A5.V	0.0003	0.00235	6275	7236	962	37	0	76	None
A5	A5.V;A5.composite	0.045	0.13905	3421	3604	184	5	0	221	None
A5	A5.V;A5.composite	0	0	7274	7874	601	12	0	221	None
A5	A5.III;A5.composite	0.0025	0.0115	7775	8327	553	10	0	181	None
A5	A5.I.insert;A5.composite	0.0001	0.00083	8329	9090	762	12	0	182	None
A5	A5.III;A5.composite	0	0	9308	9476	169	18	0	181	None
A5	A5.I.insert;A5.composite	0	0	9308	9553	246	28	0	182	None
A5	A5.IV;A5.composite	0	0	9308	9553	246	28	0	178	None
A5	A5.II;A5.composite	0	0	9308	9553	246	28	0	174	None
A5	A5.III;A5.composite	0.0109	0.03957	9478	9553	76	9	0	181	None
A7	A7.II;A7.IV	0.0004	0.00194	2304	3398	1095	37	0	73	None
A7	A7.III;A7.IV	0	0.00013	2304	3571	1268	44	0	74	None
A7	A7.I;A7.IV	0.0001	0.00067	2315	3515	1201	42	0	70	None
A7	A7.II;A7.IV	0	0.00001	3989	4384	396	54	0	73	None
A7	A7.I;A7.IV	0	0.00002	3991	4384	394	53	0	70	None
A7	A7.III;A7.IV	0	0	3991	4438	448	65	0	74	None
A7	A7.I;A7.II	0	0.00009	3991	7209	3219	113	0	31	None
A7	A7.III;A7.IV	0.0006	0.00322	6256	7209	954	35	0	74	None
A7	A7.II;A7.IV	0.0006	0.00388	6256	7209	954	35	0	73	None
A7	A7.I;A7.IV	0.0006	0.00484	6256	7228	973	36	0	70	None
A7	A7.I;A7.composite	0.0222	0.06146	3428	4128	701	9	0	169	None
A7	A7.II;A7.composite	0.0297	0.07748	7266	7844	579	8	0	178	None
A7	A7.IV;A7.composite	0	0	7266	8661	1396	14	0	220	None
A7	A7.III;A7.composite	0	0.00026	8663	9082	420	12	0	182	None
A7	A7.I;A7.composite	0	0	8715	9269	555	26	0	169	None
A7	A7.III;A7.composite	0	0.00002	9084	9269	186	14	0	182	None
A7	A7.III;A7.composite	0	0	9300	9468	169	18	0	182	None
A7	A7.I;A7.composite	0	0	9300	9468	169	18	0	169	None
A7	A7.II;A7.composite	0	0	9300	9545	246	28	0	178	None
A7	A7.III;A7.composite	0.0036	0.01345	9470	9545	76	9	0	182	None
A7	A7.I;A7.composite	0.0222	0.06146	9470	9545	76	9	0	169	None
B1	B1.I;B1.IV	0.0007	0.00391	2853	3138	286	32	0	78	None
B1	B1.III;B1.IV	0.0018	0.00631	2853	3632	780	37	0	66	None
B1	B1.II;B1.IV	0.0054	0.01609	2870	3666	797	34	0	66	None
B1	B1.II;B1.IV	0	0	3668	7740	4073	121	0	66	None
B1	B1.I;B1.IV	0.002	0.00838	3985	4300	316	30	0	78	None
B1	B1.III;B1.IV	0.0002	0.00038	3985	4388	404	46	0	66	None
B1	B1.I;B1.IV	0.0408	0.12025	4390	6270	1881	23	0	78	None
B1	B1.III;B1.IV	0	0	4390	7268	2879	64	0	66	None
B1	B1.II;B1.composite	0.0007	0.00314	7271	8319	1049	11	0	172	None
B1	B1.IV;B1.composite	0.0002	0.00032	8187	8919	733	7	0	220	None
B1	B1.I;B1.composite	0	0	8715	9545	831	56	0	170	None
B1	B1.II;B1.composite	0.037	0.11466	9147	9268	122	8	0	172	None
B1	B1.III;B1.composite	0	0	9299	9545	247	28	0	179	None
B1	B1.II;B1.composite	0	0	9299	9545	247	28	0	172	None
B2	B2.V;B2.VI	0.0004	0.00248	2658	3419	762	36	0	79	None
B2	B2.II;B2.VI	0.0008	0.00523	2790	3419	630	35	0	77	None
B2	B2.I;B2.VI	0.0297	0.1333	2870	3138	269	28	0	72	None

Supplementary Table S2.14. Gene conversion events detected in the *Sdic* region for eight strains of *D. melanogaster* according to GenConv

		Sim	BC	Coordinates		Offsets	Num	Num	Tot	MisM
Strain	Genes Involved	*p*-value	*p*-value	Begin	End	Len	Poly	Dif	Difs	Pen.
B2	B2.IV;B2.VI	0.0036	0.01686	2870	3419	550	30	0	81	None
B2	B2.III;B2.VI	0.0049	0.02321	2870	3419	550	30	0	79	None
B2	B2.III;B2.V	0.0042	0.01981	3723	7236	3514	116	0	21	None
B2	B2.I;B2.VI	0.0096	0.04914	3986	4301	316	31	0	72	None
B2	B2.II;B2.VI	0	0.00007	3986	4389	404	47	0	77	None
B2	B2.III;B2.VI	0	0	3986	6273	2288	70	0	79	None
B2	B2.V;B2.VI	0	0	3986	6273	2288	70	0	79	None
B2	B2.IV;B2.VI	0	0	3988	6273	2286	69	0	81	None
B2	B2.III;B2.IV	0.0037	0.01705	3988	7236	3249	108	0	23	None
B2	B2.IV;B2.V	0.0007	0.00432	3988	7658	3671	113	0	25	None
B2	B2.I;B2.VI	0.0007	0.00477	4310	6273	1964	38	0	72	None
B2	B2.I;B2.IV	0.001	0.00581	4310	7116	2807	73	0	38	None
B2	B2.I;B2.III	0.0057	0.02681	4310	7116	2807	73	0	33	None
B2	B2.I;B2.V	0.0057	0.02681	4310	7116	2807	73	0	33	None
B2	B2.II;B2.VI	0	0	4391	7271	2881	62	0	77	None
B2	B2.I;B2.VI	0.0041	0.01809	6275	7116	842	34	0	72	None
B2	B2.IV;B2.VI	0	0.00078	6275	7236	962	38	0	81	None
B2	B2.V;B2.VI	0	0.00118	6275	7236	962	38	0	79	None
B2	B2.III;B2.VI	0	0.00081	6275	7271	997	39	0	79	None
B2	B2.VI;B2.composite	0.0056	0.02656	3606	3802	197	6	0	220	None
B2	B2.II;B2.composite	0.001	0.00532	7274	8322	1049	12	0	171	None
B2	B2.I;B2.composite	0	0	8666	9548	883	56	0	165	None
B2	B2.IV;B2.composite	0	0	9150	9471	322	27	0	181	None
B2	B2.III;B2.composite	0	0	9302	9548	247	29	0	172	None
B2	B2.II;B2.composite	0	0	9302	9548	247	29	0	171	None
B2	B2.V;B2.composite	0	0	9303	9548	246	28	0	169	None
B2	B2.IV;B2.composite	0.0113	0.05079	9473	9548	76	9	0	181	None
B3	B3.II;B3.IV.insertion	0	0.00004	2300	2988	689	29	0	142	None
B3	B3.III;B3.IV.insertion	0	0.00013	2443	2988	546	26	0	147	None
B3	B3.I;B3.IV.insertion	0.0009	0.00174	2790	2988	199	22	0	146	None
B3	B3.II;B3.IV.insertion	0	0	4016	4411	396	55	0	142	None
B3	B3.III;B3.IV.insertion	0	0	4018	4411	394	54	0	147	None
B3	B3.I;B3.IV.insertion	0	0	4018	6641	2624	106	0	146	None
B3	B3.II;B3.III	0.0003	0.0012	4018	7624	3607	117	0	32	None
B3	B3.III;B3.IV.insertion	0	0	4413	6499	2087	34	0	147	None
B3	B3.II;B3.IV.insertion	0	0	4413	6499	2087	34	0	142	None
B3	B3.III;B3.IV.insertion	0.0197	0.06563	6501	6641	141	16	0	147	None
B3	B3.II;B3.IV.insertion	0.0319	0.1009	6501	6641	141	16	0	142	None
B3	B3.I;B3.II	0.0045	0.01443	9333	9600	268	91	0	33	None
B3	B3.II;B3.III	0.0064	0.01928	9333	9600	268	91	0	32	None
B3	B3.IV.insertion;B3.composite	0.003	0.00901	3291	3634	344	5	0	291	None
B3	B3.IV.insertion;B3.composite	0.0414	0.11335	7298	7749	452	4	0	291	None
B3	B3.III;B3.composite	0.007	0.01974	9126	9249	124	14	0	176	None
B3	B3.I;B3.composite	0	0	9333	9504	172	65	0	179	None
B3	B3.II;B3.composite	0	0	9333	9504	172	65	0	179	None
B3	B3.III;B3.composite	0	0	9333	9504	172	65	0	176	None
B3	B3.I;B3.composite	0	0	9506	9600	95	25	0	179	None
B3	B3.II;B3.composite	0	0	9506	9600	95	25	0	179	None
B3	B3.III;B3.composite	0	0	9506	9600	95	25	0	176	None
B6	B6.II;B6.III	0.0002	0.00109	2857	3581	725	38	0	67	None
B6	B6.I;B6.III	0.0014	0.00577	2874	3426	553	32	0	69	None
B6	B6.I;B6.III	0	0.00002	3896	4385	490	48	0	69	None

Supplementary Table S2.14. Gene conversion events detected in the *Sdic* region for eight strains of *D. melanogaster* according to GenConv



		Sim	BC	Coordinates		Offsets	Num	Num	Tot	MisM
Strain	Genes Involved	*p*-value	*p*-value	Begin	End	Len	Poly	Dif	Difs	Pen.
B6	B6.II;B6.III	0.0018	0.01154	4286	4385	100	31	0	67	None
B6	B6.I;B6.II	0.0087	0.03597	4286	8873	4588	100	0	18	None
B6	B6.I;B6.III	0	0	4387	8576	4190	66	0	69	None
B6	B6.II;B6.III	0	0	4387	8576	4190	66	0	67	None
B6	B6.III;B6.composite	0.0004	0.00142	8250	8916	667	5	0	219	None
B6	B6.II;B6.composite	0	0.00019	8660	9079	420	11	0	177	None
B6	B6.II;B6.composite	0.0185	0.0538	9144	9213	70	7	0	177	None
B6	B6.I;B6.composite	0.0033	0.01503	9144	9266	123	8	0	176	None
B6	B6.II;B6.composite	0	0	9297	9465	169	18	0	177	None
B6	B6.I;B6.composite	0	0	9297	9465	169	18	0	176	None
B6	B6.II;B6.composite	0.0009	0.00321	9467	9542	76	9	0	177	None
B6	B6.I;B6.composite	0.0011	0.00373	9467	9542	76	9	0	176	None

Only inner fragments (or events) are considered.

Sim *p*-value, probability based on 10,000 permutations.

BC *p*-value, Bonferroni-corrected KA (BLAST-like) *p*-values.

Len, tract length.

Num Poly, number of polymorphic sites in the fragment.

Num Dif, number of mismatches within the fragment.

Tot Difs, total number of mismatches between two sequences.

MisM Pen, penalty per mismatch for the two sequences involved.

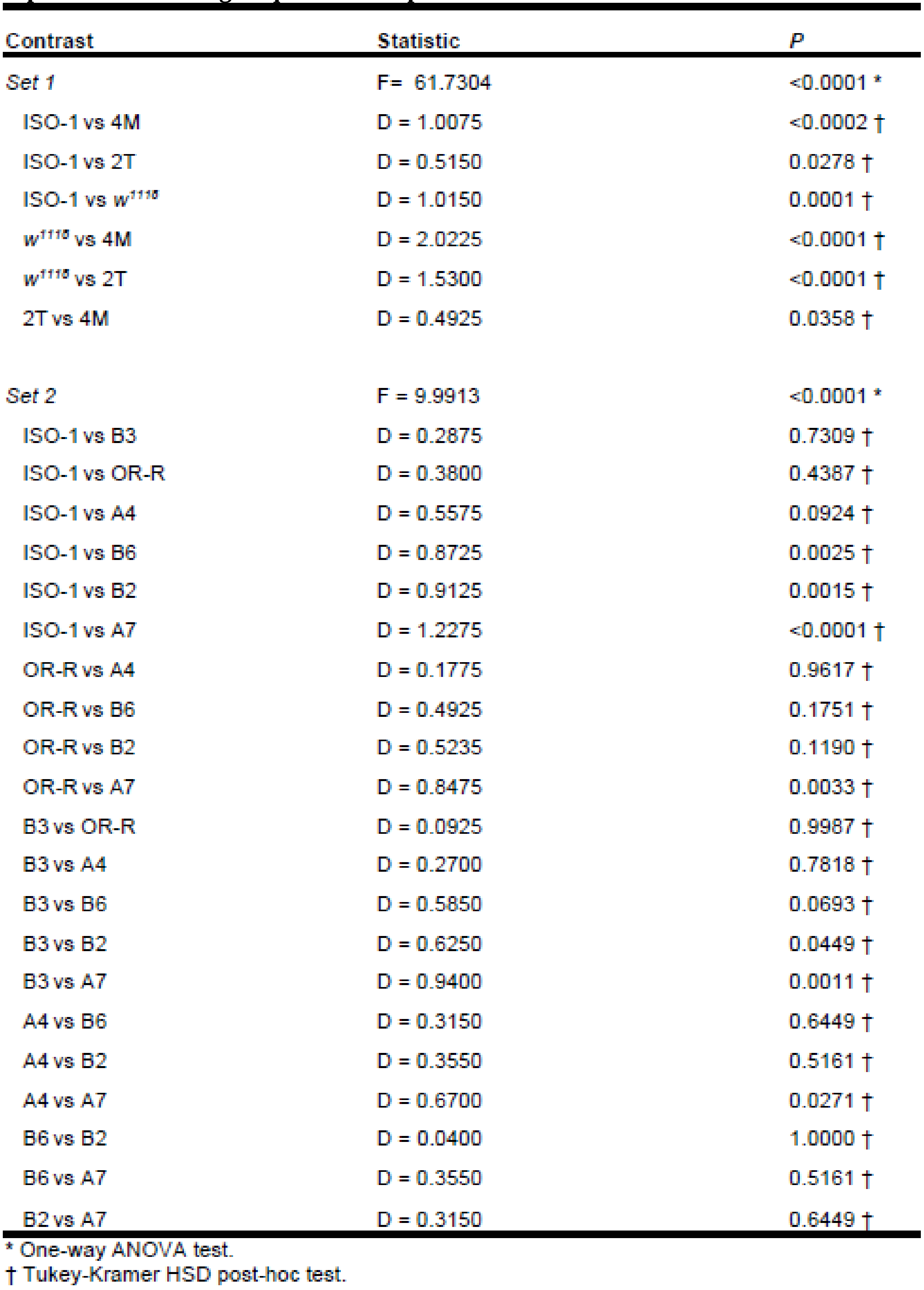
Supplementary Table S2.15. Evolution mode across partitions of the *Sdic* repeat as delineated with ACG

Partition *	Branch †	LRT	*P*-value (FDR)	Node and Subtree †
				omposite)Node12)Node6)Node4,ISO1_Berlin_composite)Node3)Node1
P5.0a	43	65.480	< 1.00E-010	[Node16] Internal Branch Rooting (A4_II,((((A4_I,(ISO1_Berlin_SdicI_2,B1_I)Node23)Node21,B2_I)Node20,(ISO1_Berlin_SdicIV_3,ISO1_Berlin_SdicII_C)Node27)Node19,((((A7_III,(((B3_IV_insertion,B3_I)Node37,(((A5_I_insert,B1_IV)Node42,(A5_V,A5_II)Node45)Node41,(B2_II,B1_II)Node48)Node40)Node36,(((A7_IV,A4_V)Node53,B6_III)Node52,A4_III)Node51)Node35)Node33,(B2_VI,B1_III)Node58)Node32,(B2_III,(((((ISO1_Berlin_SdicVI_1,((B3_II,(A7_II,ISO1_Berlin_SdicV_4)Node71)Node69,B3_III)Node68)Node66,(((A5_IV,A7_I)Node77,B6_I)Node76,(A5_III,B6_II)Node81)Node75)Node65,B2_V)Node64,B2_IV)Node63)Node61)Node31,(ISO1_Berlin_SdicIII_B,A4_IV)Node86)Node30)Node18)Node16
P5.0a	7	25.591	4.22E-07	[Node1] Internal Branch Rooting ((((B2_composite,((B3_composite,A7_composite)Node7,(B6_composite,B1_composite)Node10)Node6)Node4,A5_composite)Node3,ISO1_Berlin_composite)Node2,A4_composite)Node1
P6	42	22.381	2.24E-06	[Node76] Internal Branch Rooting ((((B1_IV,B6_III)Node79,B2_VI)Node78,(A4_V,(A7_IV,ISO1_Berlin_SdicVI_1)Node85)Node83)Node77,A5_V)Node76
P6.2	42	32.350	1.29E-08	[Node75] Internal Branch Rooting (ISO1_Berlin_SdicVI_1,((((B2_VI,A4_V)Node80,A7_IV)Node79,A5_V)Node78,(B6_III,B1_IV)Node85)Node77)Node75
P6.2	88	236.574	< 1.00E-010	[B3_IV_insertion] Leaf node B3_IV_insertion

* As shown in Fig. S8.

† The branch number in Hyphy and its associated node and subtree are relative to the PARTITION SPECIFIC gene tree.

Supplementary Table S2.16. Statistical support for differences in *Sdic male* expression according to qRT-PCR experiments



Contrast	Statistic	*P*
Set 1	F= 61.7304	<0.0001 *
ISO-1 vs 4M	D = 1.0075	<0.0002 †
ISO-1 vs 2T	D = 0.5150	0.0278 †
ISO-1 vs w^{1118}	D = 1.0150	0.0001 †
w^{1118} vs 4M	D = 2.0225	<0.0001 †
w^{1118} vs 2T	D = 1.5300	<0.0001 †
2T vs 4M	D = 0.4925	0.0358 †
Set 2	F = 9.9913	<0.0001 *
ISO-1 vs B3	D = 0.2875	0.7309 †
ISO-1 vs OR-R	D = 0.3800	0.4387 †
ISO-1 vs A4	D = 0.5575	0.0924 †
ISO-1 vs B6	D = 0.8725	0.0025 †
ISO-1 vs B2	D = 0.9125	0.0015 †
ISO-1 vs A7	D = 1.2275	<0.0001 †
OR-R vs A4	D = 0.1775	0.9617 †
OR-R vs B6	D = 0.4925	0.1751 †
OR-R vs B2	D = 0.5235	0.1190 †
OR-R vs A7	D = 0.8475	0.0033 †
B3 vs OR-R	D = 0.0925	0.9987 †
B3 vs A4	D = 0.2700	0.7818 †
B3 vs B6	D = 0.5850	0.0693 †
B3 vs B2	D = 0.6250	0.0449 †
B3 vs A7	D = 0.9400	0.0011 †
A4 vs B6	D = 0.3150	0.6449 †
A4 vs B2	D = 0.3550	0.5161 †
A4 vs A7	D = 0.6700	0.0271 †
B6 vs B2	D = 0.0400	1.0000 †
B6 vs A7	D = 0.3550	0.5161 †
B2 vs A7	D = 0.3150	0.6449 †

* One-way ANOVA test.

† Tukey-Kramer HSD post-hoc test.

Supplementary Table S2.17. Statistical support for differences in sperm competitive ability in offense assays

Competing Experimental Males	Statistic	*P* *
Strain set 1		
w^{1118} vs 2T	D = 5.0432	0.6627
w^{1118} vs A^{-}	D = 18.4266	0.0007
w^{1118} vs B^{+}	D = 5.1064	0.625
B^{+} vs 2T	D = -0.6481	0.9991
B^{+} vs A^{-}	D = 13.9056	0.0263
A^{-} vs 2T	D = -15.8999	0.0103
Strain set 2		
w^{1118} vs 4M	D = 19.1928	0.0016
w^{1118} vs E^{-}	D =18.8472	0.0062
w^{1118} vs I^{+}	D = 1.4688	0.9897
I^{+} vs 4M	D = 17.7019	0.0048
I^{+} vs E^{-}	D = 16.7507	0.0166
E^{-} vs 4M	D = 6.3411	0.7216

* According to the Stell-Dwass method.

SUPPLEMENTARY REFERENCES

Abyzov A, Urban AE, Snyder M, Gerstein M 2011. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research* 21: 974-984.

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al.* 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287: 2185-2195.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389-3402.

Benjamini Y, Hochberg Y 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* 57: 289-300.

Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology* 33: 623-630.

Bray JR, Curtis JT 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs* 27: 326-349.

Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ 2016. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.*

Chakraborty M, Emerson JJ, Macdonald SJ, Long AD 2019. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat Commun* 10: 4872.

Chakraborty M, VanKuren NW, Zhao R, Zhang X, Kalsow S, Emerson JJ 2018. Hidden genetic variation shapes the structure of functional elements in *Drosophila*. *Nat Genet* 50: 20-25.

Clifton BD, Librado P, Yeh SD, Solares ES, Real DA, Jayasekera SU, Zhang W, Shi M, Park RV, Magie RD, *et al.* 2017. Rapid Functional and Sequence Differentiation of a Tandemly Repeated Species-Specific Multigene Family in *Drosophila*. *Mol Biol Evol* 34: 51-65.

dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, Emmert DB, Gelbart WM, FlyBase C 2015. FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res* 43: D690-697.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, *et al.* 2011. The developmental transcriptome of *Drosophila melanogaster*. *Nature* 471: 473-479.

Grenier JK, Arguello JR, Moreira MC, Gottipati S, Mohammed J, Hackett SR, Boughton R, Greenberg AJ, Clark AG 2015. Global diversity lines - a five-continent reference panel of sequenced *Drosophila melanogaster* strains. *G3* (Bethesda, Md) 5: 593-603.

Hu TT, Eisen MB, Thornton KR, Andolfatto P 2013. A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. *Genome research* 23: 89-98.

Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, Chin CS, Rapicavoli NA, Rank DR, Li J, *et al.* 2014. Long-read, whole-genome shotgun sequence data for five model organisms. *Sci Data* 1: 140045.

King EG, Macdonald SJ, Long AD 2012a. Properties and power of the *Drosophila* Synthetic Population Resource for the routine dissection of complex traits. *Genetics* 191: 935-949.

King EG, Merkes CM, McNeil CL, Hoofer SR, Sen S, Broman KW, Long AD, Macdonald SJ 2012b. Genetic dissection of a model complex trait using the *Drosophila* Synthetic Population Resource. *Genome research* 22: 1558-1566.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* 35: 1547-1549.

Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE 2016a. A Thousand Fly Genomes: An Expanded *Drosophila* Genome Nexus. *Mol Biol Evol* 33: 3308-3313.

Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE 2016b. A Thousand Fly Genomes: An Expanded *Drosophila* Genome Nexus. Mol Biol Evol 33: 3308-3313.

Langley CH, Stevens K, Cardeno C, Lee YC, Schrider DR, Pool JE, Langley SA, Suarez C, Corbett-Detig RB, Kolaczkowski B*, et al.* 2012. Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics* 192: 533-598.

Ma J, An K, Zhou JB, Wu NS, Wang Y, Ye ZQ, Wu YD 2019. WDSPdb: an updated resource for WD40 proteins. *Bioinformatics* (Oxford, England).

Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL 1998. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature* 396: 572-575.

Parks AL, Cook KR, Belvin M, Dompe NA, Fawcett R, Huppert K, Tan LR, Winter CG, Bogart KP, Deal JE*, et al.* 2004. Systematic generation of high-resolution deletion coverage of the *Drosophila melanogaster* genome. *Nat Genet* 36: 288-292.

Ranz J, Clifton B 2019. Characterization and evolutionary dynamics of complex regions in eukaryotic genomes. *Sci China Life Sci.*

Ryder E, Blows F, Ashburner M, Bautista-Llacer R, Coulson D, Drummond J, Webster J, Gubb D, Gunton N, Johnson G*, et al.* 2004. The DrosDel collection: a set of P-element insertions for generating custom chromosomal aberrations in *Drosophila melanogaster*. *Genetics* 167: 797-813.

Sawyer S 1989. Statistical tests for detecting gene conversion. *Mol Biol Evol* 6: 526-538.

Solares EA, Chakraborty M, Miller DE, Kalsow S, Hall K, Perera AG, Emerson JJ, Hawley RS 2018. Rapid Low-Cost Assembly of the *Drosophila melanogaster* Reference Genome Using Low-Coverage, Long-Read Sequencing. *G3* (Bethesda, Md) 8: 3143-3154.

Yeh SD, Do T, Abbassi M, Ranz JM 2012a. Functional relevance of the newly evolved sperm dynein intermediate chain multigene family in *Drosophila melanogaster* males. *Commun Integr Biol* 5: 462-465.

Yeh SD, Do T, Chan C, Cordova A, Carranza F, Yamamoto EA, Abbassi M, Gandasetiawan KA, Librado P, Damia E*, et al.* 2012b. Functional evidence that a recently evolved *Drosophila* sperm-specific gene boosts sperm competition. *Proceedings of the National Academy of Sciences of the United States of America* 109: 2043-2048.

CHAPTER 3

Paralog Transcriptional Divergence in the *D. melanogaster*-specific Tandem Multigene Family *Sdic*

ABSTRACT

Understanding the interplay between evolving functional elements of tandem gene duplicates and their probability of retention in the genome of a species requires detailed characterizations of individual paralogs prior or soon after fixation. The repetitive nature and high sequence similarity among paralogs of young multigene families has previously limited the ability to resolve the sequence of individual paralogs. The *D. melanogaster*-specific multigene family *Sperm-specific dynein intermediate chain* (*Sdic*) is a recent tandem expansion of a defective copy of the essential gene *short wing* (*sw*), which encodes an intermediate chain subunit of the cytoplasmic dynein motor protein complex. The *Sdic* region has been reconstructed at the nucleotide level across multiple reference-quality genome assemblies from geographically diverse strains containing different *Sdic* copy numbers and paralog identities. Quantifying the expression of individual paralogs by targeting copy-specific differences has enabled an accurate functional characterization of this multigene family at the intra- and inter-population levels. Here we used RNA-seq data to quantify the expression of *Sdic* paralogs in testes, male heads, and ovaries from four populations, paying special attention to the impact of *cis*- and *trans*-acting regulatory variation. Contrary to previously reported, transcripts from all *Sdic* paralogs examined were detected in testes but not male heads or ovaries. Unlike seen using whole bodies, we find evidence of a positive correlation between *Sdic* copy number and the aggregate expression level of all *Sdic* paralogs in these strains. We detected evidence of differential expression among paralogs within each strain, with a paralog associated with a transposable element insertion showing the highest expression. Further, we documented a negative correlation between *sw* and aggregate *Sdic*

expression, which suggests a possible stoichiometric constraint on total cytoplasmic dynein intermediate chain dosage required for protein complex assembly. At the regulatory level, we found no indication that the promoter type nor the position of a paralog within the *Sdic* cluster can explain its expression level relative to other paralogs in the same strain. In contrast, through RNA-sequencing testis, male heads, and male accessory glands from a set of synthetic genotypes with identical genomes except for different *Y* chromosomes, we detected a weak but significant effect of *trans*-regulatory variation associated with the *Y* chromosome on the expression of *Sdic* in male accessory glands. This work highlights the importance of combining molecular and sequencing approaches to obtain paralog-specific information for generating a more nuanced portrait of how recently originated multigene families functionally evolve along their path to fixation and consolidation in the genome.

INTRODUCTION

Novel gene functions most commonly originate through duplication of existing genes followed by functional divergence among the retained duplicates at the coding and/or expression levels. While most duplicates are quickly lost or decay as pseudogenes through *nonfunctionalization* (Force et al 1999), various scenarios, such as advantageous alterations in total gene dosage or functional differentiation among the duplicates via *neofunctionalization* and *subfunctionalization*, have been proposed to explain the retention and fixation of gene duplicates (see Innan & Kondrashov 2010 for a comprehensive list of evolutionary scenarios; Kuzmin et al 2022). The *structural and functional entanglement model* proposes paralogs will follow these various evolutionary trajectories based on the extent of their structural and functional entanglement, *i.e.*, their structural ability to partition their distinct functions due to constraints on overlapping subsets of functions (Kuzmin et al 2020). While this and other models can propose mechanisms by which new functions arise, comprehensive functional characterizations of young polymorphic duplicates that test specific hypotheses that evaluate the applicability of these models to the evolution of the paralogs of particular gene families are lacking. Therefore, how gene families overcome the immediate consequences of gene duplication, *i.e.*, dosage increase, and potentially accumulate the molecular diversity required for novel functions, while being impacted by genetic drift and natural selection along the path to fixation, remains largely uncharacterized.

Tandemly arranged multigene families primarily originate through DNA-based duplication events mediated by non-allelic homologous recombination (NAHR) events (Hastings et al 2009). Recent gene duplications unique to individual species have been shown to impact organismal fitness and contribute to phenotypic change in that species (Yeh et al 2012; Jugulam et al. 2014; Florio et al 2015; Fiddes et al 2018; Chakraborty et al. 2019), playing key roles in adaptation, phenotypic diversification, and genetic innovation (Brown et al. 1998; Newcomb et al. 2005; Perry et al. 2007; Jugulam et al. 2014). Uncovering the mechanisms that shape the functional attributes of individual paralogs within these families during their early evolutionary stages has been precluded by three major difficulties. First, repetitive regions composed of multiple highly similar tandem repeats, *i.e.*, *structurally complex genomic regions*, remain refractory to accurate sequence reconstruction in even ‘reference quality’ genome assemblies (Clifton et al 2017; 2020). Second, the genomic regions often display copy number variation (CNV), involving duplicates with high sequence identity that result from NAHR and gene conversion events (Clifton et al 2020; Loehlin et al 2021). Third, the rules that govern the expression of gene duplicates as they age are still not well understood (Kondrashov 2010; Rody et al 2017; Teufel et al 2018; Loehlin et al 2021; Kuzmin et al 2022). Overall, these difficulties have resulted in a scarcity of intraspecific studies that can properly evaluate paralog diversity and functional dynamics of tandem multigene families at the early stages of their formation and consolidation in eukaryotic genomes.

Studies quantifying total mRNA expression from identical or almost identical complete duplicates in tandem (Hayward et al 2017; Konrad et al 2018; Clifton et al 2020; Zhang et al 2022) or inserted as transgenes (Loehlin & Carroll 2016) showed a trend of a >2-fold increase in transcript level, inconsistent with the previously expected *2-fold*, *i.e.*, the *dosage-additivity hypothesis* (Loehlin et al 2021). Further, the combined transcript abundance of older duplicate pairs that have accumulated multiple mutations in *cis*- and affect *trans*-factors typically show <2-fold expression, both within the same species (Cardoso-Moreira et al 2016; Rogers et al 2017) and between distinct species (Lan & Pritchard 2016). Indeed, most young tandem duplicate pairs express at the single-copy level through *dosage sharing* (Lan & Pritchard 2016), as expression changes as low as 1.2-fold can affect fitness (Yan et al 2002; Loehlin et al 2021). In addition, a study comparing expression of the tandem gene family *Acsx1* in *D. melanogaster* revealed that while doublets showed >2-fold expression compared to singlets, triplets did not differ in level of expression from doublets (Loehlin et al 2021). In general, it has been postulated that the chance of a new duplicate persisting in a population is dependent on whether the magnitude in dosage change associated with the initial sequentially identical duplication event is enough to have a measurable effect on fitness (Loehlin et al 2021; Konrad et al 2018; Zhang et al 2022). Protein complex subunits are particularly sensitive to dosage changes, as stoichiometry plays an important role in proper protein complex assembly (Zhang et al 2022). Collectively, these findings point to a complex relationship between total dosage of the product encoded by tandemly arranged multigene families, number of paralogs, age of the paralogs, and protein function.

Sperm-specific dynein intermediate chain (*Sdic*) is a newly evolved tandem multigene family present only in *D. melanogaster*, so therefore originated after the split of the *melanogaster* lineage from the *simulans* clade ~1.4 Ma (Nurminsky et al. 1998; Obbard et al. 2012). This species-specific multigene family is one of the few genetic factors known to influence sperm competition (Civetta & Ranz 2019), *i.e.,* a form of sexual selection that biases fertilization at the postcopulatory level, through an impact on sperm competitive ability (Yeh et al 2012; Jayaswal et al 2018). Two studies that combined the annotation of reference-quality genome assemblies from different populations with the implementation of qPCR and read-depth analyses (Clifton et al. 2017, 2020), helped delineate a faithful reconstruction of the *Sdic* region in the reference strain of *D. melanogaster* (ISO-1) as well as reveal extensive CNV at this region, which is composed of a single fixed paralog and multiple segregating paralogs (Clifton et al 2020). The original *Sdic* copy originated from a segmental duplication on the *X* chromosome involving two adjacent genes, *short wing* (*sw*) and *Annexin B10* (*AnxB10*), in which the central genes fused into a chimeric entity that essentially encodes a defective form of the sw protein, a cytoplasmic dynein intermediate chain, *i.e.*, a regulatory subunit of the cytoplasmic dynein protein complex (Nurminsky et al. 1998; Kardon & Vale 2009). The repetitive nature and high sequence similarity among *Sdic* paralogs and the flanking parental genes has likely facilitated recurrent NAHR events, resulting in repeated contractions and expansions of the tandem array (Clifton et al 2020; Hastings et al. 2009), as well as rampant gene conversion, which contributes to high sequence identity levels among the repeats (Clifton et al 2017). All *Sdic* paralogs reliably annotated show no sequence features indicative of pseudogenization. Nevertheless, no correlation between copy number (CN) and total *Sdic* expression in male whole-bodies was found across a geographically diverse set of six inbred strains, pointing to regulatory variation as a more relevant factor in influencing the expression of the *Sdic* family compared to the number of active copies (Clifton et al 2020). However, whole-body expression profiling can mask tissue-specific expression patterns, leaving it unclear how the apparent buffering mechanisms that maintain the lack of correlation between *Sdic* mRNA abundance and copy number hold at the tissue level.

The distinct functional attributes of each paralog are another key factor that can impact the probability of paralog retention. *Sdic* paralogs are not only expressed in the male germline but also, although to a lesser extent, in various somatic tissues of both sexes throughout the lifetime – including male accessory glands, heads of both sexes, and ovaries–, which is compatible with ongoing temporal and spatial paralog functional specialization (Clifton et al 2017). The genetic basis of this functional divergence remains unclear. *Cis*-acting variation among duplicates is known to influence most novel expression patterns, with *trans* variation impacting a smaller but relevant proportion of duplicates (Cridland et al 2020). Intraspecific *cis* variation found among *Sdic* paralogs affects promoter sequences, the predicted miRNA binding site composition in 3’UTRs, TE insertions, and premature stop codons (Clifton at al. 2017, 2020), the latter having the potential to impede copy functionality by activating the nonsense-mediated decay pathway (Catalan et al. 2016; Hug et al. 2016; Scott et al. 2016). *Cis*-regulatory mechanisms can also differentially affect the expression of a duplicate depending on its position within the tandem array (Loehlin et al 2021). Further, *trans*-acting factors are also speculated to impact interpopulation *Sdic* expression variation (Clifton et al. 2020). One such *trans*-acting mechanism is known as *Y-linked regulatory variation* (YRV), *i.e.*, the effect of *Y* chromosome variation on genome-wide patterns of expression diversity in male reproductive (Lemos et al 2008; Jiang et al 2010; Sackton et al 2011; Ågren et al 2020) and somatic tissues (Branco et al 2017; Ågren et al 2020). The impact of YRV on expression variation is not exerted through the ~20 mostly monomorphic protein coding genes resident on the *Y* chromosome of *D. melanogaster* (Zurovcova & Eanes 1999; Chang and Larracuente 2019), but through variation in the repetitive, noncoding heterochromatic regions of this chromosome (Lyckegaard & Clark 1989; Lemos et al. 2008, 2010; Sackton et al. 2011). The extent to which *Sdic* paralogs are diverging in their expression attributes across populations at the tissue level remains unexplored, as does a more precise evaluation of the role of *cis*- and *trans*-acting factors on the regulation of *Sdic* expression. For example, no precise analysis of the impact of *Y*-linked regulatory variation on the tissue level expression of the individual *Sdic* paralogs has been performed using an accurate annotation of the *Sdic* region (Branco et al 2017; Wang et al 2018).

Here we perform a detailed interpopulation gene expression analysis to gain key insights into the molecular basis of the regulation and variation of *Sdic* expression attributes. Using RNA-seq, we identify and quantify RNA expression specific to individual *Sdic* paralogs and their parental gene *sw* within male heads, testes, and ovaries from a set of geographically diverse strains that differ in *Sdic* CN and paralog composition. We also scrutinize the impact of YRV on this gene family by quantifying *Sdic* expression across a set of strains differing only in their *Y* chromosome origin in whole bodies using qRT-PCR and in male heads, accessory glands, and testes using RNA-seq. Our results highlight the importance of integrating precise paralog-specific sequence information with tissue-level expression data to obtain accurate portraits of how the functional attributes of multigene families evolve.

RESULTS

Aggregate *Sdic* expression correlates with *Sdic* copy number in testes but not in ovaries or heads

Our previous qRT-PCR survey of expression variation across a set of isogenic strains of diverse geographically origin differing in the number of *Sdic* copies revealed that *Sdic* CNV was not positively correlated with aggregate *Sdic* expression (*Sdic*All), *i.e.*, the total expression when all *Sdic* copies are considered (Clifton et al 2020). In the qRT-PCR assays performed, aggregate expression was estimated using primers designed to target a fraction of *Sdic* coding sequence in which there is no nucleotide variation across paralogs both within and between strains. Nevertheless, conclusions from whole-body assays are limited as they can mask interpopulation differentiation in gene expression across tissues. To gain detailed knowledge about *Sdic* expression at the tissue level, we sequenced the transcriptome of testes, ovaries, and male heads across four isogenic strains from diverse continental origins (Panel I; table 3.1). For this, we used Illumina PE-100 RNA-sequencing. All the strains investigated possess a reference-quality genome assembly (Chakraborty et al 2018; Chakraborty et al 2019), in which the *Sdic* region has been precisely annotated, exhibiting CNV (3-6 copies) and varying compositions of *Sdic* paratypes and promoters (Clifton et al. 2020; fig. 3.1).

We tested for a positive correlation between the aggregate *Sdic* transcript abundance and CN across the three tissues examined. For that, we used a computational pipeline (Clifton et al. 2017; Clifton et al. 2020) that screens the sequencing reads for the presence of sequence intervals with perfect matches to a given sequence of interest (supplementary table S7). In this case, the sequence searched corresponded to the *de novo* evolved exon 1 of *Sdic*, which is identical across paralogs and strains and absent from the parental genes *sw* and *AnxB10*. Contrary to previous reports (Clifton et al 2017), we found just a few or no reads supporting *Sdic* expression in ovaries and male heads, a pattern consistent across strains. Not surprisingly, we found evidence for *Sdic* expression in testes but also a strong positive relationship between *Sdic* CN and total *Sdic* expression (fig. 3.2; r^2 = 0.8604, *P* = 1.386e-05). This result highlights the importance of tissue-level surveys of gene expression, as opposed to whole-body organisms which can mask biologically relevant patterns of expression.

A negative correlation between *sw* and aggregate *Sdic* expression suggests a possible stoichiometric constraint on total cytoplasmic dynein intermediate chain dosage required for protein complex assembly

Sdic paralogs share similar sequences with their parental gene *sw* at regions that encode protein-protein interaction domains, in particular regions that interacts with other subunits of the dynein complex (Jones et al 2014; Clifton et al 2017, 2020). Since *sw*'s protein function is dosage dependent (Boylan et al 2000), and the Sdic protein could be competing with sw for protein interactions with dynein complex subunits, we quantified expression of *sw* in these strains. Greater expression of *Sdic* correlated with greater expression of *sw* could be suggestive of a selective pressure on maintaining individual dosages driven by competition (Wei et al 2019), while a negative correlation would suggest that total dynein intermediate chain dosage is constrained by the stoichiometry of protein complex assembly. We found substantially lower expression of *sw* in relation to *Sdic*, as well as that both genes were differentially expressed across strains (fig. 3.3A, supplementary table S3.1; two-way ANOVA followed by pairwise Tukey HSD tests). Further, we found a negative correlation between aggregate *Sdic* and *sw* expression levels (fig. 3.3B; r^2 = 0.3977, *P* = 0.02791). Together, this does not support our hypothesis that *Sdic* regulates *sw* at the post-translational level in testes through competitive exclusion, instead it appears that *Sdic* could regulate *sw* through constraints imposed by protein complex assembly stoichiometry. Proteomic data is necessary but currently lacking to unambiguously test these hypotheses at the effective functional level.

Divergent expression of *Sdic* paralogs is present across strains

Using the same sequencing outputs and leveraging on the precise knowledge of the nucleotide differences among *Sdic* paralogs (supplementary table S3.7), we estimated the expression level of each paralog and determined whether there were statistically significant differences among them for each strain. We detected minimal or no *Sdic* expression in male heads and ovaries, which precluded our ability to study *Sdic* paralog functional divergence outside of the testis. In testis, we report differences in expression among the paralogs in all four strains (fig. 3.4, supplementary table S3.2). It should be noted that the low expression detected for paralogs *SdicIII_B* in ISO-1 and *SdicII* in A4 could be artefacts of their motif sequences overlapping the ends of their 3'UTRs, which could have been already degraded in some transcripts by the exosome (Tourrière et al 2002).

In ISO-1, *SdicII_C* (paratype *l*) has significantly higher expression than all other paralogs, while *SdicIV_3* (paratype *j*) is significantly lower than both *SdicI_2* (paratype *k*) and *SdicIV_4* (paratype *c*). Also in B6, *SdicI* (paratype *l*) has significantly higher expression than the other two paralogs. In A4, *SdicII* (paratype *c*) is expressed at a significantly lower level than the other paralogs, with *SdicIII* (paratype *h*) also showing significantly higher expression than *SdicIV* (paratype *c*). Lastly, in A7, the paralog with a premature stop codon induced by a 17.5kb TE insertion in the fourth exon, *SdicIII* (paratype *a*), has significantly higher expression than all other paralogs except for *SdicII* (paratype *l*). Further, the fixed paralog across strains, *Sdic1*-like (_1 paralogs in fig. 3.4), is not the most highly expressed paralog in any of the strains. Positionally within the *Sdic* cluster, with the exception of B6 (fig. 3.4B), none of the outermost *Sdic* paralogs exhibit the highest expression level.

No clear pattern between paralogs and promoter type (fig. 3.1A) is apparent at this time, which precludes the ability to make informed inferences about the influence of the promoter on *Sdic* paralog expression. As a result, we opted for comparing the normalized expression of all the *Sdic* paralogs, excluding ISO-1_*SdicB* and A4_*SdicII*, paralogs for which motif overlap with the end of the 3'UTR made quantification unreliable (supplementary table S3.7). Overall, there is no trend in expression for the different paralogs, as grouped by the protein version coded, *i.e.*, by paratype, as in Clifton et al 2020 (fig. 3.5). Notably, the *Sdic* paralog with the 17.5 kb TE insertion and a premature stop codon, A7_*SdicIII* (paratype *a*) shows higher expression than all other paralogs analyzed, except for B6_*SdicI* and A4_*SdicIII*. Second, while ISO-1 shows the highest aggregate expression across panel I (fig. 3.3A), the individual *Sdic* paralogs in ISO-1 show the lowest individual expression levels across the entire panel (fig. 3.5, supplementary table S3.3).

The *Y* chromosome regulates aggregate *Sdic* expression level differentially across tissues

The *Y* chromosome has been shown to act as a trans-acting factor with the capability to regulate the expression of 20-40% of the genes in testes, and even somatic tissues, of *D. melanogaster* (Lemos et al 2008; Jiang et al 2010). Here, we evaluated the effect of different *Y* chromosomes on *Sdic* expression in a common genetic background, specifically that of the strain 4361. For that, we first generated six *Y* chromosome substitution lines following a previously established mating scheme (Panel II in table 1; Materials and Methods; Lemos et al 2008). Next, we performed two types of expression analyses. A reliable assembly of the *Sdic* region in the 4361 strain does not currently exist. Therefore, we limited our analysis to regions conserved across the *Sdic* paralogs reliably annotated (Clifton et al 2020), *i.e.*, exon 1 for detection of aggregate *Sdic* expression, and the last exon of the fixed paralog *Sdic1*-like, which is significantly remodeled compared to the other *Sdic* paralogs and *sw*.

In the first analysis, we assayed aggregate *Sdic* and *Sdic1*-like expression in male whole-bodies using qRT-PCR. We found statistically significant differences in total *Sdic* expression within the panel, with A7y showing greater expression than ORRy (fig. 3.6A, supplementary table S3.4). No difference in expression was found for *Sdic1*-like alone. This result suggests that *Y* chromosome has a regulatory impact on the expression of the *Sdic* multigene family contributing to interpopulation expression differences, although this effect does not necessarily affect each *Sdic* paralog.

To increase our ability to detect the regulatory effect of the *Y* chromosome, we performed additional PE-100 RNA-sequencing at the tissue level. We surveyed expression of different *Sdic* paralogs in testes and male heads across four strains (4361, yA4, yA7, yB6), and in accessory glands of two strains (4361, yA7). Using the above-mentioned computational pipeline, we quantified RNA-seq reads containing perfect matches to sequences along the regions targeted in our qRT-PCR assay for *Sdic*All and *sw*. We found that neither aggregate *Sdic* nor *sw* expression is differentially expressed in testis across the *Y* chromosome substitution panel (fig. 3.7A, supplementary table S3.5). In agreement with panel I, we find minimal or no *Sdic* expression in male heads, although the *Y* chromosome does show an inconsistent effect on *sw* expression in this anatomical part. The strain A4y shows a significantly increased average expression compared to 4361while A7y shows significantly decreased expression compared to 4361 (fig. 3.7B, supplementary table S3.5). Interestingly, we found that the *Y* chromosome has an impact on the total expression of *Sdic*, but not *sw*, in male accessory glands, with A7y showing decreased expression compared to 4361 (fig. 3.7C, supplementary table S3.5).

DISCUSSION

In contrast to RNA-based duplicates, which often recruit novel *cis*-regulatory sequences relative to those present in the original gene, full DNA-based duplicates have a lower probability to evolve new functional attributes (Chen et al 2013; Assis & Bachtrog 2013). The evolutionary dynamics of most DNA-based duplicates has been studied in interspecific analyses focused on pairs of tandemly arrayed paralogs (Cardoso-Moreira et al 2016; Loehlin & Carroll 2016; Rogers et al 2017; Loehlin et al 2021; Zhang et al 2022;) or on those produced through whole genome duplication or aneuploidy (Song et al 2020; Desvignes et al 2021; Gillard et al 2021; Shi et al 2021). Rarely has the functional evolution of recently expanded, tandemly arranged gene families composed of more than two paralogs been studied at the population level, while also having precise a sequence annotation of the individual paralogs (Clifton et al 2017). Here, we have performed a population, tissue-level characterization of the expression levels of individual paralogs in the tandemly expanded gene family *Sdic*, which is unique to *D. melanogaster*, and has been precisely annotated across a set of reference quality assemblies. We have done so by examining the effects of *cis*- and *trans*- acting regulatory variation across a set of strains of geographically diverse origins, paying special attention to the *trans*-acting impact of *Y* chromosome variation, *i.e.*, YRV, in a controlled genomic environment.

We previously reported no correlation between total *Sdic* expression and *Sdic* copy number in male whole-bodies across a panel of geographically diverse isogenic lines, which harbor the most common *Sdic* copy numbers for this multigene family in natural populations (Clifton et al 2020). This lack of correlation was interpreted as the result of variation in expression modifiers acting in *cis* and *trans*. Here, we more precisely assayed *Sdic* expression in different tissues. The positive correlation between total *Sdic* expression and CN documented (fig. 3.2) suggests that *Sdic* dosage could be under positive selection in testis while also highlighting the importance of tissue level analysis of gene expression in order to reveal biologically meaningful patterns. Further, another analysis of duplicate expression, also conducted with whole-body samples, found no differences in total expression between strains harboring two or three tandem repeats of an enzyme coding gene (Loehlin et al 2021). Here we show total *Sdic* expression significantly increases from three to four to five, but not from five to six, copies (fig. 3.3); although, in both these cases, CNV takes place in different genetic backgrounds, so it is not possible to disentangle the effects of CNV and genetic background. Dosage constraints based on stoichiometry are likely different in different tissues and dependent on whether the protein assembles into a complex or not. Protein-protein interaction information is required to better identify any potential stabilizing mechanism on gene expression beyond a certain threshold or pertaining to protein complex assembly.

Testis-specific expression is typical of newly evolved genes, often thought to be the result of particularities of this tissue, namely a particularly permissive chromatin and the simplicity of promoter sequences required for expression (Kaessmann 2010; Guschanski et al 2017; Witt et al 2021). The *de novo* acquisition of a testis-specific promoter element likely explains the greater expression of *Sdic* in testis relative to *sw* seen in all strains, which is the case when considering the aggregate expression of *Sdic* and when considering the expression of each paralog relative to *sw*. We find that both aggregate *Sdic* and *sw* expression vary significantly across our *Sdic* CNV panel of strains, showing a statistically significant negative correlation between expression levels**.** This conflicts with the hypothesis that higher levels of *sw* expression should require more *Sdic* expression to regulate any dosage dependent function of *sw* in testis. Protein complex subunits are particularly sensitive to dosage constraints (Zhang et al 2022). A negative correlation could be due to an unidentified mechanism that represses *sw* while over-expressing *Sdic*, which has been postulated as an important mechanism for maintaining dosage balance in biological systems (Tu et al 2016). This could be a mechanism that maintains total dynein intermediate chain dosage within a limit that does not significantly impede cytoplasmic dynein protein complex assembly and functionality. Lastly, and contrary to our previous survey of RNA-seq expression data (Clifton et al. 2017), we did not detect *Sdic* expression in male heads or ovaries. This could be due to our strict read counting approach not detecting low expression, or mistakes in our previous RNA-seq read counting methodology (Clifton et al 2017). The absence of *Sdic* expression data outside testis for our *Sdic* CNV panel precluded our ability to study expression changes associated with functional divergence through tissue-specificity, as well as to measure any influence *Sdic* CN could have on total *Sdic* expression outside testis.

Consistent with the pattern of functional divergence across tissues previously documented at the expression level (Clifton et al 2017), the *Sdic* paralogs vary significantly in their testis expression level in all strains analyzed, showing no consistent trend for promoter type or coded protein variant (fig. 3.4, fig. 3.5). This is still consistent with *cis* regulatory differences being present among the paralogs of each strain, although the identity of those differences is not apparent at this time. Precise identification of the *cis* regulatory impact of individual *Sdic* paralogs on one another will require quantification of *Sdic* expression in genetically engineered lines with different individual copies removed from the array. Further, it is not known if these tissue-level expression differences among paralogs are maintained at the effective functional level of the protein or are stabilized through post-translational buffering mechanisms. Precise quantification of the individual paralogs at the proteomic level is needed to test for the presence of these buffering mechanisms.

Having accurately annotated sequence information of the *Sdic* region has allowed us to test relevant aspects associated with the expression properties of a recently formed, tandemly arranged multigene family. In testes, we found evidence of expression for all *Sdic* paralogs in each of the strains, including those that show the presence of a premature stop codon or unusual features such as a TE insertion (A7_*SdicIII*). While we have no way to compare promoter types here while controlling for other forms of *cis* variation, in A7 (fig. 3.4D) all *Sdic* paralogs are driven by the same promoter and show similar expression levels with the exception of the paralog harboring a ~17.5 kb mdg1 TE insertion, the most highly expressed *Sdic* paralog quantified. By a conservative estimate, rare allele of large effect (RALE) transposable element insertions in or near transcripts in *D. melanogaster* were associated with reductions in gene expression contrary to the case here (Cridland et al 2015). TE insertions can modify expression through inducing loss of accessibility of the transcriptional machinery to *cis*-regulatory DNA elements or by influencing expression at the post-transcriptional level through interactions at the 3'UTR (Goubert et al 2020). Increased expression of this paralog suggests a repressive cis-regulatory element may be present within the *Sdic* array. The lack of consistent differences between *Sdic* paralogs with different promoters in the same strain, suggests that the *Sdic* promoters likely play a minimal role as *cis* regulators of *Sdic* expression, at least in the testis. It is not known if the sequence differences among the promoters affect binding of *Sdic*-modulating transcription factors, such as *modulo* (Mikhaylova et al 2006), however transcription factors play a decreased role in modulating gene expression in the permissive chromatin environment of the testis (Witt et al 2021). Controlled experiments using genetically modified single copy *Sdic* regions in identical genomic backgrounds but driven by different *Sdic* promoters will be needed to properly evaluate how promoter evolution contributes to the functional divergence of *Sdic* paralogs. Lastly, we find no evidence of positional effects, *i.e.*, the position of the repeat within the tandem array, acting on expression of the *Sdic* paralogs. In none of the strains, *Sdic1*-like paralog, the only fixed and mostly functionally diverged paralog from *sw*, shows the highest or lowest expression. The action of positional effects on young tandem gene families should be more precisely identified using genetically engineered lines with shuffled orders of paralogs within the same genomic background.

The *Y* chromosome impacts male fitness and fertility (Carvalho et al 2001). The presence of *Sdic* impacts sperm competition (Yeh et al 2012) and *Sdic* expression is maximal in testis and high in male accessory glands (Clifton et al 2017) –consistent with the *out of the testis hypothesis* for the origin of new genes (Kaessmann 2010)–. Based on these premises, we hypothesized that the *Y* chromosome could act as a *trans* regulator of *Sdic* expression. No analysis of the tissue level impact of YRV has been conducted using an analytical pipeline effective enough in distinguishing unambiguously *Sdic* from *sw* expression. Our panel of *Y* chromosome substitution lines did not identify any *trans* regulatory action of YRV on *Sdic* expression in testis or male heads but demonstrated that YRV impacts total *Sdic* expression in accessory glands, a somatic tissue with a role in reproduction, and also *sw* expression in heads but not in testis or accessory glands. These results are in line with those documented with similar experiments involving *Y* chromosome substitutions lines in which the authors detected significant effects on the expression variation of this chromosome on somatic tissues or on genes primarily expressed in somatic tissues (Lemos et al 2008; Branco et al 2017; Wang et al 2018; Ågren et al 2020).

Our work highlights the importance of combining molecular and sequencing approaches to obtain paralog-specific information for generating a more nuanced portrait of how recently originated multigene families functionally evolve along their path to fixation and consolidation in the genome. Nevertheless, having a full understanding of how specific genetic changes in promoters, 3’UTRs, and other *cis*-regulatory motifs impact the expression of different paralogs will require the generation of synthetic genotypes, ultimately harboring paralogs of different ages, to uncover how regulatory mechanisms that diversify gene functions evolve as gene duplicates age.

MATERIALS AND METHODS

Fly husbandry and strains used

We used a combination of *D. melanogaster* strains, including the reference strain and others with wild-type genotypes of diverse geographical origin and variable *Sdic* copy number (panel I) (King et al. 2012), as well as a set of *Y* chromosome substitution lines (panel II) (table 1). Flies were reared on dextrose-cornmeal-yeast medium in narrow polystyrene *Drosophila* vials at room temperature (~25^{o}C) under 24 h fluorescent light. Adult virgins were collected within 6-8 h of eclosion, sorted by sex, and then cultured separately by sex in groups of $\leq$50 individuals until sacrificed. All sorting, scoring, collecting, counting, and manipulation of flies was performed under CO_2 anesthesia.

Generation of *Y* chromosome substitution lines

Crosses to generate the *Y* chromosome substitution lines (panel II; table 1) were done following the mating scheme as described (Lemos et al 2008). *Y* chromosomes were chosen such that a wide variation in origin was surveyed. All strains share the same genomic background, Bloomington Stock Center strain #4361, with the exception of *Y* chromosomes derived from different donor strains. To buffer the effects of newly generated dominant mutations during the construction of the synthetic genotypes, each line was generated from multiple G0 males (65-75 depending on the strain), from which 70 G1 males per strain were used for the subsequent cross. Lastly, 30 G2 males per strain were pooled to generate the final *Y* chromosome substituted strains. All crosses were done with an equal number of males and 4361 females.

Material collection and tissue dissections

Material collections from panels I and II took place at separate time periods. For panel I, at 5 d post-eclosion, virgin adults were systematically sacrificed and had their tissues (testes, ovaries, and male heads) dissected. For panel II, adult naïve males were collected at 5 d post-eclosion and tissues (testes, male accessory glands, and male heads) were dissected from naïve adults aged to 4-6 d post-eclosion. Tissue dissection was performed under a stereoscope in 1×PBS (phosphate-buffered saline) solution, while dissected tissues were stored in ice-cold 1×PBS. Following dissection, 1×PBS was replaced with TRIzol reagent (ThermoFisher) using a micropipettor. Tissues were then completely homogenized using a 1.5 mL motorized pestle, flash-frozen in liquid nitrogen, and immediately transferred to a -80 C freezer for storage until used for RNA extractions. Heads were collected similarly, except prior to storing in ice-cold 1×PBS, flies were CO_2 anesthetized and collected into a 15 mL centrifuge tube (Corning). The centrifuge tube was then flash-frozen in liquid nitrogen and vigorously agitated by vortexing and strong flicking and shaking to decapitate and dismember the flies. The contents of the centrifuge tubes were poured onto a petri dish over ice from which the heads were quickly collected under a stereoscope and placed in ice-cold 1×PBS before being transferred to TRIzol. No tissue was kept in 1×PBS for longer than 2 h. Dissections were done separately for each strain, tissue, and sex to avoid possible cross-contamination. Tissues were dissected within specific timeframes to minimize unintended variation, which is particularly relevant in the case of heads due to circadian rhythms. In this case, within1 h window from 3:00-4:00 pm (panel I) or 1:30-2:30 pm (panel II).

RNA extractions

Immediately prior to RNA extraction, tissues previously homogenized in TRIzol were pooled to make the desired number of individuals per replicate: 25 male whole bodies, 20 pairs of ovaries, ~200 heads, 100 pairs of accessory glands, and 60 and 100 pairs of testes for panels I and II, respectively. Four replicates were extracted for each sample type. Total RNA was extracted using chloroform following manufacturer instructions for TRIzol (ThermoFisher). DNA traces were subsequently eliminated using the RNeasy mini kit with DNase I (Qiagen). RNA integrity, purity, and concentration were assessed by gel electrophoresis (Aranda et al 2012), a Nanodrop-8000 spectrophotometer (ThermoFisher), and a Qubit RNA BR assay kit (ThermoFisher), respectively. Following extraction, RNA was immediately stored at -80ºC until used for cDNA synthesis or submitted for RNA sequencing.

qRT-PCR analysis of Y substitution panel

For each whole-body sample from panel II, 1.0 µg of total RNA from four biological replicates was converted to 20 µl of cDNA using the SuperScript IV First-Strand Synthesis System with an RNase inhibitor (ThermoFisher). Successful reverse transcriptase reactions were confirmed through successful RT–PCR of the housekeeping gene *Gapdh2* using 2X Apex Taq RED Master Mix (Apex Bioresearch Products) from 1 µl of cDNA (supplementary table S3.6). 20 µl qRT-PCR reactions were performed in all experiments using 1 µl of 1:10 diluted cDNA in 200 µl 96-well plates (Bio-Rad) with using a CFX-96 1000 touch real-time instrument (Bio-Rad) and PowerUP SYBR Green Master Mix (Applied Biosystems) over 40 cycles. Four biological replicates were quantified per strain. Primer efficiencies were determined using a 1:5 dilution standard curve (5X,1X,1:5X,1:25X,1:125X). Expression estimates were obtained accounting for variable primer efficiencies for the genes of interest (*Sdic*All, *Sdic1*-like) and the reference gene *clot* (*cl*) (Pfaffl 2001). 4361 samples were used as the calibrator for all comparisons. Primer sets used are listed in supplementary table S3.6. Primer design for *Sdic* took into consideration sequence similarities and differences with *sw* and *AnxB10* to confidently survey solely *Sdic* expression. To estimate the combined expression of all *Sdic* paralogs, the *Sdic*All primers target a region with perfect sequence conservation across all paralogs and strains to prevent any paralog or population bias. Likewise, the priming sites for *Sdic1*-like target a region that is conserved across all *Sdic1*-like paralogs reliably annotated. All samples tested for the same primer set were run on the same plate. Material from 4361 male whole bodies was used as to generate the standard curves for calculating primer efficiencies. delta-Cq values were used in the statistical analyses.

RNA sequencing

Samples from panel I were sequenced separately from panel II samples. Prior to sequencing, RNA integrity was further estimated using the RNA 6000 Nano Chip Kit (Agilent Technologies) with an Agilent 2100 Bioanalyzer. For each sample, the three out of the four replicates with the highest RIN values were submitted for RNA sequencing at the UCI Genomics High Throughput Facility (GHTF). Ribodepleted, strand-specific paired-end libraries were prepared according to the Illumina TruSeq Total RNA stranded protocol. The input quantity for total RNA was 500 ng and rRNA was depleted using ribo-zero rRNA gold removal kit (human/mouse/rat). The rRNA depleted RNA was chemically fragmented for three minutes. First strand synthesis used random primers and reverse transcriptase to make cDNA. After second strand synthesis the ds cDNA was cleaned using AMPure XP beads and the cDNA was end repaired and then the 3’ ends were adenylated. Illumina barcoded adapters were ligated on the ends and the adapter ligated fragments were enriched by nine cycles of PCR. The resulting libraries were validated by qPCR and sized by Agilent Bioanalyzer DNA high sensitivity chip. The concentrations for the libraries were normalized and then multiplexed together. The multiplexed libraries were sequenced on paired-end 100 cycles chemistry on a NovaSeq 6000 instrument.

RNA-seq read processing and quantification of gene expression

Quality control and preprocessing of RNA-seq reads were performed using HTStream (https://github.com/s4hts/HTStream; last accessed February 14, 2022), including removal of known *D. melanogaster* rRNA-related sequences as presented in NCBI, PCR duplicates, adapter sequences, reads shorter than 50 nt, and filtered for low-quality bases using a sliding window approach that required a window size of 10 nt and a minimum quality score of 20. Gene expression was examined for the entire *Sdic* multigene gene family, the fixed paralog (*Sdic1*-like), and *Sdic*'s parental gene *sw* across all five strains, as well as 14 individual *Sdic* paralogs from panel I (supplementary table S3.7). Expression levels were estimated as the number of RNA-seq reads per sample with perfect matches to gene- or paralog-specific motifs. These motifs contain core motifs ranging from 10-18 nt that identify a particular paralog, sets of paralogs, or different genes depending on the case. These core motifs are ultimately extended to 130 nt. Counts were generated using a custom script that examines all the reads in each library for the presence of a given motif (Clifton et al 2020). Normalized counts are expressed as RPKM (Mortazavi et al 2008), *i.e.*, the number of perfect matches detected divided by the total number of RNA-seq reads generated in each library. We used ≥11 reads in all three replicates as a threshold for dubbing a gene as expressed.

Statistical analyses

One-way or two-way ANOVAs were implemented to identify differences in gene expression across strains for a given gene and between genes depending on the test. For tests with statistically significant *P*-values, post-hoc Tukey-Kramer HSD tests were performed to identify statistically significant pairwise comparisons. Gene expression correlations were calculated using the Pearson's product-moment correlation test. All the analyses were performed in R.

ACKNOWLEDGEMENTS

We are in debt to J.J. Emerson and Brandon Gaut for feedback on early versions of this manuscript and Seung-Ah Chung and the UCI Genomics High Throughput Facility (GHTF) for sequencing our samples. We thank Xiao-Qin Xia and Wanting Zhang for technical help with counting RNA-seq reads. We thank Sarah Bedford, Brian Luu, Matthew Raft, and Andres Vargas for help with tissue collections for pilot experiments. We also thank Ali Mortazavi for granting us access to his qRT–PCR instrument and to the University of California, Irvine High-Performance Computing cluster for facilitating our analyses. This work was supported by University of California, Irvine intramural funds and National Science Foundation Grant (MCB-2129845) (to J.M.R), and U.S. Department of Education GAANN fellowship funds (to B.D.C).

REFERENCES

Ågren JA, Munasinghe M, and Clark AG. 2020. Mitochondrial-*Y* chromosome epistasis in *Drosophila melanogaster*. *Proc. R. Soc. B*. 28720200469

Aranda PS, LaJoie DM, Jorcyk CL. 2012. Bleach Gel: A simple agarose gel for analyzing RNA quality. *Electrophoresis*. 33(2):366-369

Assis R, Bachtrog D. 2013. Neofunctionalization of young duplicate genes in *Drosophila*. *Proc Natl Acad Sci U S A*. 110(43):17409-14.

Branco AT, Schilling L, Silkaitis K, Dowling DK, Lemos B. 2017. Reproductive activity triggers accelerated male mortality and decreases lifespan: genetic and gene expression determinants in *Drosophila*. *Heredity*

Boylan K, Serr M, Hays T. 2000. A molecular genetic analysis of the interaction between the cytoplasmic dynein intermediate chain and the glued (dynactin) complex. *Mol Biol Cell*. 11(11):3791–3803.

Brown CJ, Todd KM, Rosenzweig RF. 1998. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Mol Biol Evol*. 15(8):931–942.

Cardoso-Moreira M, Arguello JR, Gottipati S, Harshman LG, Grenier JK, Clark AG. 2016. Evidence for the fixation of gene duplications by positive selection in *Drosophila*. *Genome Res*. 26(6):787–798.

Carvalho AB, Dobo BA, Vibranovski MD, Clark AG. 2001. Identification of five new genes on the *Y* chromosome of *Drosophila melanogaster*. *PNAS*

Catalan A, Glaser-Schmitt A, Argyridou E, Duchen P, Parsch J. 2016. An indel polymorphism in the *MtnA* 3’ untranslated region is associated with gene expression variation and local adaptation in *Drosophila melanogaster*. *PLoS Genet.* 12(4):e1005987.

Chakraborty M, VanKuren NW, Zhao R, Zhang X, Kalsow S, Emerson JJ. 2018. Hidden genetic variation shapes the structure of functional elements in *Drosophila*. *Nature Genetics*. 50:20-25

Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. 2019. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat Commun*. 10(1):4872.

Chang C, Larracuente A. 2019. Heterochromatin-Enriched Assemblies Reveal the Sequence and Organization of the *Drosophila melanogaster Y* Chromosome. *Genetics*. 211:333-348

Chen S, Krinsky B, Long M. 2013. New genes as drivers of phenotypic evolution. *Nat Rev Genet*. 14:645–660

Civetta A, Ranz JM. 2019. Genetic Factors Influencing Sperm Competition. *Front Genet*. 13;10:820.

Clifton BD, Librado P, Yeh SD, Solares ES, Real DA, Jayasekera SU, Zhang W, Shi M, Park RV, Magie RD, Ma H, Xia X, Marco A, Rozas J, Ranz JM. 2017. Rapid functional and sequence differentiation of a tandemly repeated species-specific multigene family in *Drosophila*. *Mol Biol Evol*. 34(1):51–65

Clifton BD, Jimenez J, Kimura A, Chahine Z, Librado P, Sanchez-Gracia A, Abbassi M, Carranza F, Chan C, Marchetti M, Zhang W, Shi M, Vu C, Yeh S, Fanti L, Xia X, Rozas J, Ranz JM. 2020. Understanding the early evolutionary stages of a tandem *D. melanogaster*-specific gene family: a structural and functional population study. *Mol Biol Evol*. 37(9):2584–2600

Cridland JM, Thornton KR, Long AD. 2015. Gene Expression Variation in *Drosophila melanogaster* Due to Rare Transposable Element Insertion Alleles of Large Effect. *Genetics*.199(1):85–93

Cridland JM, Majane AC, Sheehy HK, Begun DJ. 2020. Polymorphism and Divergence of Novel Gene Expression Patterns in *Drosophila melanogaster*. *Genetics*

Desvignes T, Sydes J, Montfort J, Bobe J, Postlethwait JH. 2021. Evolution after Whole-Genome Duplication: Teleost MicroRNAs. *Mol Biol Evol*. 38(8):3308-3331

Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, van den Bout A, Bishara A, Rosenkrantz JI, Lorig-Roach R, Field AR, Maeussler M, Russo L, Bhaduri A, Nowakowski TJ, Pollen AA, Dougherty ML, Nuttle X, Addor MC, Zwolinski S, Katzman S, Kriegstein A, Eichler EE, Salama SR, Jacobs FMJ, Haussler D. 2018. Human-specific *NOTCH2NL* genes affect Notch signaling and cortical neurogenesis. *Cell*. 173(6): 1356–1369.e22.

Florio M, Albert M, Taverna E, Namba T, Brandl H, Lewitus E, Haffner C, Sykes A, Wong FK, Peters J, Guhr E, Klemroth S, Prüfer K, Kelso J, Naumaan R, Nüsslein I, Dahl A, Lachmann R, Pääbo S, Huttner WB. 2015. Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science*. 347(6229):1465-70

Force A, Lynch M, Pickett FB, Amores A, Yan Y, Postlehwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*. 151:1531–154

Gillard GB, Grønvold L, Røsæg LL. Holen MM, Monsen O, Koop BF, Rondeau EB, Gundappa K, Mendoza J, Macqueen DJ, Rohlfs RV, Sandve SR, Hvidsten TR. 2021. Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. *Genome Biol*. (22)103

Goubert C, Zevallos NA, Feschotte C. 2020. Contribution of unfixed transposable element insertions to human regulatory variation. *Phil. Trans. R. Soc*. B375:20190331

Guschanski K, Warnefors M, Kaessmann H. 2017. The evolution of duplicate gene expression in mammalian organs. *Genome Research*. 27:1461-1474

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet*. 10(8):551–564.

Hayward CPM, Liang M, Tasneem S, Soomro A, Waye JS, Paterson AD, Rivard GE, Wilson MD. 2017. The duplication mutation of Quebec platelet disorder dysregulates PLAU, but not C10orf55, selectively increasing production of normal PLAU transcripts by megakaryocytes but not granulocytes. *PLoS One*. 2017;12(3):e0173991.

Hug N, Longman D, Caceres JF. 2016. Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res*. 44(4):1483–1495.

Jones P, Chang H, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetiew M, Yong S, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. PMID: 24451626

Jayaswal V, Jimenez J, Magie R, Nguyen K, Clifton B, Yeh S, Ranz JM. 2018. A species-specific multigene family mediates differential sperm displacement in *Drosophila melanogaster*. *Evolution*. 72(2):399–403.

Jiang PP, Hartl DL, Lemos B. 2010. Y not a dead end: epistatic interactions between Y-linked regulatory polymorphisms and genetic background affect global gene expression in *Drosophila melanogaster*. *Genetics* 186 (1), 109-118

Jugulam M, Niehues K, Godar AS, Koo DH, Danilova T, Friebe B, Sehgal S, Varanasi VK, Wiersma A, Westra P, et al. 2014. Tandem amplification of a chromosomal segment harboring 5-enolpyruvylshikimate-3-phosphate synthase locus confers glyphosate resistance in *Kochia scoparia*. *Plant Physiol*. 166(3):1200–1207.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*. 11:97-108

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res*. 20(10):1313–1326

Kardon JR, Vale RD. 2009. Regulators of the cytoplasmic dynein intermediate chain. *Nat Rev Mol Cell Biol*. 10:854–865

King EG, Macdonald SJ, Long AD. 2012. Properties and power of the *Drosophila* Synthetic Population Resource for the routine dissection of complex traits. *Genetics.* 191(3):935–949.

Konrad A, Flibotte S, Taylor J, Waterston RH, Moerman DG, Bergthorsson U, Katju V. 2018. Mutational and transcriptional landscape of spontaneous gene duplications and deletions in Caenorhabditis elegans. *Proc Natl Acad Sci USA*. 115(28): 7386–7391.

Kondrashov FA. 2010. Gene Dosage and Duplication. *Evolution After Gene Duplication*. *Wiley-Blackwell*

Kuzmin E, Vandersluis B, Ba ANN, Wang W, Koch EN, Usaj M, Khmelinskii A, Usaj MM, Van Leeuwen J, Kraus O, Tresenrider A, Pryszlak M, Hu M, Varriano B, Costanzo M, Knop M, Moses A, Myers CL, Andrews BJ, Boones C. 2020. Exploring whole-genome duplicate gene retention with complex genetic interaction analysis. *Science*. 368(6498):eaaz5667.

Kuzmin E, Taylor JS, Boone C. 2022. Retention of duplication genes in evolution. *Trends in Genetics*. 38(1):59-72

Lan X, Pritchard JK. 2016. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science*. 352(6288):1009–1013.

Lemos B, Araripe LO, Hartl DL. 2008. Polymorphic *Y* Chromosomes Harbor Cryptic Variation with Manifold Functional Consequences. *Science*

Lemos B., Branco A. T., Hartl D. L., 2010. Epigenetic effects of polymorphic *Y* chromosomes modulate chromatin components, immune response, and sexual conflict. *Proc. Natl. Acad. Sci. USA* 107: 15826–15831

Loehlin DW, Carroll SB. 2016. Expression of tandem gene duplicates is often greater than twofold. *Proc Natl Acad Sci USA.* 113(21):5988–5992.

Loehlin DW, Kim JY, Paster CO. 2021. A tandem duplication in *Drosophila melanogaster* shows enhanced expression beyond the gene copy number. *Genetics*.

Lyckegaard EM, Clark AG. 1989. Ribosomal DNA and *Stellate* gene copy number variation on the *Y* chromosome of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* 86: 1944–1948

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 5:621–628.

Mikhaylova LM, Boutanaev AM, Nurminsky DI. 2006. Transcriptional regulation by *Modulo* integrates meiosis and spermatid differentiation in male germ line. *PNAS*

Newcomb RD, Gleeson DM, Yong CG, Russell RJ, Oakeshott JG. 2005. Multiple mutations and gene duplications conferring organophosphorus insecticide resistance have been selected at the *Rop-1* locus of the sheep blowfly, *Lucilia cuprina*. *J Mol Evol*. 60:207–220.

Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL. 1998. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature*. 396(6711):572–575.

Obbard DJ, Maclennan J, Kim KW, Rambaut A, O’Grady PM, Jiggins FM. 2012. Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. *Mol Biol Evol*. 29(11):3459–3473.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA,Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet*. 39:1256–1260.

Pfaffl MW. 2001. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res*. 29:e45.

Rody HVS, Baute GJ, Rieseberg LH, Oliveira LO. 2017. Both mechanism and age of duplications contribute to biased gene retention patterns in plants. *BMC Genomics*. 18(46)

Rogers RL, Shao L, Thornton KR. 2017. Tandem duplications lead to novel expression patterns through exon shuffling in *Drosophila yakuba*. *PLoS Genet*. 13(5):e1006795.

Sackton TB, Montenegro H, Hartl DL, Lemos B. 2011. Interspecific *Y* chromosome introgressions disrupt testis-specific gene expression and male reproductive phenotypes in *Drosophila*. *PNAS* 108 (41), 17046-17051

Scott EC, Gardner EJ, Masood A, Chuang NT, Vertino PM, Devine SE. 2016. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res*. 26(6):745–755.

Shi X, Yang H, Chen C, Hou J, Hanson KM, Albert PS, Ji T, Cheng J, Birchler JA. 2021. Genomic imbalance determines positive and negative modulation of gene expression in diploid maize. *The Plant Cell*. 33(4):917–939

Song MJ, Potter BI, Doyle JJ, Coate JE. 2020. Gene Balance Predicts Transcriptional Responses Immediately Following Ploidy Change in Arabidopsis thaliana. *The Plant Cell*. 32(5):1434–1448

Teufel AI, Johnson MM, Laurent JM, Kachroo AH, Marcottee EM, Claus OW. 2018. The Many Nuanced Evolutionary Consequences of Duplicated Genes. *Mol. Biol. Evol.* 36(3):304-314

Tourrière H, Chebli K, Tazi J. 2002. mRNA degradation machines in eukaryotic cells. *Biochimie*. 84(8):821-37

Tu X, Wang Y, Zhang M, Wu J. 2016. Using Formal Concept Analysis to Identify Negative Correlations in Gene Expression Data. *IEEE/ACM Trans Comput Biol Bioinform*. 13(2):380-91

Wang M, Branco AT, Lemos B. 2018. The *Y* Chromosome Modulates Splicing and Sex-Biased Intron Retention Rates in *Drosophila*. *Genetics*. 208(3):1057-1067

Wei L, Yuan Y, Hu T, Li S, Cheng T, Lei J, Xie Z, Zhang MQ, Wang X. 2019. Regulation by competition: a hidden layer of gene regulatory network. *Quantitative Biology*. 7(2):110-121

Witt E, Svetec N, Benjamin S, Zhao L. 2021. Transcription Factors Drive Opposite Relationships between Gene Age and Tissue Specificity in Male and Female *Drosophila* Gonads. *Molecular Biology and Evolution*. 38(5):2104–2115

Yan H, Dobbie Z, Gruber SB, Markowitz S, Romans K, Giardiello FM, Kinzler KW, Vogelstein B. 2002. Small changes in expression affect predisposition to tumorigenesis. *Nat Genet* 30:25–26

Yeh SD, Do T, Chan C, Cordova A, Carranza F, Yamamoto EA, Abbassi M, Gandasetiawan KA, Librado P, Damia E, et al. 2012. Functional evidence that a recently evolved *Drosophila* sperm-specific gene boosts sperm competition. *Proc Natl Acad Sci USA*. 109(6):2043–2048.

Zhang D, Leng L, Chen C, Huang J, Zhang Y, Yuan H, Ma C, Chen H, Zhang Y. 2022. Dosage sensitivity and exon shuffling shape the landscape of polymorphic duplicates in *Drosophila* and humans. *Nature Ecology and Evolution*. 6:273-287

Zurovcova M., Eanes W. F., 1999. Lack of nucleotide polymorphism in the Y-linked sperm flagellar dynein gene Dhc-Yh3 of *Drosophila melanogaster* and *D. simulans*. *Genetics* 153: 1709–1715

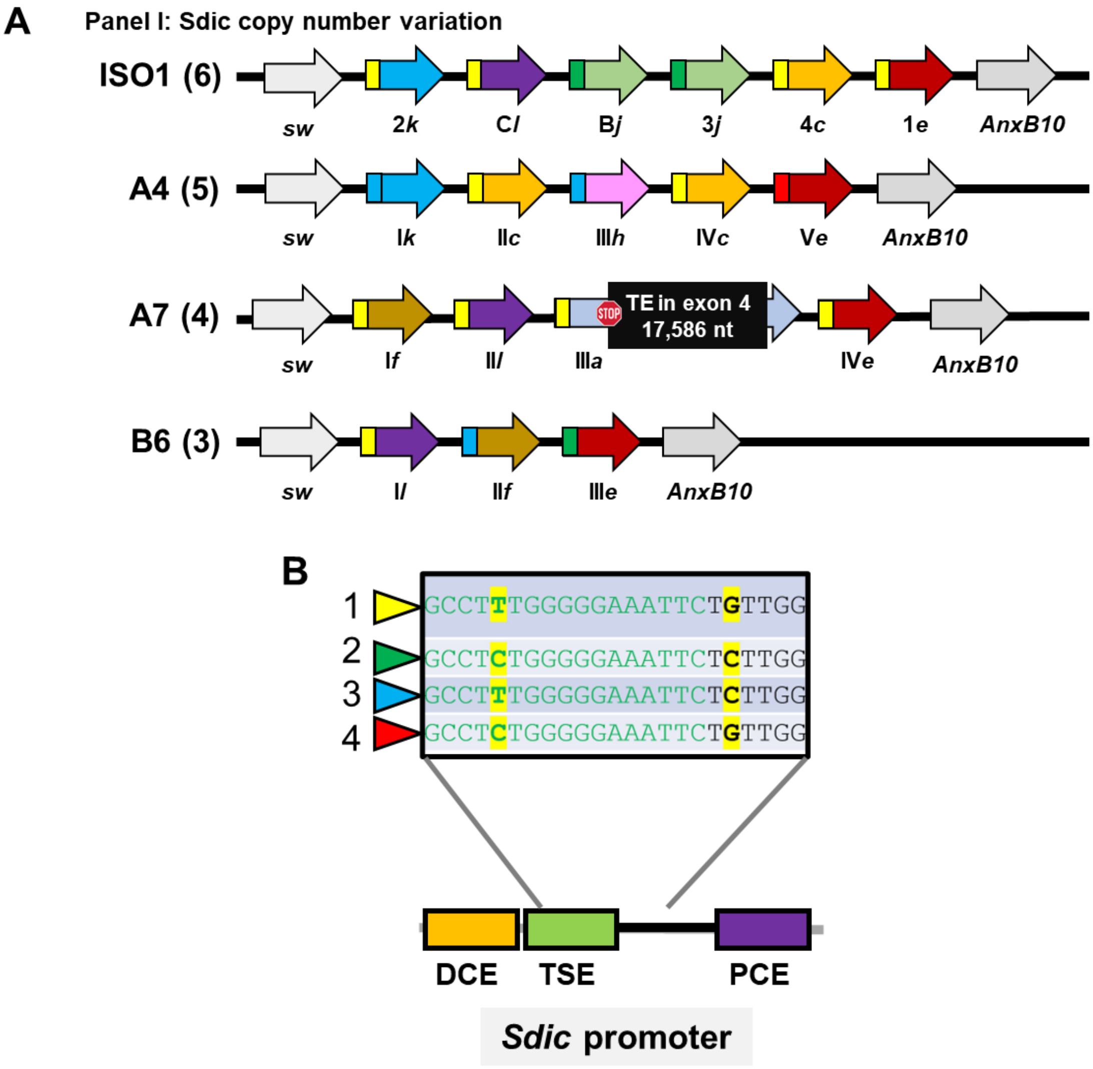


Figure 3.1. The structural and sequence diversity of the *Sdic* region. (A) Structure of the *Sdic* region in panel I. Strain name and number of *Sdic* copies for each region is displayed on the left. The regions are shown as they are arranged from telomere (left) to centromere (right). The different paralogs (arrows) are color coded based on the version of the Sdic protein they code. Paralogs coding the same version of the Sdic protein are said to represent the same paratype. Paratypes are labeled as reported (Clifton et al 2020). Dark red paratype, the fixed paralog referred to as *Sdic1*-like. Promoters (boxes) are color coded according to (B). (B) *Sdic* promoter sequence variation. Two nucleotide sites are variable across the 18 promoters, resulting in four types. Structural organization of the *Sdic* promoter is labelled according to Nurminsky et al. 1998. DCE, distal core element; TSE, testis-specific core element; PCE, proximal core element.

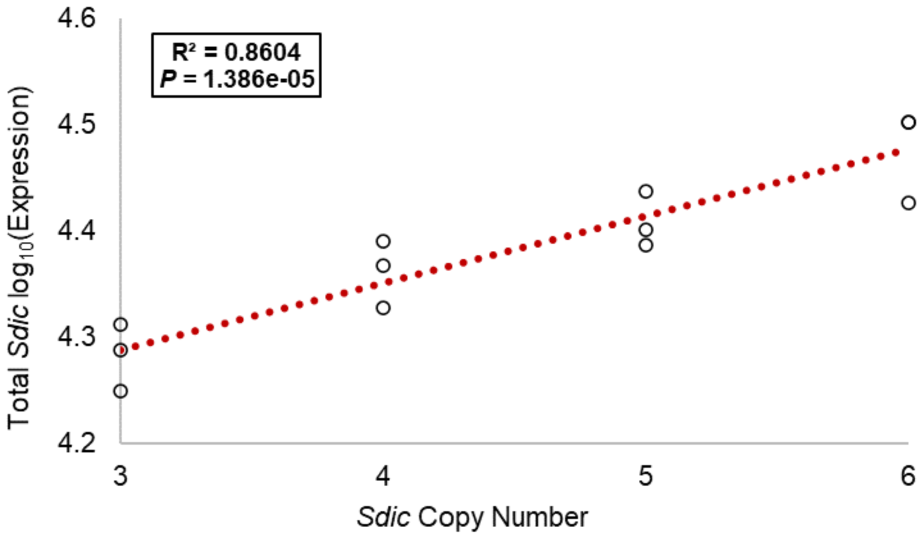


Figure 3.2. Aggregate *Sdic* expression in testis and *Sdic* CN are positively correlated. Gene expression is provided as $\log_{10}$(RPKM). The coefficient of determination (r^2) and its corresponding *P*-value are shown at the top**.** Red dotted line, linear regression line**.**

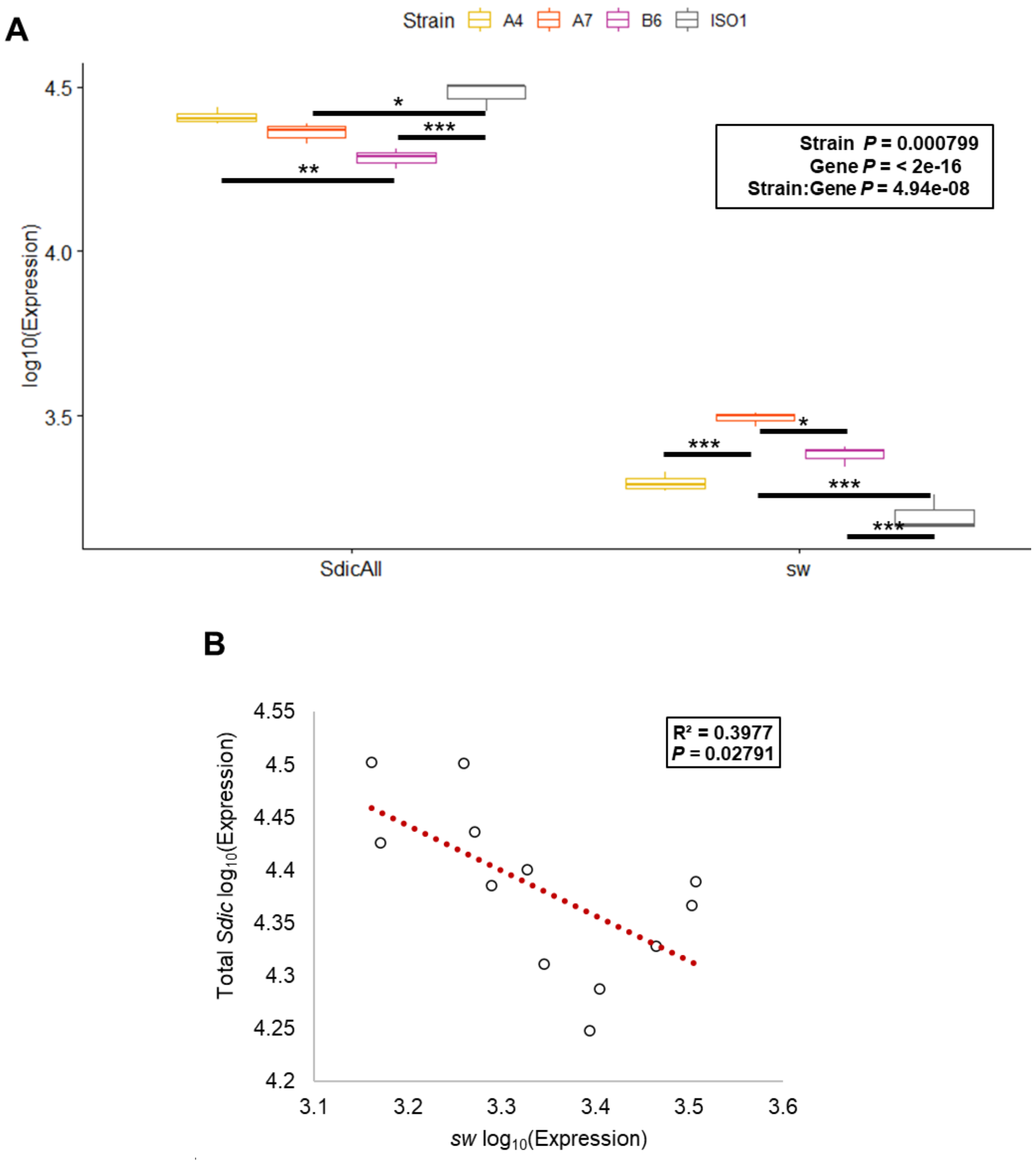


Figure 3.3. Total expression of all *Sdic* paralogs is greater than, but negatively correlated with, *sw* expression, which both show expression variation across testis. (A) Two-way ANOVA for aggregate *Sdic* (*Sdic*All) vs *sw* expression in testis. Box plots represent the median surrounded by quartiles. *P*-values from a two-way ANOVA are displayed at the top right. Black bars connect significant pairwise comparisons (Tukey HSD) (not shown for *Sdic*All and *sw* comparisons in (A)). *, $P < 0.05$; **, $P < 0.005$; ***, $P < 0.0005$. Statistical values for all comparisons are listed in supplementary table S3.1. (B) Correlation between expression of *sw* and aggregate *Sdic*. The coefficient of determination (r^2) and its corresponding *P*-value are shown at the top. Red dotted line, linear regression line. Gene expression is provided as $\log_{10}$(RPKM). All the strains from panel I are shown.

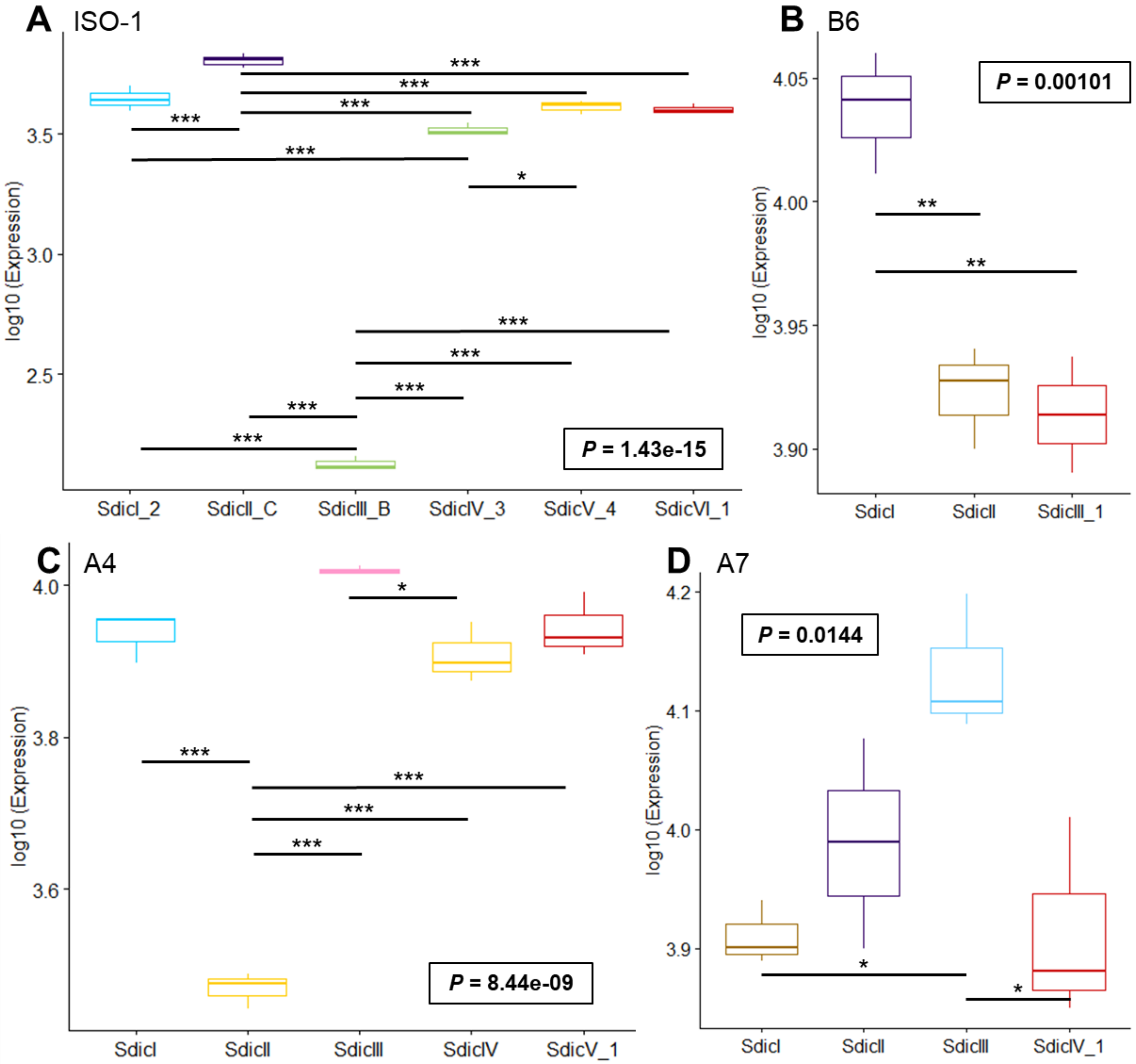


Figure 3.4. Individual *Sdic* paralogs show significant expression differences within strains. Expression of the individual *Sdic* paralogs in ISO-1 (A), B6 (B), A4 (C), and A7 (D) is provided as $\log_{10}$(RPKM). Box plots represent the median surrounded by quartiles. *Sdic* paralogs are shown from left to right as they are arranged along the cluster from telomere to centromere in each strain. Paralogs are colored as in fig. 3.1A. Black boxes contain one-way ANOVA *P*-values. Black bars connect significant pairwise comparisons (Tukey HSD). *, $P < 0.05$; **, $P < 0.005$; ***, $P < 0.0005$. Statistical values for all comparisons are listed in supplementary table S3.2.

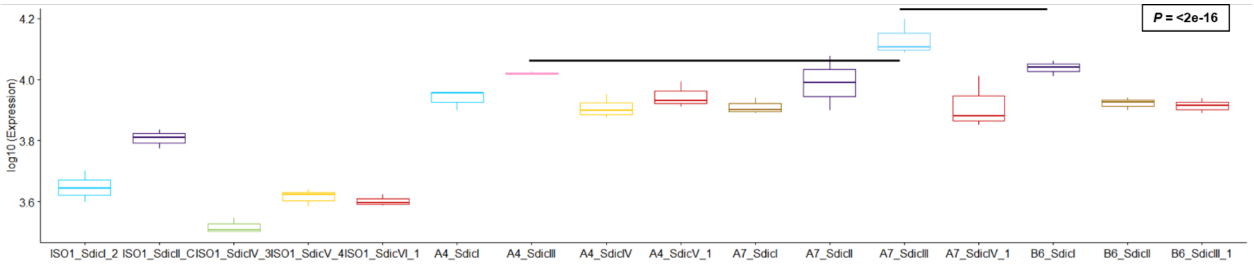


Figure 3.5. The expression comparison of all *Sdic* paralogs in testis shows the highest expression level for the TE harboring paralog, A7_*SdicIII*. Paralog expression is normalized across strains as RPKM. Expression is shown as $\log_{10}$(RPKM). Box plots represent the median surrounded by quartiles. Paratypes are colored as in fig. 3.1A. Detecting expression of paralogs ISO-1_*SdicB* and A4_*SdicII* was problematic, so they are omitted here**.** The one-way ANOVA *P*-value is shown in the black box. Black bars connect insignificant pairwise comparisons (Tukey HSD) involving A7_*SdicII*. Statistical values for all comparisons are listed in supplementary table S3.3.

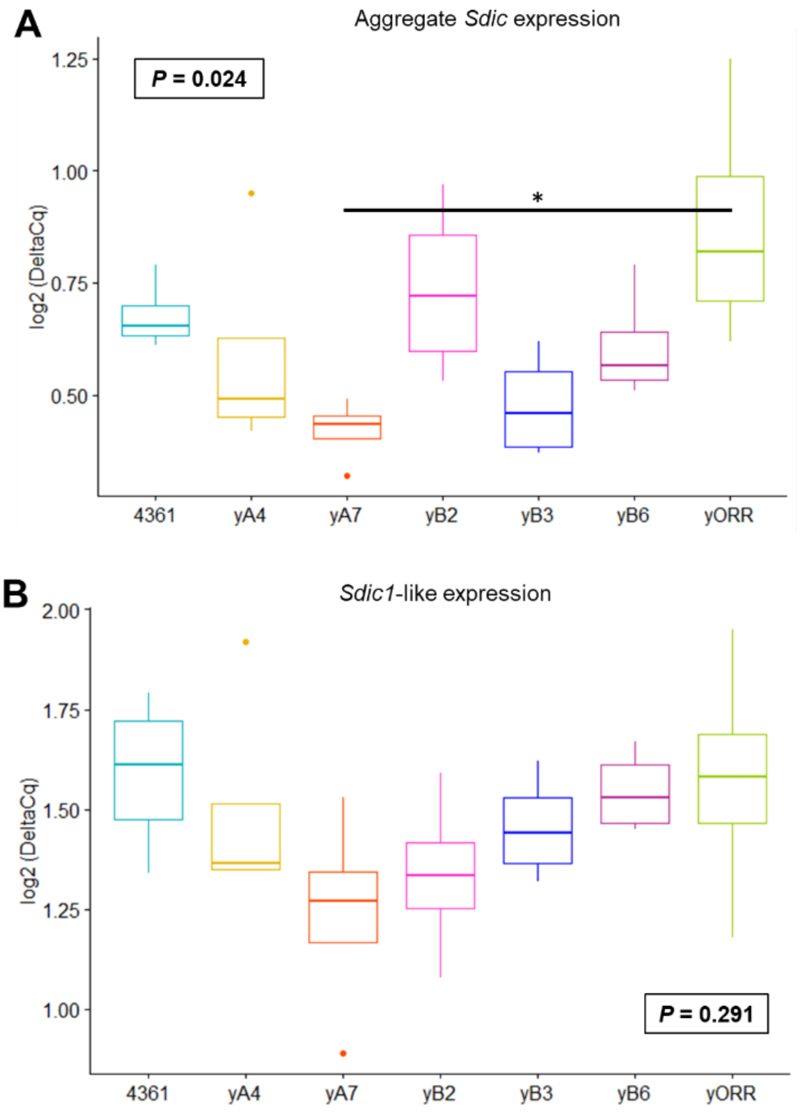


Figure 3.6. *Y* chromosome origin impacts the combined expression of the *Sdic* paralogs, but not *Sdic1*-like alone, in male whole-bodies across. Aggregate *Sdic* (A) and *Sdic1*-like (B) expression levels in male whole-bodies from each of the *Y* chromosome substitution strains from panel II are plotted as their qRT-PCR Delta-Cq values, *i.e.*, lower Delta-Cq values symbolize higher gene expression. Box plots represent the median surrounded by quartiles. Black boxes contain one-way ANOVA *P*-values. Black bars connect significant pairwise comparisons (Tukey HSD). *, *P* < 0.05. Statistical values for all comparisons are listed in supplementary table S3.4.

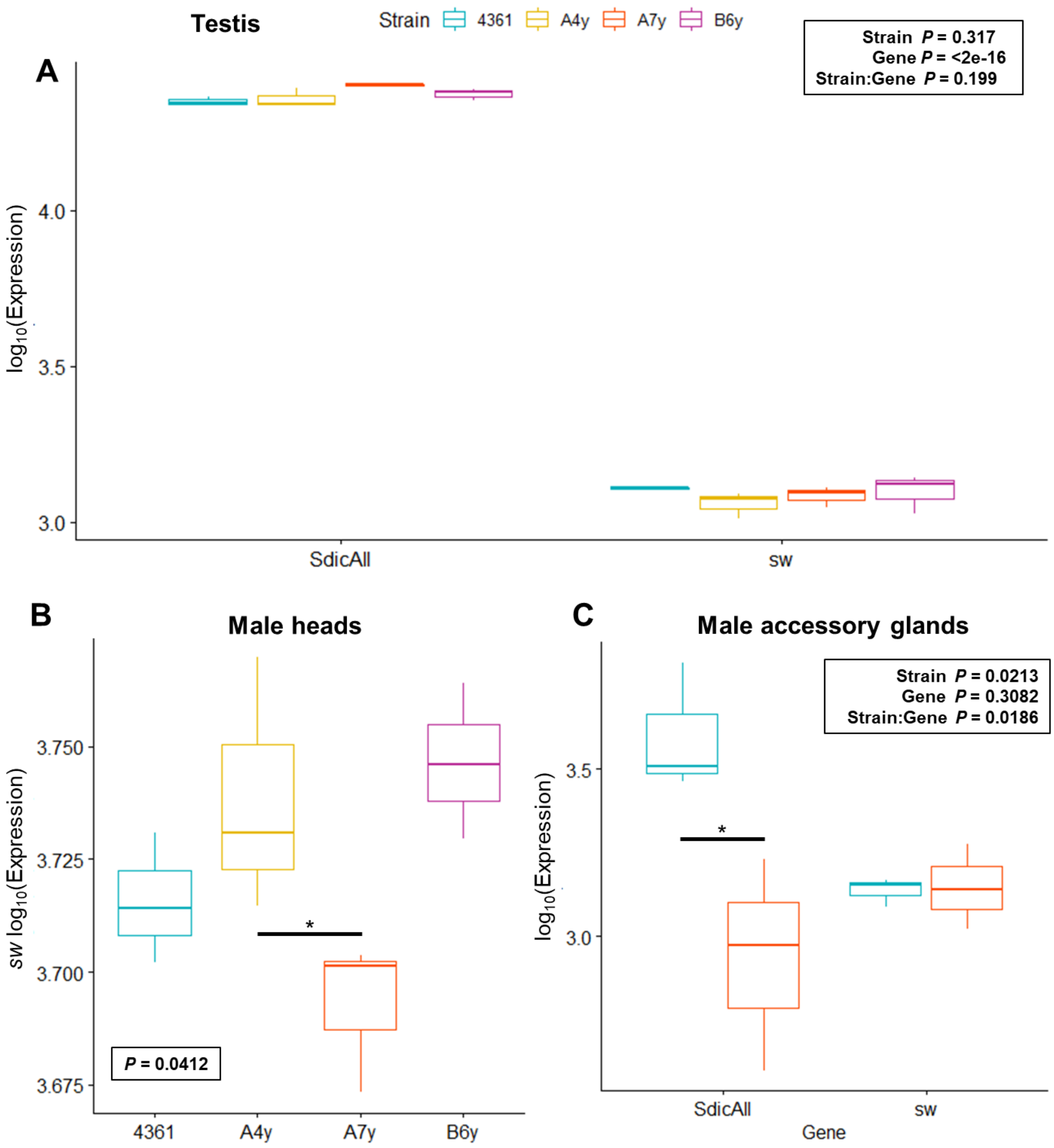


Figure 3.7. *Y*-linked regulatory variation impacts the aggregate expression of *Sdic* in accessory glands, *sw* in male heads, and neither in testis. (A) Two-way ANOVA comparing aggregate *Sdic* (*Sdic*All) and *sw* testis expression across panel II. (B) One-way ANOVA comparing *sw* male head expression across panel II. *Sdic* expression was not detected in male heads, so is omitted here. (C) Two-way ANOVA comparing aggregate *Sdic* (*Sdic*All) and *sw* male accessory gland expression across panel II. Expression is plotted as $\log_{10}$(RPKM). Box plots represent the mean surrounded by quartiles. The tissue under comparison is shown at the top. *P*-values are displayed at the top right. Black bars connect significant pairwise comparisons (Tukey HSD) (not shown for *Sdic*All and *sw* comparisons in (A)). *, $P < 0.05$. Statistical values for all comparisons are listed in supplementary table S3.5.

Table 3.1. *D. melanogaster* strains used in this study

Strain ID	Strain origin	*Sdic* CN *	Expression Assays †
I. *Sdic* CNV panel			
ISO-1	*D. melanogaster* reference strain	6	RNA-seq (T, O, H)
A4	Kariba Dam, South Africa	5	RNA-seq (T, O, H)
A7	Ken-Ting, Taiwan	4	RNA-seq (T, O, H)
B6	Ica, Peru	3	RNA-seq (T, O, H)
II. *Y* chromosome substitution panel			
4361: *y*[1]; *bw*[1]; *e*[4]; *ci*[1] *ey*[R]	Bloomington Stock Center	?	qRT-PCR WB, RNA-seq (T, AG, H)
A4y	Ranz Lab	?	qRT-PCR WB, RNA-seq (T, H)
A7y	Ranz Lab	?	qRT-PCR WB, RNA-seq (T, AG, H)
B2y	Ranz Lab	?	qRT-PCR WB
B3y	Ranz Lab	?	qRT-PCR WB
B6y	Ranz Lab	?	qRT-PCR WB, RNA-seq (T, H)
ORRy	Ranz Lab	?	qRT-PCR WB

***** CN, copy number (Clifton et al. 2020)

?, unknown.

† Tissues profiled: WB, whole-body; T, testis; O, ovary; H, male head; AG, male accessory glands.

Supplementary Table S3.1. Two-way ANOVA to test differences in aggregate *Sdic* vs *sw* expression and for each gene across strains

Test	**Contrast**	***P***
Two-way ANOVA	Aggregate *Sdic* vs *sw* in Testis (Strain)	0.000799
Two-way ANOVA	Aggregate *Sdic* vs *sw* in Testis (Gene)	< 2E-16
Two-way ANOVA	Aggregate *Sdic* vs *sw* in Testis (Strain:Gene)	4.94E-08
Tukey HSD	A7:SdicAll-A4:SdicAll	0.7282992
Tukey HSD	B6:SdicAll-A4:SdicAll	0.0085454
Tukey HSD	ISO1:SdicAll-A4:SdicAll	0.3005130
Tukey HSD	A4:sw-A4:SdicAll	0.0000000
Tukey HSD	A7:sw-A4:SdicAll	0.0000000
Tukey HSD	B6:sw-A4:SdicAll	0.0000000
Tukey HSD	ISO1:sw-A4:SdicAll	0.0000000
Tukey HSD	B6:SdicAll-A7:SdicAll	0.1766026
Tukey HSD	ISO1:SdicAll-A7:SdicAll	0.0166568
Tukey HSD	A4:sw-A7:SdicAll	0.0000000
Tukey HSD	A7:sw-A7:SdicAll	0.0000000
Tukey HSD	B6:sw-A7:SdicAll	0.0000000
Tukey HSD	ISO1:sw-A7:SdicAll	0.0000000
Tukey HSD	ISO1:SdicAll-B6:SdicAll	0.0000941
Tukey HSD	A4:sw-B6:SdicAll	0.0000000
Tukey HSD	A7:sw-B6:SdicAll	0.0000000
Tukey HSD	B6:sw-B6:SdicAll	0.0000000
Tukey HSD	ISO1:sw-B6:SdicAll	0.0000000
Tukey HSD	A4:sw-ISO1:SdicAll	0.0000000
Tukey HSD	A7:sw-ISO1:SdicAll	0.0000000
Tukey HSD	B6:sw-ISO1:SdicAll	0.0000000
Tukey HSD	ISO1:sw-ISO1:SdicAll	0.0000000
Tukey HSD	A7:sw-A4:sw	0.0000861
Tukey HSD	B6:sw-A4:sw	0.1195794
Tukey HSD	ISO1:sw-A4:sw	0.0506411
Tukey HSD	B6:sw-A7:sw	0.0235793
Tukey HSD	ISO1:sw-A7:sw	0.0000004
Tukey HSD	ISO1:sw-B6:sw	0.0001751

Supplementary Table S3.2. One-way ANOVAs to test differences in testes expression levels among paralogs in four strains

Test	Contrast	*P*
One-way ANOVA	ISO-1 paralogs in Testis	1.43E-15
Tukey HSD	SdicII_C-SdicI_2	0.0006291
Tukey HSD	SdicIII_B-SdicI_2	0.0000000
Tukey HSD	SdicIV_3-SdicI_2	0.0038375
Tukey HSD	SdicV_4-SdicI_2	0.8246113
Tukey HSD	SdicVI_1-SdicI_2	0.5686617
Tukey HSD	SdicIII_B-SdicII_C	0.0000000
Tukey HSD	SdicIV_3-SdicII_C	0.0000016
Tukey HSD	SdicV_4-SdicII_C	0.0001132
Tukey HSD	SdicVI_1-SdicII_C	0.0000609
Tukey HSD	SdicIV_3-SdicIII_B	0.0000000
Tukey HSD	SdicV_4-SdicIII_B	0.0000000
Tukey HSD	SdicVI_1-SdicIII_B	0.0000000
Tukey HSD	SdicV_4-SdicIV_3	0.0288513
Tukey HSD	SdicVI_1-SdicIV_3	0.0634103
Tukey HSD	SdicVI_1-SdicV_4	0.9964439
One-way ANOVA	A4 paralogs in Testis	8.44E-09
Tukey HSD	SdicII-SdicI	0.0000000
Tukey HSD	SdicIII-SdicI	0.0566038
Tukey HSD	SdicIV-SdicI	0.7932879
Tukey HSD	SdicV_1-SdicI	0.9981389
Tukey HSD	SdicIII-SdicII	0.0000000
Tukey HSD	SdicIV-SdicII	0.0000001
Tukey HSD	SdicV_1-SdicII	0.0000000
Tukey HSD	SdicIV-SdicIII	0.0101247
Tukey HSD	SdicV_1-SdicIII	0.0887373
Tukey HSD	SdicV_1-SdicIV	0.6348283
One-way ANOVA	A7 paralogs in Testis	0.0144000
Tukey HSD	SdicII-SdicI	0.5435150
Tukey HSD	SdicIII-SdicI	0.0189197
Tukey HSD	SdicIV_1-SdicI	0.9999064
Tukey HSD	SdicIII-SdicII	0.1298477
Tukey HSD	SdicIV_1-SdicII	0.5775890
Tukey HSD	SdicIV_1-SdicIII	0.0205796
One-way ANOVA	B6 paralogs in Testis	0.0010100
Tukey HSD	SdicII-SdicI	0.0021339
Tukey HSD	SdicIII_1-SdicI	0.0014542
Tukey HSD	SdicIII_1-SdicII	0.8914300

Supplementary Table S3.3. One-way ANOVA to test differences in expression in testes among paralogs across all panel I strains

Test	**Contrast**	***P***
One-way ANOVA	All *Sdic* paralogs (except ISO-1_B, A4_II) in testis	<2e-16
Tukey HSD	ISO1_SdicII_C-ISO1_SdicI_2	0.0073280
Tukey HSD	ISO1_SdicIV_3-ISO1_SdicI_2	0.0621458
Tukey HSD	ISO1_SdicV_4-ISO1_SdicI_2	0.9998877
Tukey HSD	ISO1_SdicVI_1-ISO1_SdicI_2	0.9956393
Tukey HSD	A4_SdicI-ISO1_SdicI_2	0.0000003
Tukey HSD	A4_SdicIII-ISO1_SdicI_2	0.0000000
Tukey HSD	A4_SdicIV-ISO1_SdicI_2	0.0000029
Tukey HSD	A4_SdicV_1-ISO1_SdicI_2	0.0000002
Tukey HSD	A7_SdicI-ISO1_SdicI_2	0.0000024
Tukey HSD	A7_SdicII-ISO1_SdicI_2	0.0000000
Tukey HSD	A7_SdicIII-ISO1_SdicI_2	0.0000000
Tukey HSD	A7_SdicIV_1-ISO1_SdicI_2	0.0000018
Tukey HSD	B6_SdicI-ISO1_SdicI_2	0.0000000
Tukey HSD	B6_SdicII-ISO1_SdicI_2	0.0000009
Tukey HSD	B6_SdicIII_1-ISO1_SdicI_2	0.0000018
Tukey HSD	ISO1_SdicIV_3-ISO1_SdicII_C	0.0000003
Tukey HSD	ISO1_SdicV_4-ISO1_SdicII_C	0.0006370
Tukey HSD	ISO1_SdicVI_1-ISO1_SdicII_C	0.0002412
Tukey HSD	A4_SdicI-ISO1_SdicII_C	0.0590018
Tukey HSD	A4_SdicIII-ISO1_SdicII_C	0.0001172
Tukey HSD	A4_SdicIV-ISO1_SdicII_C	0.3050321
Tukey HSD	A4_SdicV_1-ISO1_SdicII_C	0.0357390
Tukey HSD	A7_SdicI-ISO1_SdicII_C	0.2703040
Tukey HSD	A7_SdicII-ISO1_SdicII_C	0.0014052
Tukey HSD	A7_SdicIII-ISO1_SdicII_C	0.0000000
Tukey HSD	A7_SdicIV_1-ISO1_SdicII_C	0.2267482
Tukey HSD	B6_SdicI-ISO1_SdicII_C	0.0000299
Tukey HSD	B6_SdicII-ISO1_SdicII_C	0.1413059
Tukey HSD	B6_SdicIII_1-ISO1_SdicII_C	0.2264934
Tukey HSD	ISO1_SdicV_4-ISO1_SdicIV_3	0.3604757
Tukey HSD	ISO1_SdicVI_1-ISO1_SdicIV_3	0.5766886
Tukey HSD	A4_SdicI-ISO1_SdicIV_3	0.0000000
Tukey HSD	A4_SdicIII-ISO1_SdicIV_3	0.0000000
Tukey HSD	A4_SdicIV-ISO1_SdicIV_3	0.0000000
Tukey HSD	A4_SdicV_1-ISO1_SdicIV_3	0.0000000
Tukey HSD	A7_SdicI-ISO1_SdicIV_3	0.0000000
Tukey HSD	A7_SdicII-ISO1_SdicIV_3	0.0000000
Tukey HSD	A7_SdicIII-ISO1_SdicIV_3	0.0000000
Tukey HSD	A7_SdicIV_1-ISO1_SdicIV_3	0.0000000
Tukey HSD	B6_SdicI-ISO1_SdicIV_3	0.0000000
Tukey HSD	B6_SdicII-ISO1_SdicIV_3	0.0000000
Tukey HSD	B6_SdicIII_1-ISO1_SdicIV_3	0.0000000

Supplementary Table S3.3. One-way ANOVA to test differences in expression in testes among paralogs across all panel I strains

Test	Contrast	*P*
Tukey HSD	ISO1_SdicVI_1-ISO1_SdicV_4	1.0000000
Tukey HSD	A4_SdicI-ISO1_SdicV_4	0.0000000
Tukey HSD	A4_SdicIII-ISO1_SdicV_4	0.0000000
Tukey HSD	A4_SdicIV-ISO1_SdicV_4	0.0000003
Tukey HSD	A4_SdicV_1-ISO1_SdicV_4	0.0000000
Tukey HSD	A7_SdicI-ISO1_SdicV_4	0.0000002
Tukey HSD	A7_SdicII-ISO1_SdicV_4	0.0000000
Tukey HSD	A7_SdicIII-ISO1_SdicV_4	0.0000000
Tukey HSD	A7_SdicIV_1-ISO1_SdicV_4	0.0000002
Tukey HSD	B6_SdicI-ISO1_SdicV_4	0.0000000
Tukey HSD	B6_SdicII-ISO1_SdicV_4	0.0000001
Tukey HSD	B6_SdicIII_1-ISO1_SdicV_4	0.0000002
Tukey HSD	A4_SdicI-ISO1_SdicVI_1	0.0000000
Tukey HSD	A4_SdicIII-ISO1_SdicVI_1	0.0000000
Tukey HSD	A4_SdicIV-ISO1_SdicVI_1	0.0000001
Tukey HSD	A4_SdicV_1-ISO1_SdicVI_1	0.0000000
Tukey HSD	A7_SdicI-ISO1_SdicVI_1	0.0000001
Tukey HSD	A7_SdicII-ISO1_SdicVI_1	0.0000000
Tukey HSD	A7_SdicIII-ISO1_SdicVI_1	0.0000000
Tukey HSD	A7_SdicIV_1-ISO1_SdicVI_1	0.0000001
Tukey HSD	B6_SdicI-ISO1_SdicVI_1	0.0000000
Tukey HSD	B6_SdicII-ISO1_SdicVI_1	0.0000000
Tukey HSD	B6_SdicIII_1-ISO1_SdicVI_1	0.0000001
Tukey HSD	A4_SdicIII-A4_SdicI	0.6071905
Tukey HSD	A4_SdicIV-A4_SdicI	0.9999653
Tukey HSD	A4_SdicV_1-A4_SdicI	1.0000000
Tukey HSD	A7_SdicI-A4_SdicI	0.9999892
Tukey HSD	A7_SdicII-A4_SdicI	0.9817987
Tukey HSD	A7_SdicIII-A4_SdicI	0.0005220
Tukey HSD	A7_SdicIV_1-A4_SdicI	0.9999983
Tukey HSD	B6_SdicI-A4_SdicI	0.3102531
Tukey HSD	B6_SdicII-A4_SdicI	1.0000000
Tukey HSD	B6_SdicIII_1-A4_SdicI	0.9999984
Tukey HSD	A4_SdicIV-A4_SdicIII	0.1714499
Tukey HSD	A4_SdicV_1-A4_SdicIII	0.7416789
Tukey HSD	A7_SdicI-A4_SdicIII	0.1966536
Tukey HSD	A7_SdicII-A4_SdicIII	0.9998961
Tukey HSD	A7_SdicIII-A4_SdicIII	0.1837558
Tukey HSD	A7_SdicIV_1-A4_SdicIII	0.2360766
Tukey HSD	B6_SdicI-A4_SdicIII	1.0000000
Tukey HSD	B6_SdicII-A4_SdicIII	0.3565547
Tukey HSD	B6_SdicIII_1-A4_SdicIII	0.2363391
Tukey HSD	A4_SdicV_1-A4_SdicIV	0.9994407

Supplementary Table S3.3. One-way ANOVA to test differences in expression in testes among paralogs across all panel I strains

Test	Contrast	*P*
Tukey HSD	A7_SdicI-A4_SdicIV	1.0000000
Tukey HSD	A7_SdicII-A4_SdicIV	0.6563922
Tukey HSD	A7_SdicIII-A4_SdicIV	0.0000532
Tukey HSD	A7_SdicIV_1-A4_SdicIV	1.0000000
Tukey HSD	B6_SdicI-A4_SdicIV	0.0604081
Tukey HSD	B6_SdicII-A4_SdicIV	1.0000000
Tukey HSD	B6_SdicIII_1-A4_SdicIV	1.0000000
Tukey HSD	A7_SdicI-A4_SdicV_1	0.9997563
Tukey HSD	A7_SdicII-A4_SdicV_1	0.9959455
Tukey HSD	A7_SdicIII-A4_SdicV_1	0.0009419
Tukey HSD	A7_SdicIV_1-A4_SdicV_1	0.9999345
Tukey HSD	B6_SdicI-A4_SdicV_1	0.4300950
Tukey HSD	B6_SdicII-A4_SdicV_1	0.9999992
Tukey HSD	B6_SdicIII_1-A4_SdicV_1	0.9999350
Tukey HSD	A7_SdicII-A7_SdicI	0.7015730
Tukey HSD	A7_SdicIII-A7_SdicI	0.0000650
Tukey HSD	A7_SdicIV_1-A7_SdicI	1.0000000
Tukey HSD	B6_SdicI-A7_SdicI	0.0710312
Tukey HSD	B6_SdicII-A7_SdicI	1.0000000
Tukey HSD	B6_SdicIII_1-A7_SdicI	1.0000000
Tukey HSD	A7_SdicIII-A7_SdicII	0.0250770
Tukey HSD	A7_SdicIV_1-A7_SdicII	0.7610193
Tukey HSD	B6_SdicI-A7_SdicII	0.9882785
Tukey HSD	B6_SdicII-A7_SdicII	0.8826781
Tukey HSD	B6_SdicIII_1-A7_SdicII	0.7613751
Tukey HSD	A7_SdicIV_1-A7_SdicIII	0.0000856
Tukey HSD	B6_SdicI-A7_SdicIII	0.4200955
Tukey HSD	B6_SdicII-A7_SdicIII	0.0001699
Tukey HSD	B6_SdicIII_1-A7_SdicIII	0.0000858
Tukey HSD	B6_SdicI-A7_SdicIV_1	0.0884584
Tukey HSD	B6_SdicII-A7_SdicIV_1	1.0000000
Tukey HSD	B6_SdicIII_1-A7_SdicIV_1	1.0000000
Tukey HSD	B6_SdicII-B6_SdicI	0.1480883
Tukey HSD	B6_SdicIII_1-B6_SdicI	0.0885778
Tukey HSD	B6_SdicIII_1-B6_SdicII	1.0000000

Supplementary Table S3.4. One-way ANOVA to test differences in aggregate *Sdic* and *Sdic1*-like whole body expression across panel II

Test	**Contrast**	***P***
One-way ANOVA	Aggregate *Sdic* in male whole-bodies	0.024
Tukey HSD	A4y:SdicAll-4361:SdicAll	0.9893271
Tukey HSD	A7y:SdicAll-4361:SdicAll	0.4011493
Tukey HSD	B2y:SdicAll-4361:SdicAll	0.9990741
Tukey HSD	B3y:SdicAll-4361:SdicAll	0.6765911
Tukey HSD	B6y:SdicAll-4361:SdicAll	0.9972237
Tukey HSD	ORRy:SdicAll-4361:SdicAll	0.6765911
Tukey HSD	A7y:SdicAll-A4y:SdicAll	0.8208430
Tukey HSD	B2y:SdicAll-A4y:SdicAll	0.8906520
Tukey HSD	B3y:SdicAll-A4y:SdicAll	0.9707221
Tukey HSD	B6y:SdicAll-A4y:SdicAll	0.9999981
Tukey HSD	ORRy:SdicAll-A4y:SdicAll	0.2728866
Tukey HSD	B2y:SdicAll-A7y:SdicAll	0.1956247
Tukey HSD	B3y:SdicAll-A7y:SdicAll	0.9990741
Tukey HSD	B6y:SdicAll-A7y:SdicAll	0.7354191
Tukey HSD	ORRy:SdicAll-A7y:SdicAll	0.0196544
Tukey HSD	B3y:SdicAll-B2y:SdicAll	0.4011493
Tukey HSD	B6y:SdicAll-B2y:SdicAll	0.9415855
Tukey HSD	ORRy:SdicAll-B2y:SdicAll	0.9052100
Tukey HSD	B6y:SdicAll-B3y:SdicAll	0.9362797
Tukey HSD	ORRy:SdicAll-B3y:SdicAll	0.0528372
Tukey HSD	ORRy:SdicAll-B6y:SdicAll	0.3482904
One-way ANOVA	*Sdic1*-like in male whole-bodies	0.291

Supplementary Table S3.5. ANOVAs to test differences in aggregate *Sdic* vs *sw* expression and for each gene across strains in testes, accessory glands, and male heads

Test	Comparison	*P*
Two-way ANOVA	Aggregate *Sdic* vs *sw* in Testis (Strain)	0.319
Two-way ANOVA	Aggregate *Sdic* vs *sw* in Testis (Gene)	<2E-16
Two-way ANOVA	Aggregate *Sdic* vs *sw* in Testis (Strain:Gene)	0.199
Tukey HSD	4361:sw-4361:SdicAll	0.0000000
Tukey HSD	A4y:sw-A4y:SdicAll	0.0000000
Tukey HSD	A7y:sw-A7y:SdicAll	0.0000000
Tukey HSD	B6y:sw-B6y:SdicAll	0.0000000
Two-way ANOVA	Aggregate *Sdic* vs *sw* in Accessory Glands (Strain)	0.0213
Two-way ANOVA	Aggregate *Sdic* vs *sw* in Accessory Glands (Gene)	0.3082
Two-way ANOVA	Aggregate *Sdic* vs *sw* in Accessory Glands (Strain:Gene)	0.0186
Tukey HSD	A7y:SdicAll-4361:SdicAll	0.0145638
Tukey HSD	A7y:sw-4361:SdicAll	0.0897361
Tukey HSD	4361:sw-A7y:SdicAll	0.6161207
Tukey HSD	A7y:sw-4361:sw	0.9999008
One-way ANOVA	*sw* in Male Head	0.0412
Tukey HSD	4361:sw-A4y:sw	0.5352380
Tukey HSD	4361:sw-A7y:sw	0.5284924
Tukey HSD	4361:sw-B6y:sw	0.3023218
Tukey HSD	A4y:sw-A7y:sw	0.0877543
Tukey HSD	A4y:sw-B6y:sw	0.9572215
Tukey HSD	A7y:sw-B6y:sw	0.0434636

Supplementary Table S3.6. PCR primer sets used

Amplicon	Ta (C)	Primer efficiency, R^2	Size (nt)	Forward Primer (5’-3’)	Reverse Primer (5’-3’)	Experiment
*Sdic*All	60	96.8%, 0.995	76	CGTATTCTACTTTGAGCGGCG*	GGAATGTTCGTAGCCTGCAC	qRT-PCR
Sdic1-like	60	92.7%, 0.999	195	TCTGGTCGCTAAAGGACACC*	CGTCGTACACGTACAGCTTGC	qRT-PCR
clot	60	99.9%, 0.988	82	GAGCGGGCATACTGGAAG	GCAACAGAGTGGGCAAGAAG	qRT-PCR
Gapdh2	52	n/a	761	CAAGCAAGCCGATAGATAAAC*	GTCAAATCGACCACGGAAA	RT-PCR

* Priming site spans a splice junction

Supplementary Table S3.7. Diagnostic motifs used to detect expression in RNA-seq datasets

Gene ID	Core Diagnostic Motif*	Extended Diagnostic Motif*
sw	GGCGTCGCGAGAAGGA	AGGCTGAGCTGGAACGCAAGAAGGCCAAGTTGGCCGCCCTGCGCGAGGAGAAGGATCGCCGGCGTCGCGAGAAGGAGATCAAGGACATGGAGGAGGCGGCCGGTCGCATTGGCGGCGGAGCAGGCATCGA
*Sdic*All	ACGTATTCTACTTTGAG	ATGGGCTTAGTACTGATTAAGTTTTTACGATCAACGTATTCTACTTTGAGCGGCGGAAAGAAACAGCCTCTCAACCTAAGCGTCTACAATGTGCAGGCTACGAACATTCCACCAAAAGAGACACTGGTCT
Sdic1-like	GAGCAGTACATCGC	GACACCAAGCCGCTGTACTCCTTTGAGCAGTACATCGCCTGGTCGCCCGTGCGACGGCAGCGGCCGCCTGGACCTGATAAAACTCAACCCAGACACGGAGCACCAGCCCTTCGCCGCGACTCCTGGACTC
ISO1_*SdicI_2*	AGCGTGGTGATG	AGCGCCAGTCTAAGGCCATTGCCATTACATCGATGGCCTTCCCGGCCAACGAGATCAATAGCGTGGTGATGGGCAGTGAGGACGGCTACGTCTACTCCGCCTCGCGCCACGGCCTGCGCTCCGGGGTCAA
ISO1 *SdicII C*	CCAAGCTGGTGGT	GACGAGCGCTGGTCGAAGAACCGCTGCATCACCAGCATGGACTGGTCCACCCACTTCCCCAAGCTGGTGGTGGGCTCGTACCACAACAACGAGGAGAGTCCGAACGAGCCGGACGGCGTGGTGATGGTGT
ISO1 *SdicIII B*	ACATATTATATT	CAAATAGGATTCATACTTGATTTTAAATTAGTGCAACTAACAAGATTGCAGAAAGATCGTTTATTGTTTATTCGACTGTCGGGCAGGCTGAAAGCAACACATAAATAAATTAAATGCTACATATTATATT
ISO1 *SdicIV 3*	GCCCAGAACTCAAAACTC	GAGTCGGTACCACTTTTGCTAATCTTACAGGAGCGACCTTGAATGAAATTTAATTTGTATTTTTGTATCTTTTGTGATCCCGCTACTGTGTATAGCCCAGAACTCAAAACTCAACCGCAGTCCAAGTGCT
ISO1_*SdicV_4*	CCCATCTTAGTGAG	TGCATCGGCGACGAGGCCGGCAAGCTGTACGTGTACGACGTGGCCGAGAACCTGGCGCAGCCATCGCGCGACGAATGGTCGCGGTTCAACACCCATCTTAGTGAGATCAAGATGAACCAGAGCGATGAGG
A4 *SdicI*	AGTCTAAGGCC	TCCTGGTCGCTGGACATGCTGTCGCAACCACAGGACACGCTGGAGCTGCAGCAGCGCCAGTCTAAGGCCATTGCCATTACATCGATGGCCTTCCCGGCCAACGAGATCAATAGCCTGGTGATGGGCAGTG
A4 *SdicII*	CCAGTTCTGTTCCCAC	AAAGCCCCCAAGAAATCAGGGGAAACTGCAAAATTCTCTAGATGACCACAGTAACTGTAACTGTATTATTTTGTTACTCAATATTGGTTTCATTTCATAGCTATTTTCCCAGTTCTGTTCCCAC
A4 *SdicIII*	GCGCGACGAGAT	TTGGACCCATCCGGTCTGCACGTGTGCATCGGCGACGAGGCCGGCAAGCTGTACGTGTACGACGTGGCCGAGAACCTGGCGCAGCCATCGCGCGACGAGATCAAGATGAACCAGAGCGATGAGGTCTAG
A4_*SdicIV*	GGCCGTCAAG	TCCGGTCTGCACGTGTGCATCGGCGACGAGGCCGTCAAGCTGTACGTGTACGACGTGGCCGAGAACCTGGCGCAGCCATCGCGCGACGAATGGTCGCGGTTCAACACCCATCTTAGTGAGATCAAGATGA
A7 *SdicI*	TAGCGTGGTGATGGG	GCAGCGCCAGTCTAAGGCCATTGCCATTACATCGATGGCCTTCCCGGCCAACGAGATCAATAGCGTGGTGATGGGCAGTGAGGACGGCTACGTCTACTCCGCCTCGCGCCACGGCCTGCGCTCCGGGGTC
A7 *SdicII*	CACGAAGGTGCCGA	CTGGACCTGTGGAACCTCAACCAAGACACGAAGGTGCCGACCGCCTCGATTGTCGTGGCGGGAGCACCAGCCCTTAACCGCGTCTCTTGGACCCCATCCGGTCTGCACGTGTGCATCGGCGACGAGGCCG
A7 *SdicIII*	ACGTCCCCCCGGCC	CTGCAGAACCTGGGCAACGGATTCACCTCCAAGCTGCCACCGGGCTATCTCACCCACGGCCTGCCCACCGTTAAGGACGTCCCCCCGGCCATCACACCACTCGAGATCAAGAAGGAGACTGAAGTGAAGA
B6_*SdicI*	GCGCGACGAGATC	TTGGACCCATCCGGTCTGCACGTGTGCATCGGCGACGAGGCCGGCAAGCTGTACGTGTACGACGTGGCCGAGAACCTGGCGCAGCCATCGCGCGACGAGATCAAGATGAACCAGAGCGATGAGGTCTAG
B6 *SdicII*	TGAGATCAAGAT	CGGCGACGAGGCCGTCAAGCTGTACGTGTACGACGTGGCCGAGAACCTGGCGCAGCCATCGCGCGACGAATGGTCGCGGTTCAACACCCATCTTAGTGAGATCAAGATGAACCAGAGCGATGAGGTCTAG

*Sequences shown correspond with the complement. Both complement and reverse complement sequences were used. In the case of different *Sdic* paralogs, the name of the strain appears first.

CONCLUSIONS

Regions harboring recently expanded gene clusters are hotspots for structural and functional change, having the potential to foster adaptive evolution (Brown et al. 1998; Newcomb et al. 2005; Perry et al. 2007; Jugulam et al. 2014). In contrast to RNA-based duplicates, which often recruit novel *cis*-regulatory sequences relative to those present in the original gene, complete DNA-based duplicates are less likely to evolve new functional attributes (Chen et al 2013; Assis & Bachtrog 2013). The evolutionary dynamics of most DNA-based duplicates has been previously studied involving either young tandem duplicates with a limited number of members (Osada and Innan 2008; Cardoso-Moreira et al 2016; Loehlin & Carroll 2016; Rogers et al 2017; Loehlin et al 2021; Zhang et al 2022), tandemly arranged families of ancient origin such as the globins or rRNA genes (Brown et al. 1972; Zimmer et al. 1980), duplicates produced through whole genome duplication or aneuploidy (Song et al 2020; Desvignes et al 2021; Gillard et al 2021; Shi et al 2021), or cases in which the functional data is limited or lacking (Moore and Purugganan 2003). Rarely has the functional evolution of recently expanded, tandemly arranged gene family composed of more than two paralogs been studied at the population level, while also having precise a sequence annotation of the individual paralogs (Clifton et al 2017; 2020).

I have generated a detailed portrait of the organization and patterns of intraspecific genetic and functional variation of arguably one of the most recently formed and structurally complex regions in the *D. melanogaster* euchromatin, *Sperm-specific dynein intermediate chain* (*Sdic*). My analysis of the *Sdic* region represents a step forward in the generation of accurate portraits of the organizational, sequence, and functional evolution of recently originated, tandemly arranged multigene families. By coupling long-read sequencing technologies (Eid et al. 2009; Chakraborty et al 2019) with RNA-seq data from multiple biological conditions (Graveley et al 2011; Brown et al 2014), and tailored analytical approaches that accommodate the particularities of members of this type of multigene family, I demonstrate that we can now perform unparalleled multilevel characterizations of structurally complex genomic regions. This methodology can be used to test relevant aspects associated with the expression properties of individual repeats within recently formed, tandemly arranged multigene families.

By analyzing raw reads from different long-sequencing read technologies (Berlin et al 2015; McCoy et al 2014) I demonstrated that the organization of the *Sdic* multigene family in the reference genome ISO-1 is incorrectly portrayed in both number and arrangement of the paralogs in the reference genome assembly presented by FlyBase (dos Santos et al. 2015). I annotated the *Sdic* region in the best *D. melanogaster* genomes available (Berlin et al 2015; Chakraborty et al 2019) and compared these to the *Sdic* copy numbers estimated through additional molecular and computational techniques, demonstrating that even in *reference-quality* genome assemblies, structurally complex genomic regions such as *Sdic* remain refractory to proper assembly and require external validation. The *Sdic* copy number variation (CNV) I discovered here provides compelling evidence that the *Sdic* region has undergone extensive structural remodeling in natural populations, adding *Sdic* to the limited list of NAHR hotspots whose evolutionary dynamics is likely to be influenced by sexual selection, although in this case at the post- rather than premating level (Karn and Laukaitis 2009; Pezer et al. 2015; Pezer et al. 2017). I found the frequency distribution for *Sdic* copy number (CN), in natural populations is far from that expected under a runaway amplification process in which additional functional copies would be correlated with higher expression, ultimately having a directional effect on the phenotype (Brown et al. 1998; Schmidt et al. 2010; Soh et al. 2014). The prevalence of individuals bearing intermediate CN values, mostly between four to eight copies, could result from a scenario of stabilizing selection, or from a mutation-drift equilibrium coupled with the action of purifying selection sculpting the range boundaries as proposed for some multigene families in mammals (Hollox 2008; Teitz et al. 2018).

For a subset of seven cosmopolitan populations from one of the panels analyzed, for which genetic changes could be tracked at both the sequence and structural levels, I annotated one structurally distinct version of the *Sdic* region per population. This level of variation results from both changes in CN and recent TE insertions. The 37 *Sdic* paralogs I annotated in these strains encode a variety of Sdic proteins which differ primarily at their carboxyl ends, where the protein sw presumably interacts with the dynein heavy chain, as inferred from its ortholog in *Dictyostelium* (*dicA*; Ma et al. 1999). In the strains that could be reliably analyzed, I discovered a remarkable diversity of Sdic proteins (*paratypes*) despite profuse gene conversion events. I found extensive structural differences associated with distinct carboxyl ends, no evidence of a particular paratype being preeminent in CN within any strain, no evidence of pseudogenization, and only one paratype present in all strains. This fixed paratype shows strong evidence of having evolved under positive selection both at coding and noncoding levels. These signatures of positive selection and the lack of evidence for pseudogenization among the *Sdic* paralogs scrutinized here provide strong support to the adaptive role of *Sdic*.

Importantly, all protein variants of Sdic and its parental gene sw possess a common *cytoplasmic dynein 1 intermediate chain 1/2* domain, suggesting Sdic could function similarly to sw. However, the lack of coiled-coil and serine-rich domains at the N-terminus of Sdic would presumably prevent the Sdic variants from interacting with the dynactin protein complex, which mediates the interaction of the dynein motor protein complex with a variety of subcellular structures (Nurminsky et al. 1998a; Ma et al. 1999). Overall, Sdic and sw might share a limited set of common interactions with other protein complex subunits and subcellular structures. I found both total *Sdic* and *sw* expression vary significantly across strains with *Sdic* CNV, showing a statistically significant negative correlation between expression levels**.** This conflicts with the hypothesis that higher levels of *sw* expression should require more *Sdic* expression to regulate any dosage dependent function of *sw* in testis. Protein complex subunits are particularly sensitive to dosage constraints (Zhang et al 2022). A negative correlation could be due to an unidentified mechanism that represses *sw* while over-expressing *Sdic*, which has been postulated as an important mechanism for maintaining dosage balance in biological systems (Tu et al 2016). This could be a mechanism that maintains total dynein intermediate chain dosage within a limit that does not significantly impede cytoplasmic dynein protein complex assembly and functionality. Testing this hypothesis will require identification of protein-protein interaction partners shared between, and unique to, Sdic and sw, as well as quantification of these proteins' expression levels.

I demonstrated that artificially doubling *Sdic* copy number within the same genomic background increased total expression by >2-fold but also found no correlation between total *Sdic* expression and *Sdic* copy number in male whole-bodies across a panel of geographically diverse isogenic lines, which was interpreted as the result of variation in expression modifiers acting in *cis* and *trans*. I more recently demonstrated a positive correlation between total *Sdic* expression and CN in testis, which suggests that *Sdic* dosage could be under positive selection in testis. This analysis highlights the importance of quantifying gene expression at the tissue-specific –or even better cell-specific– level in order to reveal biologically meaningful patterns.

The *Y* chromosome impacts male fitness and fertility (Carvalho et al 2001). The presence of *Sdic* impacts sperm competition (Yeh et al 2012) and *Sdic* expression is maximal in testis and high in male accessory glands (Clifton et al 2017) –consistent with the *out of the testis hypothesis* for the origin of new genes (Kaessmann 2010)–. Testis expression is typical of newly evolved genes, often thought to be the result of particularities of this tissue, namely a particularly permissive chromatin environment and the simplicity of promoter sequences required for expression (Kaessmann 2010; Guschanski et al 2017; Witt et al 2021). Based on these premises, I hypothesized that the *Y* chromosome could act as a *trans* regulator of *Sdic* expression. My panel of *Y* chromosome substitution lines did not identify any *trans* regulatory action of the *Y* chromosome on *Sdic* expression in testis or male heads but demonstrated an impact on total *Sdic* expression in accessory glands, a somatic tissue with a role in reproduction, and also *sw* expression in heads but not in testis or accessory glands. These results are in line with those documented with similar experiments involving *Y* chromosome substitutions lines in which the authors detected significant effects on the expression variation of this chromosome on somatic tissues or on genes primarily expressed in somatic tissues (Lemos et al 2008; Branco et al 2017; Wang et al 2018; Ågren et al 2020).

The *Sdic* multigene family shows a pattern of expression consistent with quick regulatory diversification among the paralogs. As is the case for other recently originated genes, *Sdic* was likely expressed in testes at a very early stage (Kaessmann 2010; Zhao et al. 2014). This is the only expression attribute in adults shared across all paralogs, whereas expression in females was displayed by multiple paralogs in ISO-1, varying across adult samples, including some that were inferred to be among the most recently generated in the gene family. I found evidence of expression for all *Sdic* paralogs present in four strains with *Sdic* CNV, including one containing a premature stop codon due to a TE insertion. The *Sdic* paralogs vary significantly in their expression level in all four strains, showing no consistent trend for promoter type or coded protein variant. This is still consistent with *cis* regulatory differences being present among the paralogs of each strain, although the identity of those differences is not apparent at this time. Precise identification of the *cis* regulatory impact of individual *Sdic* paralogs on one another will require quantification of *Sdic* expression in genetically engineered lines with different individual copies removed from the array. Further, it is not known if these tissue-level expression differences among paralogs are maintained at the effective functional level of the protein or are stabilized through post-translational buffering mechanisms. Precise quantification of the individual paralogs at the proteomic level is needed to test for the presence of these buffering mechanisms. The lack of consistent differences between *Sdic* paralogs with different promoters in the same strain, suggests that the *Sdic* promoters likely play a minimal role as *cis* regulators of *Sdic* expression, at least in the testis. Controlled experiments using genetically modified single copy *Sdic* regions in identical genomic backgrounds but driven by different *Sdic* promoters will be needed to properly evaluate how promoter evolution contributes to the functional divergence of *Sdic* paralogs (Jimenez-Morales et al 2020). While I found no evidence of positional effects, *i.e.*, the position of the paralog within the tandem array, acting on expression of the *Sdic* paralogs, the action of positional effects on young tandem gene families should be more precisely identified using genetically engineered lines with shuffled orders of paralogs within the same genomic background. Further, this work suggests that functional diversification of tandemly arranged duplicates may proceed through posttranscriptional regulatory changes driven by the evolution of a unique composition of miRNA binding sites (Wang & Adams 2015; Catalan et al 2016), revealing an important path for the diversification of DNA-mediated duplicates. Further work with genetically engineered lines containing single paralogs with endogenous and exogenous 3’ UTRs can further illuminate these mechanisms.

Relative to *Sdic*’s organismic impact, I measured *Sdic* expression in two strains harboring complete duplications of the *Sdic* cluster within the same genomic background, for which sperm competition data was also available. While increasing *Sdic* CN did not increase sperm competitively ability, one strain showed ~3-fold increased expression and had similar sperm competitive ability to its derived wildtype strain while the other strain show ~4-fold increased expression and significantly decreased sperm competitive ability, suggesting *Sdic* dosage might be constrained to an optimal level. At this moment, the sperm boosting effect of *Sdic* has only been seen between wildtype and complete *Sdic* cluster knockout strains. Only by testing the competitive ability of sperm from strains with intermediate *Sdic* CN values within the same genomic background will the interplay between *Sdic* dosage and fitness become clear (Kondrashov 2010). Further, the functional complexity of the *Sdic* copies I have revealed through analysis of protein domain compositions and expression profiles, questions whether the phenotypic impact of the *Sdic* region is confined to post-mating male–male competition. I hypothesized that *Sdic* expression in females could result in a sexually antagonistic effect as circumstantial evidence suggests (Innocenti & Morrow 2010), which would fit with the notion that the *X* chromosome, where *Sdic* resides, is a key genomic reservoir of sexually antagonistic genetic variation (Rice 1984; Gibson et al. 2002). My observation of female fecundity suggests that should this antagonistic effect exist, it impacts either a more subtle fertility component or a completely different type of trait than was tested here. Further, the structural features and the expression profiles exhibited by some *Sdic* paralogs, are suggestive of an Sdic protein that interacts with non-axonemal dynein, *i.e.*, cytoplasmic dynein, protein complexes present in tissues possessing both ciliated (*e.g.*, sperm) and non-ciliated cells (*e.g.*, salivary glands and imaginal discs). Whether or not Sdic interacts with axonemal dynein complexes cannot be inferred from our results, but the fact that the silencing of the whole *Sdic* family results in a significant reduction in sperm competitive ability does not discard this possibility (Yeh et al. 2012).

Regardless of *Sdic*'s organismic impact, my work shows that the amplification of *Sdic* cluster has not merely resulted increased gene dosage. Collectively, my work suggests that *Sdic* CNV observed in contemporary *D. melanogaster* populations works to secure an optimal expression level across different genomic backgrounds and sexual selection regimes while also providing a substrate for gene conversion and NAHR events to prevent accumulation of nucleotide changes along the entire *Sdic* repeat except for the carboxyl end of the proteins and 3'UTR of the transcripts (Rozen et al. 2003; Teitz et al. 2018). *Sdic* paralogs in conventional laboratory strains show evidence of expression divergence between paralogs and across life stages and anatomical parts of the adult, which is concurrent with profound 3'UTR remodeling (Mayer 2019). Further, maintaining multiple paralogs that encode different and possibly fully functional proteins is compatible with a mechanism that safeguards functional diversity at the protein level while enabling mRNA expression profile diversification (Traherne et al. 2010).

Overall, my dissertation represents a sophisticated paralog- and tissue-level characterization of a species-specific multigene family, adding to a few others such as those reported in *Homo sapiens* (Dougherty et al 2018; Fiddes et al 2018) and *Cannabis sativa* (Vergara et al 2019). This work pioneers the proper reconstruction of structurally complex genomic regions while characterizing them at the functional level. This dissertation highlights the importance of combining molecular and sequencing approaches to obtain paralog-specific information that can be used for generating a more nuanced portrait of how recent expansions at NAHR hotspots are functionally evolving along their path to fixation and consolidation in the genome as multigene families. Nevertheless, it still remains a challenge to fully understand the evolutionary implications of *Sdic* amplification. Having a full understanding of how specific genetic changes in promoters, 3' UTRs, and other *cis*-regulatory motifs impact the expression of different paralogs will require the generation of controlled synthetic genotypes harboring paralogs of different ages and features to uncover how regulatory mechanisms that diversify gene functions evolve as gene duplicates age.

REFERENCES

Ågren JA, Munasinghe M, and Clark AG. 2020. Mitochondrial-*Y* chromosome epistasis in *Drosophila melanogaster*. *Proc. R. Soc. B*. 28720200469

Assis R, Bachtrog D. 2013. Neofunctionalization of young duplicate genes in *Drosophila*. *Proc Natl Acad Sci U S A*. 110(43):17409-14. doi: 10.1073/pnas.1313759110

Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol*. 33:623–630.11

Branco AT, Schilling L, Silkaitis K, Dowling DK, Lemos B. 2017. Reproductive activity triggers accelerated male mortality and decreases lifespan: genetic and gene expression determinants in *Drosophila*. *Heredity*

Brown DD, Wensink PC, Jordan E. 1972. A comparison of the ribosomal DNA's of *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. *J Mol Biol* 63:57–73.

Brown CJ, Todd KM, Rosenzweig RF. 1998. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Mol Biol Evol*. 15(8):931–942.

Brown JB, Boley N, Eisman R, MayGE, StoiberMH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, et al. 2014. Diversity and dynamics of the *Drosophila* transcriptome. *Nature*. 512:393–399.

Cardoso-Moreira M, Arguello JR, Gottipati S, Harshman LG, Grenier JK, Clark AG. 2016. Evidence for the fixation of gene duplications by positive selection in *Drosophila*. *Genome Res*. 26(6):787–798.

Carvalho AB, Dobo BA, Vibranovski MD, Clark AG. 2001. Identification of five new genes on the *Y* chromosome of *Drosophila melanogaster*. PNAS

Catalan A, Glaser-Schmitt A, Argyridou E, Duchen P, Parsch J. 2016. An indel polymorphismin the *MtnA* 3' untranslated region is associated with gene expression variation and local adaptation in *Drosophila melanogaster*. *PLoS Genet*. 12(4):e1005987.

Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. 2019. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat Commun*. 10(1):4872.

Chen S, Krinsky B, Long M. 2013. New genes as drivers of phenotypic evolution. *Nat Rev Genet*. 14:645–660

Clifton BD, Librado P, Yeh SD, Solares ES, Real DA, Jayasekera SU, Zhang W, Shi M, Park RV, Magie RD, Ma H, Xia X, Marco A, Rozas J, Ranz JM. 2017. Rapid functional and sequence differentiation of a tandemly repeated species-specific multigene family in *Drosophila*. *Mol Biol Evol*. 34(1):51–65

Clifton BD, Jimenez J, Kimura A, Chahine Z, Librado P, Sanchez-Gracia A, Abbassi M, Carranza F, Chan C, Marchetti M, Zhang W, Shi M, Vu C, Yeh S, Fanti L, Xia X, Rozas J, Ranz JM. 2020. Understanding the early evolutionary stages of a tandem *D. melanogaster*-specific gene family: a structural and functional population study. *Mol Biol Evol*. 37(9):2584–2600

Desvignes T, Sydes J, Montfort J, Bobe J, Postlethwait JH. 2021. Evolution after Whole-Genome Duplication: Teleost MicroRNAs. *Mol Biol Evol*. 38(8):3308-3331

dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, Emmert DB, Gelbart WM, FlyBase C, the FlyBase Consortium. 2015. FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res*. 43(D1):D690–D697.

Dougherty ML, Underwood JG, Nelson BJ, Tseng E, Munson KM, Penn O, Nowakowski TJ, Pollen AA, Eichler EE. 2018. Transcriptional fates of human-specific segmental duplications in brain. *Genome Res*. 28(10):1566-1576.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–138.

Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, van den Bout A, Bishara A, Rosenkrantz JI, Lorig-Roach R, Field AR, Maeussler M, Russo L, Bhaduri A, Nowakowski TJ, Pollen AA, Dougherty ML, Nuttle X, Addor MC, Zwolinski S, Katzman S, Kriegstein A, Eichler EE, Salama SR, Jacobs FMJ, Haussler D. 2018. Human-specific *NOTCH2NL* genes affect Notch signaling and cortical neurogenesis. *Cell*. 173(6): 1356–1369.e22.

Gibson JR, Chippindale AK, Rice WR. 2002. The *X* chromosome is a hot spot for sexually antagonistic fitness variation. *Proc Biol Sci*. 269:499–505.

Gillard GB, Grønvold L, Røsæg LL. Holen MM, Monsen O, Koop BF, Rondeau EB, Gundappa K, Mendoza J, Macqueen DJ, Rohlfs RV, Sandve SR, Hvidsten TR. 2021. Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. *Genome Biol*. (22)103

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van BarenMJ, Boley N, Booth BW, et al. 2011. The developmental transcriptome of *Drosophila melanogaster*. *Nature*. 471:473–479.

Guschanski K, Warnefors M, Kaessmann H. 2017. The evolution of duplicate gene expression in mammalian organs. *Genome Research*. 27:1461-1474

Hollox EJ. 2008. Copy number variation of beta-defensins and relevance to disease. *Cytogenet Genome Res*. 123(1–4):148–155.

Innocenti P, Morrow EH. 2010. The sexually antagonistic genes of *Drosophila melanogaster*. PLoS Biol. 8:e1000335.

Jimenez-Morales E, Aguilar-Herandez V, Aguilar-Henonin L, Guzman P. 2018. Molecular basis for neofunctionalization of duplicated E3 ubiquitin ligases underlying adaptation to drought tolerance in *Arabidopsis thaliana*. *The Plant Journal*. 104:474-492.

Jugulam M, Niehues K, Godar AS, Koo DH, Danilova T, Friebe B, Sehgal S, Varanasi VK, Wiersma A, Westra P, et al. 2014. Tandem amplification of a chromosomal segment harboring 5-enolpyruvylshikimate-3-phosphate synthase locus confers glyphosate resistance in *Kochia scoparia*. *Plant Physiol*. 166(3):1200–1207.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res*. 20(10):1313–1326

Karn RC, Laukaitis CM. 2009. The mechanism of expansion and the volatility it created in three pheromone gene clusters in the mouse (*Mus musculus*) genome. *Genome Biol Evol*. 1:494–503.

Kondrashov FA. 2010. Gene Dosage and Duplication. Evolution After Gene Duplication. *Wiley-Blackwell*

Lemos B, Araripe LO, Hartl DL. 2008. Polymorphic *Y* Chromosomes Harbor Cryptic Variation with Manifold Functional Consequences. *Science*

Loehlin DW, Carroll SB. 2016. Expression of tandem gene duplicates is often greater than twofold. *Proc Natl Acad Sci USA.* 113(21):5988–5992.

Loehlin DW, Kim JY, Paster CO. 2021. A tandem duplication in *Drosophila melanogaster* shows enhanced expression beyond the gene copy number. *Genetics*.

McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, Petrov DA, Fiston-Lavier AS. 2014. Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PLoS One*. 9:e106689.

Ma S, Trivinos-Lagos L, Graf R, Chisholm RL. 1999. Dynein intermediate chain mediated dynein-dynactin interaction is required for interphase microtubule organization and centrosome replication and separation in *Dictyostelium*. *J Cell Biol*. 147:1261–1274.

Mayer C. 2019. What are 3’ UTRs doing? *Cold Spring Harb Perspect Biol*. 11:a034728

Moore RC, Purugganan MD. 2003. The early stages of duplicate gene evolution. *Proc Natl Acad Sci USA*. 100(26):15682-7

Newcomb RD, Gleeson DM, Yong CG, Russell RJ, Oakeshott JG. 2005. Multiple mutations and gene duplications conferring organophosphorus insecticide resistance have been selected at the Rop-1 locus of the sheep blowfly, *Lucilia cuprina*. *J Mol Evol*. 60:207–220.

Nurminsky DI, Nurminskaya MV, Benevolenskaya EV, Shevelyov YY, Hartl DL, Gvozdev VA. 1998a. Cytoplasmic dynein intermediate chain isoforms with different targeting properties created by tissue-specific alternative splicing. Mol Cell Biol. 18:6816–6825.

Osada N, Innan H. 2008. Duplication and gene conversion in the *Drosophila melanogaster* genome. *PLoS Genet*. 4:e1000305.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet*. 39:1256–1260.

Pezer Z, Harr B, Teschke M, Babiker H, Tautz D. 2015. Divergence patterns of genic copy number variation in natural populations of the house mouse (*Mus musculus domesticus*) reveal three conserved genes with major population-specific expansions. *Genome Res*. 25(8):1114–1124.

Pezer Z, Chung AG, Karn RC, Laukaitis CM. 2017. Analysis of copy number variation in the *Abp* gene regions of two house mouse subspecies suggests divergence during the gene family expansions. *Genome Biol Evol*. 9(6):1393–1405.

Rice WR. 1984. Sex chromosomes and the evolution of sexual dimorphism. *Evolution.* 38:735–742.

Rogers RL, Shao L, Thornton KR. 2017. Tandem duplications lead to novel expression patterns through exon shuffling in *Drosophila yakuba*. *PLoS Genet*. 13(5):e1006795.

Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC. 2003. Abundant gene conversion between arms of palindromes in human and ape *Y* chromosomes. *Nature*. 423(6942):873–876.

Schmidt JM, Good RT, Appleton B, Sherrard J, Raymant GC, Bogwitz MR, Martin J, Daborn PJ, Goddard ME, Batterham P, et al. 2010. Copy number variation and transposable elements feature in recent, ongoing adaptation at the *Cyp6g1* locus. *PLoS Genet*. 6(6):e1000998

Shi X, Yang H, Chen C, Hou J, Hanson KM, Albert PS, Ji T, Cheng J, Birchler JA. 2021. Genomic imbalance determines positive and negative modulation of gene expression in diploid maize. *The Plant Cell*. 33(4):917–939

Soh YQS, Alfoldi J, Pyntikova T, Brown LG, Graves T, Minx PJ, Fulton RS, Kremitzki C, Koutseva N, Mueller JL, et al. 2014. Sequencing the mouse *Y* chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell*. 159(4):800–813.

Song MJ, Potter BI, Doyle JJ, Coate JE. 2020. Gene Balance Predicts Transcriptional Responses Immediately Following Ploidy Change in Arabidopsis thaliana. *The Plant Cell*. 32(5):1434–1448

Teitz LS, Pyntikova T, Skaletsky H, Page DC. 2018. Selection has countered high mutability to preserve the ancestral copy number of *Y* chromosome amplicons in diverse human lineages. *Am J Hum Genet*. 103(2):261–275.

Traherne JA, Martin M, Ward R, Ohashi M, Pellett F, Gladman D, Middleton D, Carrington M, Trowsdale J. 2010. Mechanisms of copy number variation and hybrid gene formation in the KIR immune gene complex. *Hum Mol Genet*. 19(5):737–751.

Tu X, Wang Y, Zhang M, Wu J. 2016. Using Formal Concept Analysis to Identify Negative Correlations in Gene Expression Data. *IEEE/ACM Trans Comput Biol Bioinform*. 13(2):380-91

Vergara D, Huscher EL, Keepers KG, Givens RM, Cizek CG, Torres A, Gaudino R, Kane NC. 2019. Gene copy number is associated with phytochemistry in *Cannabis sativa*. *AoB Plants*. 11(6).

Wang S, Adams KL. 2015. Duplicate gene divergence by changes in microRNA binding sites in Arabidopsis and Brassica. *Genome Biol Evol*. 7:646–655.

Wang M, Branco AT, Lemos B. 2018. The *Y* Chromosome Modulates Splicing and Sex-Biased Intron Retention Rates in *Drosophila*. *Genetics*. 208(3):1057-1067

Witt E, Svetec N, Benjamin S, Zhao L. 2021. Transcription Factors Drive Opposite Relationships between Gene Age and Tissue Specificity in Male and Female *Drosophila* Gonads. *Molecular Biology and Evolution*. 38(5):2104–2115

Yeh SD, Do T, Chan C, Cordova A, Carranza F, Yamamoto EA, Abbassi M, Gandasetiawan KA, Librado P, Damia E, et al. 2012. Functional evidence that a recently evolved *Drosophila* sperm-specific gene boosts sperm competition. *Proc Natl Acad Sci USA*. 109(6):2043–2048.

Zhang D, Leng L, Chen C, Huang J, Zhang Y, Yuan H, Ma C, Chen H, Zhang Y. 2022. Dosage sensitivity and exon shuffling shape the landscape of polymorphic duplicates in *Drosophila* and humans. *Nature Ecology and Evolution*. 6:273-287

Zhao L, Saelao P, Jones CD, Begun DJ. 2014.Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science*. 343:769–772.

Zimmer EA, Martin SL, Beverley SM, Kan YW, Wilson AC. 1980. Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proc Natl Acad Sci USA*. 77:2158–2162.