Department of Biomedical Sciences
University of Veterinary Medicine Vienna

Institute of Population Genetics
(Head: Univ.-Prof. Dr.rer.nat. Christian Schlötterer)

Towards Unraveling the Dynamics of Transposable Elements

PhD thesis submitted for the fulfilment of the requirements for the degree of
DOCTOR OF PHILOSOPHY (PhD)

University of Veterinary Medicine Vienna

Submitted by

Florian Schwarz

Vienna, October 2021

PhD Committee Members

Dipl.-Ing. Dr.nat.techn. Priv.-Doz. Robert Kofler

Institute of Population Genetics

University of Veterinary Medicine Vienna

Robert.Kofler@vetmeduni.ac.at

Prof. Dr. Qi Zhou

Life Sciences Institute

Zhejiang University, China

& ERC Group Leader

Department of Neurosciences and Developmental Biology

University of Vienna

qi.zhou@univie.ac.at

Univ.-Prof. Dr. rer.nat. Christian Schlötterer

Institute of Population Genetics

University of Veterinary Medicine Vienna

Christian.Schloetterer@vetmeduni.ac.at

List of Publications

*These authors contributed equally

List of original publications

Filip Wierzbicki*, **Florian Schwarz***, Odontsetseg Cannalonga, and Robert Kofler (2021), Novel quality metrics allow identifying and generating high-quality assemblies of piRNA clusters. **Molecular Ecology Resources**, 00, 1–20.
https://doi.org/10.1111/1755-0998.13455
Impact Factor (2019/2020): 7.059

Florian Schwarz, Filip Wierzbicki, Kirsten-André Senti, Robert Kofler (2020), Tirant Stealthily Invaded Natural *Drosophila melanogaster* Populations during the Last Century, **Molecular Biology and Evolution**, Volume 38, Issue 4, Pages 1482–1497.
https://doi.org/10.1093/molbev/msaa308
Impact Factor (2019/2020): 16.240

Acknowledgements

To acknowledge everyone who contributed directly or indirectly to this work and my PhD journey in general would sadly exceed the frame of this thesis. First and foremost I want to thank my Supervisor, Dr. Robert Kofler, for providing unlimited scientific and personal support and extraordinary trust over the whole duration of my PhD endeavor. Without you, none of this would have been possible. I always expected the term 'Doktorvater' to merely be a phrase, but I have been proven wrong. Second, I want to thank Filip Wierzbicki for his continuous support, no matter the time of day or my mood. Working with you are some of the best memories from all of my PhD. I also want to acknowledge the members of my PhD committee, Prof. Qi Zhou and Prof. Christian Schlötterer, whose scientific support has certainly elevated the quality of my work. This work would not have been possible without the scientific and personal support of Odontsetseg Cannalonga, Kirsten-André Senti, Divya Selvarayu and Elisabeth Salbaba. Ideally, I would need to thank every single member of the Institute of Population Genetics and the Graduate School of Population Genetics by name. However, I want to particularly highlight Lauri Törmä, my gym buddy and best friend I met in Vienna, Anna-Maria Langmüller, my Institute Mentor, Marta Pelizzola for her continued support and friendship, as well as all members of the Cooking group for countless great meals and so many new things I learned. I want to thank all members of the EES master program, specifically Sabrina Duncan and Alexander Hausmann, as well as Prof. Dirk Metzler and Prof. Jochen Wolf from the LMU Munich for motivating me to undergo this PhD adventure in a new city and country. I also want to thank all my friends from Munich, without the Mythbusters all of this would be pointless. This work was also only possible due to the support of my family and their unconditional love. Particularly, I want to thank my sister Angela, whose continued support, despite occasional perplexity of what I actually do for a living, I highly appreciate. Lastly, I want to dedicate this thesis to my mother Beate, the person who I owe everything to. Danke schön!

Contributions

*These authors contributed equally

List of original publications

Filip Wierzbicki*, **Florian Schwarz***, Odontsetseg Cannalonga, and Robert Kofler (2021), Novel quality metrics allow identifying and generating high-quality assemblies of piRNA clusters. **Molecular Ecology Resources**, 00, 1–20. https://doi.org/10.1111/1755-0998.13455

Author contributions:
RK, **FS**, and FW conceived this work. **FS** and OC generated the data. FW performed PCR. **FS** and FW analyzed the data. RK and FW provided software. RK, **FS**, and FW wrote the manuscript.

Florian Schwarz, Filip Wierzbicki, Kirsten-André Senti, Robert Kofler (2020), Tirant Stealthily Invaded Natural *Drosophila melanogaster* Populations during the Last Century, **Molecular Biology and Evolution**, Volume 38, Issue 4, Pages 1482–1497. https://doi.org/10.1093/molbev/msaa308

Author contributions:
FS and RK conceived the work. **FS** and FW analyzed the data. K-AS provided feedback on the manuscript. **FS** and RK wrote the manuscript.

Declaration

I confirm that I have followed the rules of good scientific practice in all aspects.

Florian Schwarz, October 2021

Zusammenfassung

Ein erheblicher Anteil des Genoms vieler Organismen beinhaltet Transposons (abk. TEs). TEs sind komplexe, breit verteilte repetitive genomische Teilelemente welche sich innerhalb eines Wirtsgenoms egoistisch vermehren. Kürzlich veröffentlichte Studien haben gezeigt, dass die Dynamiken von TEs eine wichtige Rolle in der adaptiver Evolution von Organismen spielen können. Jedoch verbleiben viele Details über die Dynamiken von TEs in natürlichen Populationen weiterhin unklar, einschließlich ihres Grades an Polymorphie sowie welche Faktoren den beobachteten Unterschieden in Häufigkeit und Vielfalt zugrunde liegen. In dieser Arbeit beabsichtige ich zu der Entschlüsselung der Dynamiken von TEs in natürlichen Populationen beizutragen. Zusätzlich beabsichtige ich neue Herangehensweisen und Methoden zu entwickeln, welche Wissenschaftler mit den nötigen Wekzeugen ausstatten um diese Dynamiken genau zu beschreiben.

Im ersten Abschnitt entwickle ich neue Methodiken um die Qualität der Vertretung von TEs und repetitiven Regionen in konstruierten Genomen (genannt 'Assemblies') zu beurteilen. Diese Methodiken zielen darauf ab, technologische Limitierungen bezüglich der genomischen Charakterisierung und der korrekten Repräsentation von TEs und repetitiven Regionen zu überwinden, welche durch den repetitiven Charakter von TEs verursacht werden. Hierbei beabsichtige ich besonders, die Untersuchung von funktionell relevanten genomischen Anordnungen von TE Insertionen zu ermöglichen, welche in das piRNA-Verteidigungssystem gegen unkontrollierte Transpositionsaktivität eingebunden sind. Diese Regionen werden 'piRNA cluster' genannt und sind in genomischen Assemblies aufgrund ihrer repetitiven Eigenschaften besonders kompliziert zu untersuchen. Zusätzlich sind bereits existierende Metriken zur Evaluierung von genomischen Assemblies größtenteils nicht aussagekräftig bezüglich der Qualität der Representation von TEs und repetitiven Regionen. Ich gehe dieses Problem an indem ich mehrere Metriken etabliere, welche die Qualität einer Assembly in Bezug auf die Repräsentation von TEs und repetitiven Regionen wiedergeben. Hierbei konzentriere ich mich speziell auf piRNA cluster. Ich wende diese Metriken an um neue genomische Assemblies mit einer hohen Qualität bezüglich TEs und piRNA clustern zu konstruieren. Zusätzlich zeige ich wie diese Metriken für eine breitere Anwendung in anderen genomischen Regionen oder Organismen erweitert werden können.

Zuletzt zeige ich auf, wie zukünftige Studien diese Methodik zur empirischen Beschreibung der Dynamiken von piRNA clustern benutzen könnten.

Im zweiten Abschnitt meiner Arbeit beschreibe ich die kürzlich und verstohlen erfolgte Invasion des LTR retrotransposons Tirant in *Drosophila melanogaster*. Zu diesem Zweck kombiniere ich den umfangreichen Katalog von weltweit gesammelten *D. melanogaster* Laborstämmen mit qualitativ hochwertigen genomischen Assemblies (inklusive derjenigen welche in Abschnitt 1 etabliert wurden) und ausgiebigem genomischen Sequenzieren (DNA, RNA und small-RNA Sequenzierung), um die Dynamiken der Invasion von Tirant in einem Detailreichtum zu beschreiben, welches beispiellos für ähnliche Studien ist. Ich entdecke dass die Invasion von Tirant in *D. melanogaster* im frühen 20. Jahrhundert stattgefunden hat und sich innerhalb kurzer Zeit in den gesamten Vorkommensbereich der Spezies ausgebreitet hat. Ich zeige zusätzlich, dass Tirant bereits zuvor in *D. melanogaster* aktiv war, aber alle aktiven Kopien vor der neuen Invasion verloren wurden. Existierende Überbleibsel von piRNAs waren scheinbar nicht dazu in der Lage die neue Invasion von Tirant zu verhindern. Ich rekonstruiere ebenfalls die Muster bereits bekannter kürzlich erfolgter Invasionen von TEs in *D. melanogaster*, wodurch ich die Genauigkeit und Stärke meines Arbeitsprozesses bestätige. *D. melanogaster* ist einer der bestuntersuchten Organismen bezüglich seiner TE-Landschaft und die Existenz von Tirant in *D. melanogaster* ist bereits länger bekannt. Trotzdem blieb der Fakt das Tirant erst kürzlich in natürliche Populationen von *D. melanogaster* eindrang bisher unentdeckt. Diese bemerkenswerte Entdeckung zeigt auf dass die Anwendung eines ähnlichen Arbeitsprozesses die Möglichkeit der Entschlüsselung der bisher kryptischen Dynamiken von TE-Aktivität in verschiedensten Organismen bietet. Hierdurch könnte schlussendlich die umfangreiche Beschreibung der Aktivität verschiedenster TE-Familien sowie die quantitative Beschreibung der Dynamiken von TEs in natürlichen Populationen ermöglicht werden.

Zusammengenommen liefert diese Arbeit einen neuen Einblick in die biologischen Besonderheiten von TE Dynamiken und etabliert zeitgleich neue Methodiken welche zukünftige weiterführende Beschreibungen von TE Dynamiken in natürlichen Populationen ermöglichen. Hiermit lege ich den Grundstein für zukünftige Forschungsbemühungen, speziell im Rahmen der Beschreibung der Dynamiken von piRNA clustern und TE Invasionen

Summary

A significant portion of the genome of many organisms is comprised of Transposable Elements (TEs). TEs are complex interspersed repetitive genomic elements that selfishly proliferate in a host genome. Recent studies have shown that TE dynamics can play important roles in the adaptive evolution of organisms. However, many details about the dynamics of TEs in natural populations still remain elusive, including their degree of polymorphism as well as the underlying causes for observed differences in abundance and activity. In this thesis, I aim to contribute to the unraveling of TE dynamics in natural populations. Additionally, I aim to establish approaches and methodologies that equip future researchers with the tools to properly describe these dynamics.

In the first section, I develop and apply novel methodologies to assess the representation quality of TEs and repetitive regions in genomic assemblies. These methodologies aim to overcome the technological limitations the repetitive nature of TEs poses towards genomic characterization and correct representation of TEs and repetitive regions. Specifically, I aim to allow the study of functionally important genomic arrays of TE insertions involved in the piRNA defense pathway against uncontrolled transposition activity. These regions, termed 'piRNA clusters', are particularly difficult to study in genomic assemblies due to the repetitive nature of the inserted TEs. Additionally, currently available metrics to evaluate the quality of genome assemblies are largely not representative for the quality of TEs and repetitive regions. I address this problem by establishing various quality metrics indicative of the assembly quality of TEs and repetitive regions, mainly focusing on piRNA cluster sequences. I apply these quality metrics to create genome assemblies with high quality regarding their TEs and piRNA cluster. Additionally, I demonstrate how these metrics can be extended for broader applications in other genomic regions or organisms. Lastly, I envision how future studies could utilize this methodology to empirically describe the dynamics of piRNA clusters.

In the second section of my thesis I describe the recent, stealthy invasion of the LTR retrotransposon Tirant in *Drosophila melanogaster* within the last century. For this purpose, I combine the extensive catalogue of *D. melanogaster* laboratory strains collected worldwide with high-quality genome assemblies (including those created in Chapter 1) and extensive sequencing

efforts (DNA, RNA and small-RNA sequencing) to describe the dynamics of the Tirant invasion in detail unprecedented by similar studies. I find that Tirant has invaded *D. melanogaster* in the early 20th century, and quickly spread through the entire species range. I also demonstrate that Tirant was previously active in *D. melanogaster*, but active copies were lost previously to the recent invasion. Existing residual piRNAs were seemingly unable to prevent the novel invasion. Additionally, I reconstruct the activity patterns of previously described recent TE invasions in *D. melanogaster*, confirming the accuracy and power of my workflow. *D. melanogaster* is one of the most well-studied organisms in terms of its TE landscape and activity and the existence of Tirant in *D. melanogaster* was previously well known. However, the fact that Tirant has only recently invaded natural populations has eluded detection thus far. This remarkable discovery illustrates that the application of a similar workflow has the potential to unravel previously cryptic dynamics of TE activity in a variety of organisms. Thus, it could finally be possible to exhaustively describe the activity of the various existing TE families and allow to quantitatively and qualitatively describe the dynamics of TE activity in natural populations.

In summary, this thesis provides novel biological insight into peculiarities of TE dynamics and establishes novel methodological approaches that will allow to further describe TE dynamics in natural population. I thus lay the groundwork for future research, particularly within the framework of describing the dynamics of piRNA cluster dynamics and TE invasions.

Table of Contents

Abbreviations

TE: Transposable Element

LTR: Long Terminal Repeat

piRNA: PIWI-interacting RNA

TSB: Transposition-Selection Balance

HD: Hybrid Dysgenesis

Introduction

Transposable Elements are not just junk DNA

Understanding the genetic code has been a major goal in biology since the discovery of particulate inheritance (Mendel, 1866). Even before the emergence of the first DNA sequencing methods (Maxam and Gilbert, 1977; Sanger et al., 1977), researchers discovered that not all DNA is comprised of genes, which are considered the basic unit of heredity (Ohno, 1972). Surprisingly, the relative percentage of coding sequence in a genome can often be relatively low, e.g. only 2% in humans (Lander et al., 2001). The remaining genomic parts were originally termed 'junk DNA' and were suggested to not perform any important function for the organism (Ohno, 1972). Today, 'junk DNA' is usually referred to as 'noncoding DNA'. Noncoding DNA is used as a collective term for different types of sequences including regulatory elements, introns, pseudogenes and different types of repetitive elements (Shanmugam et al., 2017). As various types of noncoding DNA have been proven to perform crucial functions for the organism, the notion of absence of function for these genomic regions has been heavily challenged, (Maston et al., 2006; Dunham et al., 2012; Jo and Choi, 2015).

Repetitive sequences are one of the major classes of noncoding DNA and can be differentiated into tandem repeats, also referred to as satellite DNA, and interspersed repeats (López-Flores and Garrido-Ramos, 2012). The most prominent type of interspersed repeats are Transposable Elements (TEs). TEs are defined as stretches of DNA sequence that transpose within a host organisms DNA, i.e. increase their copy number (McClintock, 1956). To achieve transposition and subsequently an increased copy number, functional TEs encode specific proteins. The nature of the transposition mechanism mediated by these proteins is commonly used to differentiate between two major classes of TEs: Those where the transposition process includes an RNA intermediate (Class I transposon, retrotransposon or RNA transposon) or those that transpose exclusively as DNA without any RNA intermediates (Class II transposon or DNA transposon) (Finnegan, 1992; Wicker et al., 2007). Retrotransposons are commonly further differentiated into TEs possessing a long terminal repeat (LTR) (LTR retrotransposons) and TEs without an LTR (non-LTR retrotransposons). TEs can be even further classified in subclasses, orders, superfamilies, families and subfamilies, based on characteristics like specific target-site duplications,

proteins, expression, transposition and their genomic sequence (Wicker et al., 2007).

TEs are commonly found in most complex organisms including virtually all eukaryotes (Feschotte and Pritham, 2007), where they can comprise up to at least 85% of genomic sequence (Schnable et al., 2009; Mayer et al., 2012; Wegrzyn et al., 2013). Despite their genomic abundance, most TEs are not believed to confer fitness advantages to their host. Extensive TE activity or genomic TE abundance can even be detrimental for the respective hosts (Kazazian, 1998). Thus, TEs are commonly characterized as selfish elements, as they prioritize their proliferation over potential fitness consequences for the host organism. This led to the proposition that TEs are true junk DNA, as TEs do not contribute to any function in the host genome (Doolittle and Sapienza, 1980). However, this proposition has been challenged by the discovery of TE insertions with beneficial fitness effects for the host organism, suggesting that TEs are a driving force of novel genomic variation and adaptive evolution (Biémont and Vieira, 2006; Schrader and Schmitz, 2019). For example, the emergence of the dark-winged phenotype in the evolution of industrial melanism in the peppered moth (*Biston betularia*), a famous textbook example of rapid adaptive evolution, was recently demonstrated to be caused by the insertion of a TE into the first intron of the *cortex* gene (Hof et al., 2016). Other examples of TEs conferring a selective advantage include e.g. a mammalian neogene family (*Mart* genes) derived from LTR retrotransposons (Brandt et al., 2005), domesticated non-LTR retrotransposons forming telomeric sequences in *Drosophila* (Pardue and DeBaryshe, 2003) and various examples of TEs involved in changes in gene regulation (Medstrand et al., 2005). However, while examples of beneficial TE insertions exist, they are likely still rare compared to the vast majority of TE insertions with neutral or deleterious fitness effects (Platt et al., 2018; Arkhipova, 2018). Overall, research on TEs has shown them to be important genomic factors whose dynamics can strongly affect host organisms, for better or for worse. Understanding the dynamics of TEs in natural populations and the interactions of TEs with their host organisms is thus an important field of research with key importance for a better understanding of the processes shaping the genomes of organisms.

The repetitive nature of TEs impedes our understanding of TE dynamics

Historically, it was assumed that TE insertions are negatively selected and removed from the genome (Charlesworth and Charlesworth, 1983). Thus, a classical model termed 'transposition-selection balance' (TSB) assumed that the number of TE insertions in a genome is derived as a balance between transposition creating new copies at a certain, constant rate and negative selection simultaneously removing TE copies from the genome (Charlesworth and Charlesworth, 1983; Charlesworth and Langley, 1989; Barrón et al., 2014). Contrary to the TSB model, empirical work revealed that the activity of many TE families is often not continuous over time. Instead, genomic signatures indicate that TE activity often occurs in bursts, with a high degree of proliferation generating many insertions in a short amount of time, followed by a period of inactivity with hardly any TE activity (Bergman and Bensasson, 2007; Tsukahara et al., 2009; Kofler et al., 2012; Barrón et al., 2014). These observations can be partially explained by the discovery of host defense mechanisms actively suppressing TE activity, thus contradicting a simple TSB model (Lee and Langley, 2010; Blumenstiel, 2011). In animals and plants, these defense mechanisms prominently include DNA methylation, histone modifications and the production of specialized small RNAs (Urrutia, 2003; Rowe et al., 2010; Nuthikattu et al., 2013; Sigman and Slotkin, 2016). A prominent small-RNA based defense pathway relies on specialized small RNAs of a size range between 23-29 nucleotides. These small RNAs are called PIWI-interacting RNAs (piRNAs) and prevent TE activity on both a transcriptional as well as a posttranscriptional level (Brennecke et al., 2007, 2008). The production of piRNAs is believed to be dependent on the insertion of a TE within specific genomic regions termed piRNA cluster, which are usually large arrays of TE insertions close to heterochromatin (Aravin et al., 2007).

It is widely assumed that the establishment of TE silencing will stop the activity of a TE. However, it is unclear which forces determine the timeframe needed to successfully establish piRNA silencing and how variable these dynamics can be for different TE families. Similarly, the relative importance of piRNA clusters for the establishment or the maintenance of piRNA silencing is still debated. Additionally, underlying evolutionary dynamics, like the degree of conservation of piRNA clusters within populations and species or the evolutionary timeframe over which TE silencing is maintained, still remain largely elusive. A major hindrance preventing

researchers from unraveling the dynamics of TE activity, and particularly the dynamics of piRNA clusters, is the repetitive nature of TEs. Analyses of TEs with several copies in a genome are often limited as many technologies can not distinguish between TE copies. For example, the design of primer sequences for Polymerase Chain Reaction necessitates a unique genomic template, which is impossible for TEs with several identical copies. Similarly, analyses based on modern state-of-the-art short-read-sequencing techniques produce average read lengths of 50-250 basepairs. These read lengths are mostly insufficient to differentiate between insertions of a TE, since copies of most TEs regularly have lengths of a few kilobases (Sedlazeck et al., 2018). Consequently, genome assemblies based on short-read sequencing data usually misrepresent TEs (Alkan et al., 2011). This misrepresentation is especially pronounced for heterochromatic regions, as they often contain large arrays of repetitive sequence. The recent advent of long-read sequencing technologies routinely producing reads longer than most TEs has elevated this issue. However, even in classic high-quality reference genomes and modern genome assemblies produced by long-read sequencing technologies, many heterochromatic, repetitive regions like piRNA clusters are still incomplete. Also, most widely used quality metrics employed to assess the quality of a genome assembly like the BUSCO (Benchmarking-Universal-Single-Copy-Orthologs) tool (Seppey et al., 2019) focus mainly on the representation of euchromatic regions while disregarding the representation of TEs and repetitive regions. Improving currently available genome assemblies to reliably assess the population dynamics of TEs thus still remains a major task to enable the detailed characterization of TE dynamics in populations and species.

Detailed descriptions of TE dynamics in natural populations are scarce

Regardless of their genomic abundance, many TEs are not precisely understood on a molecular level. For example, TE abundance, the amount of active TEs and the intensity of TE activity can all vary drastically between species (e.g. (Lee and Kim, 2014)). The total percentage of TEs within plants varies between 3% in *Utricularia gibba* (Ibarra-Laclette et al., 2013) and 85% in maize (Schnable et al., 2009). Similar trends can be observed for animals, where even closely related species show drastic differences in their genomic TE content (Blommaert et al., 2019; Wong et al., 2019). These differences in TE content are often directly correlated

with differences in genome size observed between these organisms. Thus, variation of TE content is likely a prominent contributor to the still unexplained drastic variation in genome sizes between organisms known as the C-value enigma (Moore, 1984; Canapa et al., 2015). Within a genome, larger arrays of TE insertions are usually restricted to the functionally less important heterochromatin and its peripheries (Charlesworth, 1991; Bartolomé et al., 2002). Contrarily, TE insertions in the functionally more important euchromatin are mostly singular and often stem from recent TE activity (Kaminker et al., 2002). Generally, the total amount of TEs in a genome does not necessarily correlate with the amount of active TEs, i.e. copies that retained their transposition capabilities. In humans, at least 45% of the genome is constituted of TEs (Lander et al., 2001). Yet, out of at least 800 described TE families residing in the human genome (Bao et al., 2015), only four families have retained transposition capabilities (Mills et al., 2007). Contrarily, TEs are less abundant but more active in insect species like the vinegar fly *Drosophila melanogaster*. In *D. melanogaster*, only 20% of the genome is constituted by $\sim$120 TE families (Kaminker et al., 2002; Bao et al., 2015). However, around 30% of *D. melanogaster* TE insertions consisting of most of the $\sim$120 TE families are likely still potentially active (Barrón et al., 2014; Kofler et al., 2015). It is still not entirely clear which evolutionary forces underlay such interspecific discrepancies in TE content and activity. Additionally, it is unclear how TE dynamics vary for the different TE classes families and subfamilies. To determine the underlying factors of these dynamics and derive general expectations and models for TE activity, detailed molecular characterizations of the activity of TEs in natural populations are needed.

In few organisms the TE diversity and activity, the defense mechanisms against transposition and the phenotypic consequences of TE mobilization are as exceptionally well understood as in *D. melanogaster* (Engels, 1983; Kaminker et al., 2002; Barrón et al., 2014; Kofler et al., 2015; Jakšić et al., 2017; McCullers and Steiniger, 2017). Historicaly, one of the first descriptions of phenotypic consequences of TE activity, the discovery of the Hybrid Dysgenesis (HD) systems (Bucheton et al., 1976; Kidwell et al., 1977; Blackman et al., 1987; Yannopoulos et al., 1987), was made in *D. melanogaster*. HD is an age- and environmentally-dependent (Bucheton, 1979) nonreciprocal sterility syndrome in specific crosses of *D. melanogaster*, resulting in a variety of detrimental phenotypes, e.g. sterility of F1 females due to underdeveloped gonads or drastically reduced hatching rates of F2 embryos, caused by uncontrolled TE proliferation. HD is considered

one of the first direct observations as well as one of the most prominent examples of functional consequences of TE activity. Three TEs are known to induce HD in *D. melanogaster*, the DNA transposons P-element and hobo and the non-LTR retrotransposon I-element. All three TEs are exceptionally well characterized in terms of their molecular properties (e.g. temperature-dependent activity (Bucheton, 1979) or sequence variants like the repressor element KP, which can influence the degree of P-element activity (Black et al., 1987; Lee et al., 1996; Ruiz and Carareto, 2003) as well as their evolutionary history of activity in space and time (e.g. if they are vertically or horizontally transmitted or if they had different waves of activity at distinct points in time) (Kidwell, 1983; Daniels et al., 1990a,b; Bucheton et al., 1992).

However, these TEs only represent three families out of more than hundred active TEs in *D. melanogaster*. For most TEs, even in well-studied model organisms like *D. melanogaster*, the evolutionary history and molecular mechanisms are not nearly as well described. Thus, it is yet unclear if and to what degree these different TE families show variations in properties like i) how the TE is transmitted (vertical or horizontal), ii) the degree and mode of activity (continuous or burst-like), iii) the time span of the TEs activity, iv) how host organisms defend against their activity and v) how/if host defense is maintained over extended periods of time. It is a major goalpost of TE research to characterize the natural variation of these properties among TEs and ultimately shed light on what determines these underlying evolutionary dynamics of TEs in these natural populations.

Aim of my thesis

The aim of my thesis is to improve the current understanding of TE dynamics in natural populations. The thesis is divided into two main sections, where each section contains one manuscript.

In the first part, I aim to pave the way to allow researchers to overcome technical limitations preventing comprehensive analyses of the genomic representation of TEs. I particularly focus on the representation of arrays of repetitive sequences like piRNA clusters within whole-genome assemblies. I develop experimental and computational tools assessing this quality of representation. Most importantly, I establish quality metrics to globally and/or locally assess

the assembly quality of TEs and piRNA clusters. I apply these metrics to establish an idealized assembly pipeline for *D. melanogaster* and to assess the quality of existing assemblies. Utilizing these metrics, I aim to establish a comprehensive data set of genomic assemblies to perform meaningful inferences of polymorphisms in TEs and piRNA clusters between populations. Further, I demonstrate that my methods developed in this work are broadly applicable and can be usefully extended to different species as well as different regions than piRNA clusters. Lastly, I demonstrate the usefulness of this approach by demonstrating how it allows the empirical testing of model predictions regarding the mechanisms of piRNA-mediated host defense utilizing piRNA cluster.

In the second part, I aim to characterize the evolutionary history and molecular mechanisms of the recent invasion of the LTR retrotransposon Tirant in *D. melanogaster*. I first discover the invasion due to drastic abundance differences of Tirant within two laboratory strains descending from natural populations sampled at different points in time. To test the hypothesis that Tirant invaded *D. melanogaster* populations in the timeframe between the sampling of the ancestors of these two laboratory strains, I establish a novel workflow with the aim to describe the detailed dynamics of Tirant over space and time. I aim to describe the exact dynamics of this invasion by determining i) when canonical (i.e. 'new') Tirant arrived in *D. melanogaster* populations, ii) how canonical Tirant spread during its invasion, iii) the abundance of canonical Tirant in extant populations, iv) the abundance and role of old, fragmented Tirant insertions, v) the temporal comparison with the other recent TE invasions in *D. melanogaster* (P-element, I-element, hobo), vi) the abundance and silencing properties of piRNAs derived from old vs canonical insertions and vii) if I can detect signatures that Tirant was introduced by horizontal transfer from a closely related species. I test the effectiveness of this approach by reconstructing and comparing previous findings about the other recent TE invasions in *D. melanogaster* and simultaneously establish a state-of-the-art workflow to describe the dynamics of a TE. The detection of this recent, stealthy invasion in the highly studied *D. melanogaster* strongly highlights the potential such workflows possess for the detailed study of TE dynamics in natural populations.

Results

Chapter 1

Received: 15 December 2020 | Revised: 30 April 2021 | Accepted: 14 June 2021

DOI: 10.1111/1755-0998.13455

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Novel quality metrics allow identifying and generating high-quality assemblies of piRNA clusters

Filip Wierzbicki[1,2] | Florian Schwarz[1,2] | Odontsetseg Cannalonga[1] | Robert Kofler[1]

[1]Institut für Populationsgenetik, Vetmeduni Vienna, Wien, Austria

[2]Vienna Graduate School of Population Genetics, Vetmeduni Vienna, Vienna, Austria

Correspondence
Robert Kofler, Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, 1210 Wien, Austria.
Email: rokofer@gmail.com

Funding information
Austrian Science Fund, Grant/Award Number: P30036-B25 and W1225

Abstract

In most animals, it is thought that the proliferation of a transposable element (TE) is stopped when the TE jumps into a piRNA cluster. Despite this central importance, little is known about the composition and the evolutionary dynamics of piRNA clusters. This is largely because piRNA clusters are notoriously difficult to assemble as they are frequently composed of highly repetitive DNA. With long reads, we may finally be able to obtain reliable assemblies of piRNA clusters. Unfortunately, it is unclear how to generate and identify the best assemblies, as many assembly strategies exist and standard quality metrics are ignorant of TEs. To address these problems, we introduce several novel quality metrics that assess: (a) the fraction of completely assembled piRNA clusters, (b) the quality of the assembled clusters and (c) whether an assembly captures the overall TE landscape of an organisms (i.e. the abundance, the number of SNPs and internal deletions of all TE families). The requirements for computing these metrics vary, ranging from annotations of piRNA clusters to consensus sequences of TEs and genomic sequencing data. Using these novel metrics, we evaluate the effect of assembly algorithm, polishing, read length, coverage, residual polymorphisms and finally identify strategies that yield reliable assemblies of piRNA clusters. Based on an optimized approach, we provide assemblies for the two *Drosophila melanogaster* strains Canton-S and Pi2. About 80% of known piRNA clusters were assembled in both strains. Finally, we demonstrate the generality of our approach by extending our metrics to humans and *Arabidopsis thaliana*.

KEYWORDS
Drosophila melanogaster, genome assembly, Oxford Nanopore sequencing, piRNA clusters, transposable elements

1 | INTRODUCTION

Transposable elements (TEs) are short stretches of DNA that proliferate within genomes, even if this activity reduces the fitness of the hosts (Hickey, 1982). An unconstrained proliferation of TEs could lead to an accumulation of deleterious TE insertions that may eventually drive host populations to extinction (Brookfield & Badge, 1997; Kofler, 2020). Hence, host organisms evolved elaborate mechanisms

Filip Wierzbicki and Florian Schwarz contributed equally to this work.

to combat the spread of TEs (Brennecke et al., 2007; Lewis et al., 2018; Yang et al., 2017). In mammals and invertebrates, the host defence against TEs is based on piRNAs, that is small RNAs with a size between 23 and 29 nt (Brennecke et al., 2007; Gunawardane et al., 2007). These piRNAs bind to PIWI-clade proteins that silence TEs at the transcriptional as well as the post-transcriptional level (Brennecke et al., 2007; Gunawardane et al., 2007; Le Thomas et al., 2013; Sienski et al., 2012). piRNAs are derived from discrete genomic loci, termed piRNA clusters, which may make up substantial portions of genomes (e.g. 3.5% in *D. melanogaster*). piRNA clusters play a central role in the defence against TE invasions (Bergman et al., 2006; Zanni et al., 2013). Under the currently prevailing view, the trap model, it is assumed that a newly invading TE is stopped when a copy of the TE jumps into a piRNA cluster, which triggers the production of piRNAs that silence the TE (Bergman et al., 2006; Duc et al., 2019; Goriaux et al., 2014; Malone & Hannon, 2009; Ozata et al., 2019; Yamanaka et al., 2014; Zanni et al., 2013). Despite the central importance of piRNA clusters in the defence against TEs, the composition and evolution of these regions remains poorly understood. A better understanding of these regions could shed light on important open questions in TE biology, like whether or not the trap model holds (Kofler, 2019, 2020; Mohamed et al., 2020). Our lack of knowledge comes mostly from the fact that piRNA clusters are notoriously difficult to assemble. Most piRNA clusters are located in the heterochromatin and consist of highly repetitive sequences such as TEs (Asif-Laidin et al., 2017; Brennecke et al., 2007; Zanni et al., 2013). Long-read sequencing (e.g. by Pacific Biosciences or Oxford Nanopore Technology) promises to close this gap in our understanding by enabling us to obtain complete assemblies of piRNA clusters. However, it is currently not clear which assembly strategies yield reliable assemblies of piRNA clusters, since many different assembly tools, polishing strategies, sequencing data and scaffolding approaches may be used. In fact, it is not even clear on how to identify the best assemblies, as classic quality metrics such as BUSCO and NG50 are ignorant of TEs and piRNA clusters: BUSCO (Benchmarking Universal Single-Copy Orthologs) provides the fraction of correctly assembled (i.e. 'complete') core genes (Simão et al., 2015; Waterhouse et al., 2018) and NG50 gives the size of the smallest contig out of the largest contigs that account for 50% of the reference genome (Earl et al., 2011).

Here, we address these challenges by first introducing several novel quality metrics that assess the number and the quality of the assembled piRNA clusters. Our novel quality metrics were then used to evaluate the effect of different assembly algorithms, polishing approaches, read lengths, coverages, levels of residual polymorphisms and scaffolding methods. Based on these results, we identify strategies that generate high-quality assemblies of piRNA clusters. Using such an optimized assembly strategy, we provide novel assemblies for the *Drosophila melanogaster* strains Canton-S and Pi2. Additionally, we demonstrate the generality of our approach by extending our metrics to humans and *A. thaliana*. We provide a user-friendly pipeline, a manual and a walkthrough for assessing the quality of assembled piRNA clusters.

2 | MATERIALS AND METHODS

2.1 | Sequencing

The *D. melanogaster* strains Canton-S and Pi2 were obtained from the Bloomington Drosophila Stock Center (BDSC) (Canton-S = 64349; Pi2 = 2384). The reference strain, Iso-1, was kindly provided by Dr. K.A. Senti. We performed Oxford Nanopore Sequencing, Illumina paired-end sequencing and Hi-C for Canton-S and Pi2 (Table S1). The strain Iso-1 was solely sequenced using the Illumina paired-end technology.

High molecular weight DNA for Oxford Nanopore sequencing was extracted from whole bodies of 50 female virgin flies using the Phenol-Chloroform extraction protocol described by Maniatis et al. (1982) using slightly elongated incubation times (5 min). The DNA was sheared to a mean fragment length of 20–30 kb with Covaris g-TUBEs (Covaris Inc., Woburn, MA, USA). The length of the DNA was measured with a TapeStation (4200; DNA ScreenTape, Agilent Technologies). Library preparation was performed with an input of 2–5 μg of sheared DNA following the manufacturer's protocol (kit LSK108; Oxford Nanopore Technologies; Oxford). About 1–2 μg of the libraries was run for 48–72 hours on MIN106 flow cells. The DNA concentration was measured with a Qubit fluorometer (broad-range DNA assay) (Thermo Fisher Scientific, Waltham, MA, USA), and the purity of the DNA was controlled with NanoDrop (Thermo Fisher Scientific, Waltham, MA, USA).

DNA for Illumina paired-end sequencing was extracted from whole bodies of 20–30 virgin female flies using a salt-extraction protocol (Maniatis et al., 1982). Libraries were prepared with the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs, Ipswich, MA, USA) using 1 μg DNA. Illumina sequencing was performed by the Vienna Biocenter Core Facilities on a HiSeq2500 platform (2 × 125 bp; Illumina, San Diego, CA, USA).

Hi-C was performed following the Phase Genomics Proximo Hi-C animal Kit (Phase Genomics, Seattle, WA). About 40–50 female third instar larvae were sliced with a razor blade to obtain about 80 mg of tissue. Crosslinking and library preparation were performed according to instructions. Sequencing was performed by the Vienna Biocenter Core Facilities NGS on an Illumina HiSeq2500 platform (2 × 125 bp; Illumina, San Diego, CA, USA).

2.2 | Assemblies

Short-read assemblies with Illumina paired-end reads (read length 125 and mean coverage of 30×) were performed with ABYSS (Simpson et al., 2009) (version 2.1.5; abyss-pe) using a k-mer size of 96.

Base calling of raw nanopore reads (fast5 format) was performed with either ALBACORE (version 2.3.4; Oxford Nanopore Technologies, Oxford, GB) or GUPPY (version 2.1.3; Oxford Nanopore Technologies, Oxford, GB). Summary statistics, including mean read length and the total output, were calculated with NANOPLOT (De Coster et al., 2018) (version 1.20.1).

De novo assembly of the nanopore reads was performed with four different tools: CANU (Koren et al., 2017) (version 1.7), MINIASM (Li, 2016) (version 0.3-r179), WTDBG2 (Ruan & Li, 2020) (version 2.4) and FLYE (Kolmogorov et al., 2019) (version 2.8-b1674). With CANU, raw nanopore reads were corrected and trimmed prior to the assembly (preset --nanopore-corrected). To generate assemblies with miniasm, we first aligned all reads against themselves with MINIMAP2 (Li, 2018) (version 2.16-r922) using a preset for nanopore reads (-x ava-ont). We generated the assemblies with MINIASM using default settings. The resulting assembly graph files (gfa) were transformed into fasta-files with awk. We launched WTDBG2 with the raw nanopore reads and a nanopore-specific preset ('preset2'). FLYE was launched with the raw nanopore reads with the corresponding option (--nano-raw) and default parameters.

Polishing of long-read assemblies was carried out in two steps. We first used RACON (Vaser et al., 2017) (version 1.2.1) with the raw nanopore reads mapped to the assembly (MINIMAP2; -ax map-ont; version 2.16-r922 Li, 2018) and then PILON (Walker et al., 2014) (version 1.22) with Illumina paired-end reads mapped to the assembly (BWA MEM (Li & Durbin, 2009) (version 0.7.17-r1188). The optimal number of polishing iterations was chosen based on the maximally achieved BUSCO (Benchmarking Universal Single-Copy Orthologs) values (Table S2).

Scaffolding of contigs was done with Hi-C following the SALSA2 protocol (Ghurye et al., 2019) (version 30. Nov.2018). Briefly, Hi-C reads were mapped to the assembly with BWA BWASW (Li & Durbin, 2009) (version 0.7.17-r1188), filtered (https://github.com/Arima Genomics), and duplicates were removed (PICARD-TOOLS; version 2.18.23; https://broadinstitute.github.io/picard/). The mapped reads were then used for scaffolding with SALSA2 using the parameters: diploid mode (-m yes) and restriction enzyme sequence (GATC). An assembly graph was provided. Reference-guided scaffolding was performed with RAGOO (Alonge et al., 2019) (version 1.1) based on release 6 of the *D. melanogaster* reference genome (Hoskins et al., 2015).

Random sampling of reads was performed with SEQTK (https://github.com/lh3/seqtk) (version 1.3-r106). To obtain subsets of the longest reads, we sorted all reads by length and then used the appropriate number of the first reads (i.e. the longest reads). Polishing of assemblies generated with subsets of reads was carried out with the respective subsets.

For an overview of our assembly pipeline, see Figure S1.

To visualize assemblies, we generated dotplots using NUCMER (Kurtz et al., 2004) (version 3.1). We aligned assemblies to the main chromosome arms (X, 2L, 2R, 3L, 3R and 4) of the *D. melanogaster* reference ('mumreference'; with parameters -c 1000 -l 100), created coordinate index files using DotPrep.py and visualized genome alignments with DOT (https://dnanexus.github.io/dot/).

The final assemblies were based on CANU using 100× of the longest reads. Misassemblies were identified based on Hi-C heatmaps and alignments of the assemblies to the reference genome (dotplots). Hi-C heatmaps were generated with JUICER (Durand et al., 2016) (version 1.7.6) using 'Sau3AI' as the restriction enzyme. Heatmaps were visualized and analysed with JUICEBOX (1.11.08). Potential misassemblies identified in the Hi-C heatmaps were cross-validated with long reads that were aligned to the assemblies. Breaks in the alignment of the long reads were interpreted as support of an assembly error. Contigs with misassemblies were broken with a custom script 'introduceBreaks2fasta.py'. Potential contamination (e.g. adaptor sequences) was removed using the standard tools implemented by NCBI.

2.3 | Quality of assemblies

BUSCO (Waterhouse et al., 2018) (version 3.0.2) values were based on the diptera _odb9 data set (2799 genes). QUAST (Gurevich et al., 2013) (version 5.0.2; quast-lg) was used to compute basic assembly statistics such as NG50 and the total assembly length. As reference, we used the genome of *D. melanogaster* (release 6).

Computing our TE landscape metrics (abundance of TEs, number of SNPs and internal deletions (IDs) within TEs) requires Illumina raw sequencing reads (expectations) and artificial reads generated from an assembly of interest (observations). We generated artificial reads of length 125 bp starting at each position of the assembly (yielding a uniform distribution; artificial-reads-for-assembly.py). The abundance of TEs, as well as the number of SNPs and internal deletions within TEs, was estimated with DEVIATE (Weilguny & Kofler, 2019) (version 0.3.6) to obtain both the expected values (Illumina raw reads) and the observed values (artificial reads derived from the assembly). As reference library for DEVIATE, we used the consensus sequences of TE families present in *D. melanogaster* (v9.42; we added the sequence of Mariner: M14653) (Quesneville et al., 2005). Solely SNPs and internal deletions with a minimum frequency of 2% were considered. The GC content for each TE was calculated via a custom script ('GC-content-calculator.py').

The CUSCO metric relies on the annotation of piRNA clusters of *D. melanogaster* release 5 (Brennecke et al., 2007; Hoskins et al., 2007). From the 142 annotated piRNA clusters, we excluded clusters that were annotated at the ends of chromosomes (10) and on the highly fragmented U-chromosome (46) (as flanking sequences can not be obtained for these clusters). For the remaining 86 clusters, we identified sequences flanking the clusters at both ends. These flanking sequences were required to align uniquely to release 6 of the *D. melanogaster* reference genome. For two piRNA clusters that were adjacent to each other (cluster 8 and 9), we could only obtain a pair of sequences flanking both clusters. In summary, we were able to design flanking sequences for 85 piRNA clusters. These sequences had a size between 49 and 12,567 nucleotides. To compute the CUSCO, the flanking sequences were aligned to an assembly using BWA MEM (Li & Durbin, 2009). The CUSCO was computed with the script 'cusco.py' as the fraction of complete piRNA clusters (i.e. both flanking sequences aligned to the same contig/scaffold). We furthermore distinguished between an ungapped-CUSCO and a gapped-CUSCO based on the presence of poly-N sequences between the two sequences flanking a piRNA cluster. Poly-N tracts in

assemblies were identified using the script 'find-polyN.py'. To determine whether piRNA clusters are uniquely assembled, we tested if both flanks mapped to multiple contigs/scaffolds using the script 'multi-cluster.py'.

To identify assembly errors in piRNA clusters, we aligned long reads to assemblies using MINIMAP2 (Li, 2018) (version 2.16-r922). A list of complete BUSCO genes was obtained from the BUSCO pipeline (diptera_odb9; 2799 genes). Based on these data, we computed the base coverage and the soft-clip coverage along each piRNA cluster as well as the 99% quantiles of these coverages (quantiles.py). To calculate the base-coverage quality (CQ), we divided the standard deviation of base coverage in a cluster by the median of standard deviations of complete BUSCO genes. The base coverage and the CQ values were computed with the script 'cluster-coverage-median.py' and the parameters --min-mq 15 and --min-len 1000. To calculate the soft-clip quality (ScQ), we divided the average number of soft clipped base pairs in a cluster by the median of the average numbers of soft-clipped bases in complete BUSCO genes. The soft-clip coverage and the ScQ values were computed with the script 'cluster-softclipcoverage-median.py' and the parameters --min-mq 15 and --min-len 1000. The script 'visualize.R' was used to visualize the base coverage, the soft-clip coverage, the coverage quantiles and locations of assembly gaps (i.e. poly-N sequences) for the piRNA clusters.

2.4 | PCR validation

PCRs were performed at a volume of 20 µl, with 0.05 U/µl of Firepol polymerase (Solis BioDyne, Tartu, Estonia), 2.5 mM MgCl2, 200 µM dNTPs, 0.2 µM primer and 100 ng/µl DNA. See Table S3 for all primer pairs. We used a PCR cycler (Bio Rad CFX Connect, Hercules, CA, USA) with the following program: 5-min. denaturation at 94; 30 cycles of denaturation (30 s at 94), annealing (1 min at 58) and elongation (1 min at 72), followed by 10 min of final extension at 72. The PCR products were loaded on a 1% agarose gel and ran with 120 V for 30 min in TBE buffer. The expected length of amplicons was inferred from the assemblies. Only polymorphic TE insertions for which both breakpoints agreed with the expectations were assumed to be successfully validated.

2.5 | Data analysis

To identify heterozygous SNPs, we aligned Illumina paired-end reads to release 6 of the *D. melanogaster* genome with bwa mem using default parameters. Reads with a low mapping quality were removed using SAMTOOLS (version 1.7; Li et al., 2009), a mpileup file was created (SAMTOOLS), and allele frequency estimates were obtained using mpileup2sync (POPOOLATION2; Kofler, Pandey, et al., 2011) with the parameters --fastq-type sanger --min-qual 20. The fraction of heterozygous SNPs for windows of 100 kb was computed with a custom script (polymorphicSNPs_from_sync.py). To account for sequencing errors, we solely classified SNPs with allele frequencies between 0.25 and 0.75 as segregating. Furthermore, a minimum coverage of 10 was required for each site. Windows with insufficient coverage at more than 25% of the sites were excluded. Finally, solely windows with sufficient coverage in all three samples (Pi2, Canton-S and Iso-1) were retained.

To identify the redundant contigs, we chopped assemblies into nonoverlapping fragments of 1 kb using a custom script (chopgenome.py) and aligned them to the release 6 of the *D. melanogaster* genome using BWA BWASW with default parameters (version 0.7.17-r1188; Li & Durbin, 2009). Ambiguously mapped reads were filtered with SAMTOOLS (-q 20), and a mpileup file was generated. The mean coverage for 100 kb windows was calculated using a custom script (coverage_from_pileup.py).

We used SNIFFLES (version 1.0.7; Sedlazeck, Rescheneder, et al., 2018) to identify structural variants (SVs). Such SVs may either be present or absent in the assembly (classified as deletion and insertion, respectively). We first mapped the long reads to assemblies using NGMLR (v0.2.7; Sedlazeck, Rescheneder, et al., 2018) with the parameter -x ont (ONT data as input). SVs were identified with SNIFFLES using the parameters --report_seq (obtain the sequence of SVs) and --genotype (report allele frequency estimates of SVs). The resulting vcf-file was filtered for SVs with a minimum length of 1kb. To obtain heterozygous SVs, we filtered for allele frequencies between 25% and 75%. To identify SVs caused by TEs, we aligned the sequences of SVs to the consensus sequences of TEs (Quesneville et al., 2005) using BLASTN (2.7.1+, Altschul et al., 1990).

The composition of piRNA clusters was visualized with EASYFIG (v2.2.3 08.11.2016; Sullivan et al., 2011). Annotations of TEs were obtained with REPEATMASKER (open-4.0.7; Smit et al., 2015) using the parameters: -no_is (skip checking for bacterial insertions), -nolow (skip masking low complexity regions) and *D. melanogaster* TE sequences (Quesneville et al., 2005) or *Drosophila* TE sequences (Bao et al., 2015). Synteny within piRNA clusters among the assemblies was identified with BLASTN (2.7.1+, Altschul et al., 1990). To avoid cluttering of the figure, we removed annotations of TEs smaller than 1 kb and BLASTN similarity blocks smaller than 3 kb. For *D. melanogaster*, we used assemblies from NCBI with following accession numbers: SIXD01000000 and SISJ02000000 (Ellison & Cao, 2020); bioproject PRJNA418342 (Chakraborty et al., 2019); GCA_002310755.1 and GCA_002310775.1 (Anreiter et al., 2017); JXOZ01000000 (Vicoso & Bachtrog, 2015); LYTF01000000 (Singhal et al., 2017); and JAQD01000000 (McCoy et al., 2014). All statistical analyses were done with R (version 3.4.3) (R Core Team, 2012), and visualizations were performed using the GGPLOT2 library (Wickham, 2016).

2.6 | Application in different species

To calculate the TE landscape metrics for humans, we used the repetitive sequences library for humans from RepBase (Bao et al., 2015) (version 23.10, humrep.ref) containing 1063 sequences. We

compared a short-read and a long-read-based assembly derived from the same individual (KOREF) (Cho et al., 2016; Kim et al., 2019). To obtain 'expected' values, short-read sequences of the KOREF individual were used (SRR2204705) (Cho et al., 2016). To obtain the 'observed' values, we created artificial reads for both the short-read (KOREF1.0 (Cho et al., 2016)) and the long-read assembly (KOREF PB_62x (Hi-C scaffolded) (Kim et al., 2019)). The landscape metrics were computed as described before.

To establish flanking sequences for piRNA clusters in humans, we used annotations of 168 piRNA clusters (Sarkar et al., 2014) in the human reference genome hg19. Flanking sequences were created using the scripts 'flankbeder.sh' and 'flankparser.sh'. For each piRNA cluster, the 5 kb regions flanking each cluster were first split into five regions of 1 kb each. Potential flanking sequences containing N's were removed, and the remaining sequences were aligned back to hg19 using BWA BWASW (version 0.7.17-r1188; Li & Durbin, 2009). The potential flanking sequences were required to align back to the origin (with a tolerance of 5 kb) with a minimum mapping quality (mq) of 5. For each piRNA cluster, the most proximal pair of flanking sequence meeting these criteria was retained (136 out of 168). To calculate the CUSCO values, the flanking sequences were mapped to the respective assemblies using BWA BWASW and CUSCO was calculated as described before.

We calculated CUSCO for 11 assemblies of humans (GRCh37 GCA_000001405.1 (Church et al., 2011); GRCh38.p13 GCA_000001405.28 (Schneider et al., 2017); T2T GCA_009914755.2 (Miga et al., 2020); HG00733_Phased_Diploid GCA_003634875.1; HG00514_prelim_3.0 GCA_002180035.3; Ash1.7 GCA_011064465.1 (Shumate et al., 2020); KOREF1.0 GCA_001712695.1 (Cho et al., 2016); and Hi-C scaffolded long-read assemblies of KOREF PB_30x, PB_62x, PT_27x, PT_64x (Kim et al., 2019)).

To calculate CQ and ScQ, long reads (SRR9591076) of the KOREF individual were mapped to a short- and long-read assembly (KOREF1.0 (Cho et al., 2016) and KOREF PB_62x (Hi-C scaffolded) (Kim et al., 2019) with MINIMAP2 (Li, 2018) (version 2.17-r941), using a preset for PacBio reads (-ax map-pb). Calculations of CQ and ScQ values were performed as described previously. For all human genomes, BUSCO (Seppey et al., 2019) (version 5.1.2) was computed using vertebrata_odb10.

To identify pairs of sequences flanking *KEE* (*KNOT ENGAGED ELEMENT*) regions in *A. thaliana*, we used the annotations of the 10 *KEE* regions (Grob et al., 2014) and the reference genome TAIR10. Design of flanking sequences and calculation of CUSCO were performed as described for humans. For all *A. thaliana* assemblies, BUSCO (version 3.0.2) was computed using embryophyta_odb10. We calculated CUSCO for eight different assemblies (TAIR10 GCA_000001735.1 (Lamesch et al., 2012); AthNd1_v1.0 GCA_001742845.1 (Pucker et al., 2016); AT9943.Cdm-0.scaffold GCA_904420315.1; AT1741.KBS-Mac-74. PacBio GCA_903064285.1; Arabidopsis_thaliana_Ler GCA_902460285.1 (Berardini et al., 2015); ONTmin_IT4 GCA_900303355.1; Ler Assembly GCA_001651475.1 (Zapata et al., 2016); and ASM83594v1 GCA_000835945.1 (Berlin et al., 2015)).

3 | RESULTS

3.1 | Assembly quality of piRNA clusters

Here, we aim to identify strategies that enable us to generate high-quality assemblies of piRNA clusters. Since commonly used assembly quality metrics, such as BUSCO and NG50 (Earl et al., 2011; Simão et al., 2015), are ignorant of TEs, we first developed several novel quality metrics. Our novel metrics assess whether assemblies (a) accurately reproduce the abundance and diversity of TEs (i.e. the TE landscape) of an organisms, (b) have complete piRNA clusters and (c) contain assembly errors within piRNA clusters.

To obtain a data set for demonstrating our novel metrics, we sequenced the *D. melanogaster* strain Canton-S with (a) the Oxford Nanopore long-read technology (coverage 150×, mean read length ≈ 7 kb), (b) Illumina paired-end sequencing (coverage 30×, read length 125 bp) and (c) Hi-C (coverage 530×, read length 125 bp) (Table S1).

With our first metrics, we tested whether an assembly accurately reproduces the TE content of an organism. A good representation of the TE composition is an important quality control of assemblies and a requirement for an accurate assembly of highly repetitive regions such as piRNA clusters. With these new metrics, we do however not estimate whether TE insertion sites are correct, as this would require knowledge of the true insertion sites in an organism. Instead, we infer summary statistics of the TE landscape by measuring three different features for each TE family: (a) the abundance (in reads per million: rpm), (b) the number of SNPs and (c) the number of internal deletions. A comparison of expected and observed values for these three features allows to estimate the quality of TE representation in an assembly (Figure 1). The key idea is that the expected TE landscape can be directly inferred from the Illumina raw reads without prior need to generate an assembly. We estimate the expected TE landscape with DEVIATE (Weilguny & Kofler, 2019), which aligns Illumina reads to the consensus sequences of TEs and provides estimates of the abundance (rpm) and diversity (SNPs and IDs) of each TE family (Figure 1a; Figure S2). For an assembly of interest, we compute the observed TE landscape by generating artificial reads using the assembly as template, which are then used with DEVIATE to estimate abundance and diversity of TEs (Figure 1a, Figure S3). To avoid biases and sampling noise, these artificial reads should be uniformly distributed across the assembly and have the same length as the Illumina raw reads used for inferring the expected TE landscape.

To summarize the representation of TEs across all TE families (e.g. 127 TE families in *D. melanogaster*), we perform a linear regression between the expected and the observed values (Figure 1b; Figure S4). We propose to use the slope of each regression line as a novel quality metric (Figure 1b; Figure S4). This yields, in total, three novel quality metrics (slope of abundance, SNP count and ID count) that estimate how well an assembly captures the TE landscape. High-quality assemblies that accurately reproduce the TE landscape will have regression slopes of ≈1.0 for each of the three features. Assemblies that overestimate the TE abundance will have a slope

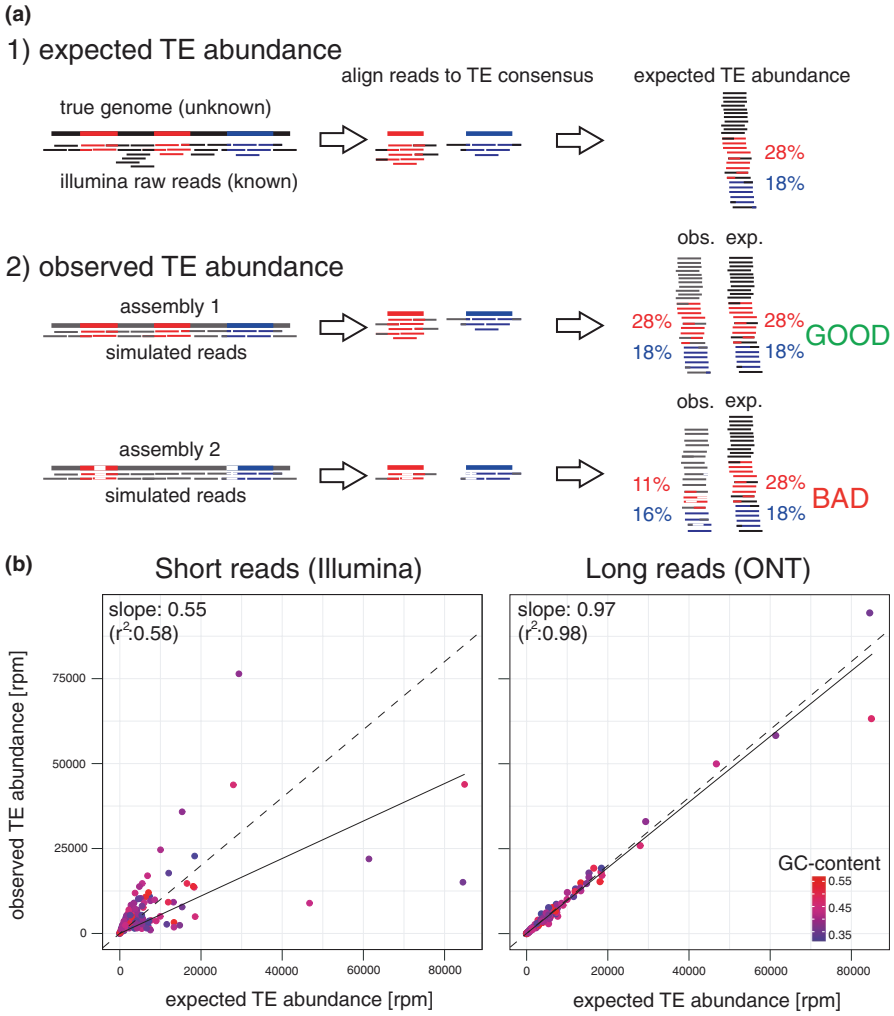

FIGURE 1 With three novel metrics, we assess how well an assembly captures the TE landscape of an organism, that is the abundance of TEs as well as the number of SNPs and internal deletions (IDs) within TEs. (a) Our metrics are based on a comparison between the expected and the observed TE landscape. We illustrate these metrics by the example of the TE abundance. The expected TE abundance (exp.) is derived by aligning Illumina raw reads to consensus sequences of TEs and counting the fraction of reads mapping to each TE family. Different TE families are shown in red and blue. The observed TE abundance (obs.) is derived by generating artificial reads from assemblies of interest, aligning these reads to the consensus sequence of TEs and counting the reads. A high-quality assembly will capture the TE abundance more accurately (good) than a low-quality assembly, for example having several assembly gaps (bad). (b) To summarize these results across all TE families, we perform a regression between the expected and the observed TE abundance. The slope of the regression represents our novel quality metric for the TE abundance. Results are shown for a short- and long-read assembly of Canton-S (30× coverage for both). Each dot represents a distinct TE family, and the dashed line shows the optimal representation of the TE abundance. Note that the long-read assembly captures the TE abundance more accurately than the short-read assembly (despite expectations being based on short reads). Similarly to the TE abundance, the slope of regression can be computed for the number of SNPs and IDs found in TEs

>1.0 and assemblies that underestimate the TE abundance a slope <1.0. To illustrate the usage of these metrics, we generated two assemblies of Canton-S: (a) an assembly based on Illumina short reads (ABySS; Simpson et al., 2009), and (b) an assembly based on ONT long reads (Canu; Koren et al., 2017) and several rounds of polishing using the long and the short reads (3× Racon and 3× Pilon; (Vaser et al., 2017; Walker et al., 2014). For both short and long reads, the coverage was 30×. The short-read assembly poorly reproduced the TE landscape (Figure 1b; Figure S4). The abundance of many families was underestimated, and the diversity of many TE families (SNPs and IDs) was overestimated (Figure 1b; Figure S4). By contrast, the long-read assembly captured the TE landscape much more accurately, with most slopes being close to the optimum (i.e. 1; Figure 1b; Figure S4). The high quality of long-read assemblies was also observed with different assembly algorithms (Figure S5) and with unpolished assemblies (Figure S4).

Next, we developed a novel metric that allows us to assess whether piRNA clusters are completely assembled. In essence, the CUSCO value (Cluster B*USCO*) estimates the fraction of completely assembled piRNA clusters (Figure 2a). Based on the reference genome of *D. melanogaster*, we identified pairs of flanking sequences for 85 out of the 142 annotated piRNA clusters of *D. melanogaster*

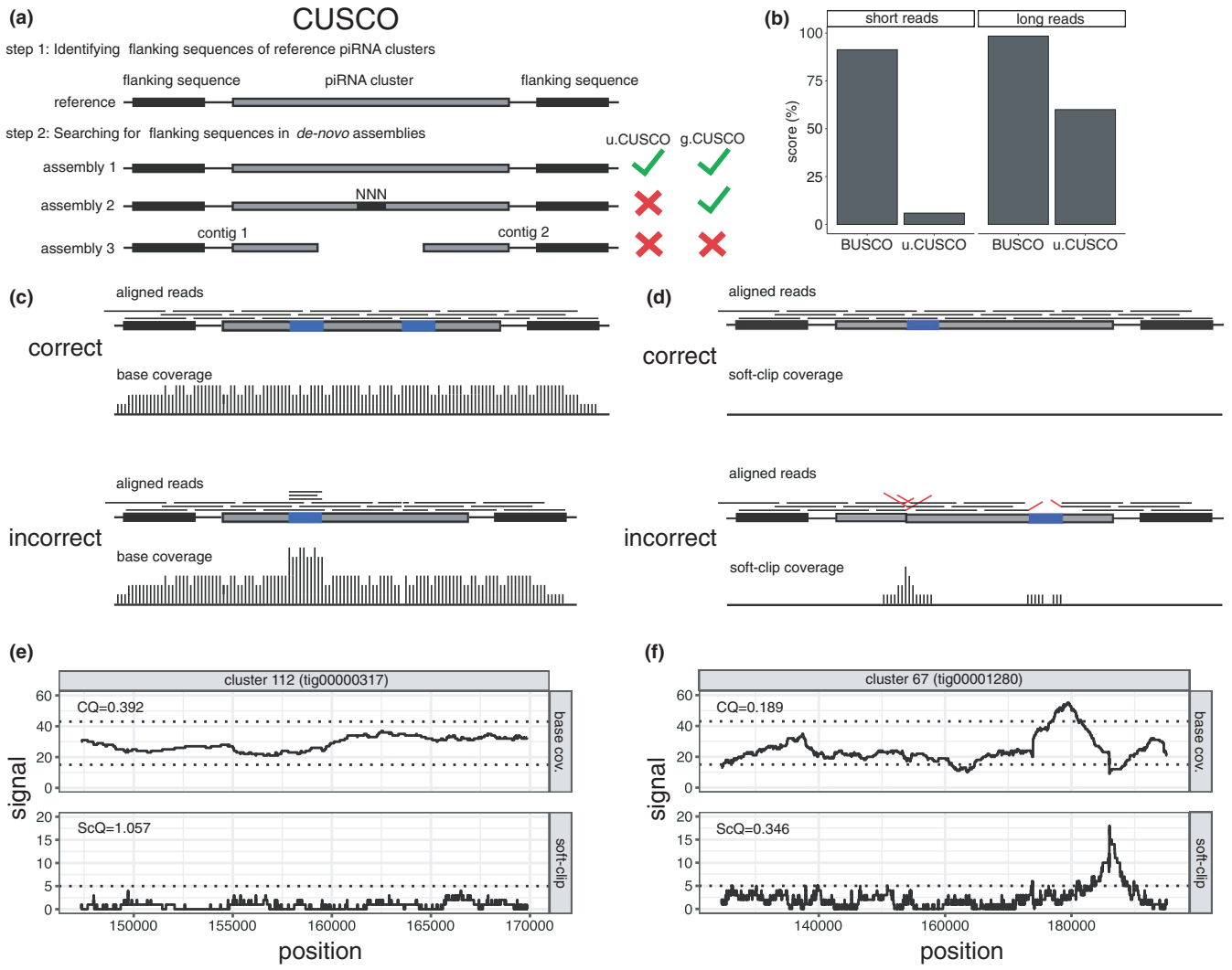

FIGURE 2 Novel metrics for assessing the quality of assembled piRNA clusters. (a) The CUSCO value (Cluster BUSCO) estimates the percent of complete piRNA clusters in an assembly of interest. Unique sequences flanking piRNA clusters are mapped to an assembly, and the CUSCO value is computed as the percentage of clusters where both flanking sequences align to the same contig. Depending on whether or not poly-N sequences (i.e. assembly gaps) are tolerated between the flanking sequences, an ungapped-CUSCO (u.CUSCO) and a gapped-CUSCO (g.CUSCO) can be computed. (b) BUSCO and CUSCO values for different assemblies of Canton-S (30x coverage for short and long reads). Although long- and short-read assemblies have similar BUSCO values, CUSCO values differ substantially. (c, d) Assembly errors in complete piRNA clusters may be identified based on (c) base-coverage heterogeneity and (d) elevated numbers of soft-clipped reads. Long reads aligned to a correct and a wrong assembly of a piRNA cluster are shown black. Red indicates not-aligned regions of long reads (i.e. soft-clipped regions). A repeat sequence is shown in blue. Example of an assembled piRNA cluster having a high (e) and low (f) quality. The clusters are from the long-read assembly of Canton-S (30×). Dotted lines show the 99% quantiles of the base coverage and of the soft-clip coverage in BUSCO genes. As a rough summary of the assembly quality for individual piRNA clusters, we compute the CQ (coverage quality) and ScQ values (soft-clip quality)

(Brennecke et al., 2007). Flanking sequences close to piRNA clusters were preferred. These flanking sequences are then mapped to an assembly of interest. Here, we consider a piRNA cluster to be 'complete' (analogous to the BUSCO terminology) when both flanking sequences align to the same contig/scaffold. We thus compute the CUSCO value as the fraction of pairs of flanking sequences aligning to the same contig/scaffold (Figure 2a). Depending on whether or not poly-N sequences (i.e. gaps in the assembly) are tolerated between the flanking sequences, an ungapped-CUSCO (i.e. contig-level) and a gapped-CUSCO (i.e. scaffold-level) can be computed (Figure 2a). Note that the ungapped-CUSCO is a subset of the gapped-CUSCO. The gapped-CUSCO thus includes both piRNA clusters with and without gaps. We ignored the length of piRNA clusters for computing CUSCO values as theoretical work suggests that piRNA clusters could be highly polymorphic: abundant presence/absence polymorphism of TE insertions in piRNA clusters may render the length of the clusters highly variable among individuals (Kelleher et al., 2018; Kofler, 2019). We illustrated the usage of CUSCO with the short- and the long-read assemblies of Canton-S (Figure 1b; Figure S4). CUSCO values differed substantially between

the short- and long-read assemblies (Figure 2b). A mere 5.88% of piRNA clusters were complete with the short reads, while 60% were complete with long reads. As we did not perform scaffolding, only the ungapped-CUSCO was calculated (Figure 2b). By contrast, both assemblies show high BUSCO values, which illustrates that BUSCO is of limited use for estimating the suitability of assemblies for an analysis of piRNA clusters (Figure 2b).

However, even when both flanking sequences align to the same contig, a piRNA cluster may still be incorrectly assembled, for example if some internal sequences are missing in the assembly. Therefore, we implemented two additional metrics that allow us to identify assembly errors in complete piRNA clusters (Figure 2c–f). Long reads aligned to an assembly of interest provide two complementary pieces of information that may allow us to identify assembly errors: the base coverage (based on aligned regions of reads) and the soft-clip coverage (based on not-aligned terminal ends of reads). Assembly errors such as repeat collapse or repeat expansions lead to marked differences in the base coverage. For example, a collapsed tandem repeat will result in an elevated coverage. An elevated heterogeneity of the base coverage is thus a hallmark of assembly errors (Figure 2c). However, the base coverage varies in all contigs, including correctly assembled ones. It is therefore necessary to distinguish base-coverage heterogeneity resulting from assembly errors from background heterogeneity. Here, we propose to use the heterogeneity of the base coverage of complete BUSCO genes as null expectation (Simão et al., 2015). BUSCO relies on genes that are conserved within a certain group, for example *Diptera* and estimates whether these genes are 'complete', 'partial' or 'missing' in an assembly. Complete BUSCO genes provide an ideal estimate of the background heterogeneity of the base coverage based on real data as complete BUSCO genes likely (a) occur as single-copy orthologs in an assembly, and (b) have few assembly errors (since the ORFs are mostly complete). We are thus relating the base-coverage heterogeneity of repetitive heterochromatic sequences (piRNA clusters) to euchromatic, conserved single-copy genes (BUSCO genes). Relying on complete BUSCO genes is however also convenient as BUSCO values are frequently computed for assessing the quality of novel assemblies anyway and a list of complete genes is provided per default by the BUSCO pipeline. We may then visualize the base coverage along piRNA clusters compared to different quantiles of the base coverage of BUSCO genes (e.g. the 99% quantile; Figure 2e,f). These quantiles are the lower and upper boundaries of the base coverage such that a certain fraction (e.g. 99%) of the base coverage of the BUSCO genes are between these boundaries. Base coverage levels exceeding or falling below these quantiles highlight potential assembly problems in piRNA clusters (Figure 2f). As a rough summary of the base-coverage heterogeneity over the entire sequence of a cluster, we may compute the base-coverage quality (CQ) for each piRNA cluster: $\mathrm{CQ} = \tilde{s}_{busco}/s_{cluster}$, where $\tilde{s}_{busco}$ is the median standard deviation of base coverages of BUSCO genes and $s_{cluster}$ the standard deviation of the base coverage of a given piRNA cluster. We used the median to guard against potential outliers in the base-coverage heterogeneity of BUSCO genes. Low CQ values (<<1.0) indicate a heterogeneous base coverage in piRNA clusters and thus highlight potential assembly problems (Figure 2e,f).

However, some assembly errors, such as deleted or misplaced sequences, might not have noticeable effects on the base coverage. These assembly problems are instead characterized by breaks in the assembly where sequences are joined in the assembly that are not joined in the genome of the organism. As a consequence, many reads spanning these breaks can only be partially aligned back to the assembly (Figure 2d). These reads are usually soft-clipped; that is, a terminal end of a read is either not aligning to any contig or aligning to an entirely different location. Soft-clipped reads can thus be used to identify assembly problems. Therefore, we propose to compute the soft-clip coverage along piRNA clusters as a complementary metric to the base-coverage heterogeneity (Figure 2d). Iterating over all reads, we compute the coverage resulting from the soft-clipped regions of reads; that is, soft-clipped regions are treated as if they were aligned to the reference (Figure 2d). Actually aligned regions of reads are ignored for computing the soft-clip coverage. As null expectation we rely on the soft-clip coverage of complete BUSCO genes. This allows us to visualize the soft-clip coverage along piRNA clusters compared to different quantiles of the soft-clip coverage based on BUSCO genes (e.g. the 99% quantile, Figure 2e,f). Note that solely an upper quantile is computed for the soft-clip coverage (a low soft-clip coverage is ideal), whereas a lower and an upper quantile is computed for the base coverage. A pronounced peak in the soft-clip coverage indicates the likely position of an assembly break (Figure 2f). The soft-clip quality (ScQ) roughly summarizes the assembly quality of a given piRNA cluster: $\mathrm{ScQ} = \tilde{c}_{busco}/c_{cluster}$, where $\tilde{c}_{busco}$ is the median of the average soft-clip coverages of BUSCO genes and $c_{cluster}$ the average soft-clip coverage of a given piRNA cluster (Figure 2e,f). Low ScQ values again highlight piRNA clusters that may contain assembly errors. To provide an estimate of quality of an assembly, we can compute the average CQ or ScQ values for all piRNA clusters in an assembly of interest. In summary, the base coverage and the soft-clip coverage can be used to estimate assembly quality at three different levels: (a) to identify errors within a piRNA clusters (e.g. elevated soft-clip coverage at a particular site), (b) to estimate the assembly quality of a particular piRNA cluster (CQ and ScQ) and (c) to estimate the overall assembly quality (average CQ and ScQ). The identification of potential assembly errors in piRNA clusters (a and b) will likely be the main application of these coverage-based metrics.

In summary, we developed novel quality metrics that enable us to estimate the assembly quality of piRNA clusters. First, the TE landscape metrics test whether an assembly accurately reproduces TE abundance and diversity (SNPs and IDs) of an organism. Second, the CUSCO estimates the fraction of complete piRNA clusters. Third, CQ and ScQ values summarize the quality of complete piRNA clusters, where the base-coverage heterogeneity and the soft-clip coverage along piRNA clusters allow us to identify the location of potential assembly problems. We made the scripts for computing our novel quality metrics and for visualizing the quality along piRNA clusters publicly available https://sourceforge.net/projects/cuscoquality/. We additionally provide the sequences flanking piRNA clusters, a manual and a walkthrough.

3.2 | Optimizing the assembly strategy

Next, we aimed to identify an assembly strategy that enables us to generate high-quality assemblies of the piRNA clusters of the *D. melanogaster* strain Canton-S. At first, we evaluated the performance of four different long-read assemblers, which rely on slightly different algorithms. Miniasm (Li, 2016) uses the overlap among reads to build contiguous sequences. Canu (Koren et al., 2017) utilizes a similar approach as miniasm. However, to reduce the error rate, Canu trims reads and generates consensus sequences of reads prior to the assembly. Wtdbg2 (Ruan & Li, 2020) uses a de Bruijn graph-based assembly algorithm, where k-mers are much larger than for short reads. Flye (Kolmogorov et al., 2019) initially generates disjointigs (concatenations of disjoint genomic segments), builds an assembly graphs and then uses reads to untangle the assembly graph. Flye was designed for an improved assembly of repetitive regions. Long reads usually have high error rates, and assemblies based on these reads may thus also contain an appreciable number of errors (Sović et al., 2016; Vaser et al., 2017). Following recommendations of previous works (Chakraborty et al., 2019; Ellison & Cao, 2020; Solares et al., 2018), we aimed to reduce the error rate by polishing the assembly with Racon (long reads) (Vaser et al., 2017) and Pilon (short reads) (Walker et al., 2014). Polishing algorithms align reads to an assembly and infer the consensus sequence (Vaser et al., 2017; Walker et al., 2014). Initially, we were concerned that this procedure could eliminate polymorphisms from TE sequences, such that the number of SNPs and IDs of TEs may be underestimated in polished assemblies. However, we found that polished assemblies capture the TE landscape slightly more accurately than unpolished assemblies (TE abundance: unpolished = 1.02, polished = 1.01; SNP metric: unpolished = 0.95, polished = 0.99; ID metric: unpolished = 0.93, polished = 0.99; Tables S4 and S5). Polishing thus enhances the suitability of assemblies for genomic analysis of TEs. We performed one to three rounds of polishing with Racon and Pilon, where the optimal number of iterations was selected based on the maximally attained BUSCO values (Table S2).

To investigate the influence of coverage on assembly quality, we evaluated the performance of each assembler with several different coverages. Reads were randomly subsampled to coverages ranging from 20 to 150× (Figure 3). Note that a minimum coverage of 20× was required for Canu and wtdbg2. To assess the quality of the assemblies, we combined our novel quality metrics with classical metrics (NG50, BUSCO and assembly length) (Figure 3; Table S6). However, we noticed that BUSCO values are very similar among the evaluated coverages and assemblers, suggesting that BUSCO is of limited use for estimating the suitability of assemblies for TE research (Table S6). When considering relevant metrics (TE landscape metrics, NG50, CUSCO, assembly length, CQ and ScQ), we found that the quality of the assembly depends on the coverage but not the assembler (ANOVA comparing linear models; model1: metric, coverage, assembler; model2: metric, coverage; model3: metric, assembler; model1 versus model2 $p = .47$; model1 versus model3 $p = .036$). When solely considering NG50 and CUSCO as metrics, the assembler (but not the coverage) had a significant influence on the assembly quality (ANOVA comparing linear models; model1 vs. model2 $p = .0004$; model1 vs. model3 $p = .11$). This indicates that the quality of assemblies depends on the assembler and the coverage. Interestingly, the best assemblies were not necessarily obtained when all reads were used (Figure 3). For example, Canu and Flye yielded the largest NG50 with a coverage of 100x ($Canu_{100x}$ = 8.1 Mbp, $Canu_{150x}$ = 3.6 Mbp, $Flye_{100x}$ = 17.2 Mbp; $Flye_{150x}$ = 10.5 Mbp) and miniasm the best representation of TEs at a coverage of 50x ($miniasm_{50x}$ = 1.01, $miniasm1_{50x}$ = 1.11). Based on our novel quality metrics (abundance, SNPs, IDs, CUSCO, CQ and ScQ), Canu and miniasm outperformed wtdbg2 and Flye at all evaluated coverages (Figure 3; Figure S6). At most coverages, Canu captured the TE abundance more accurately than miniasm, Flye and wtdbg2 (Figure 3). Assemblies generated with Canu mostly had the highest CUSCO values (Figure 3), where up to 80% of the piRNA clusters were contiguously assembled with coverages ranging from 100× to 150×. Furthermore, Canu generated the most reliable assemblies of piRNA clusters (average CQ and ScQ values; Figure S6). Although Flye yielded the highest NG50 values, it also generated the shortest assemblies (Figure 3). The Canu assemblies were the largest at most coverages and showed intermediate NG50 values (Figure 3). Overall, we conclude that Canu yielded the most contiguous (highest ungapped-CUSCO) and the most reliable (highest CQ and ScQ) assemblies of piRNA clusters (Figure 3). For the remainder of this manuscript, we thus relied on assemblies generated with Canu.

When reads are randomly sampled, large portions of the data will not be used for the assembly. These unused data may, however, still contain long reads that could be useful for improving the quality of assemblies, for example by bridging gaps between contigs. Thus, we asked if the assembly quality could be further enhanced by sampling the longest reads instead of a random subset. To test this, we sampled subsets of the longest reads with coverages ranging from 20× to 150× (Figure S7). The mean read length of these subsets ranged from 25,051 bp with 20× coverage to 7146 bp with 150× coverage (Figure S7a). Canu assemblies based on the longest reads usually have higher NG50 values than assemblies based on random reads (Figure S7b). The largest NG50 values were obtained when a coverage of 100x was used (Figure S7c). Interestingly, CUSCO values were consistently highest for assemblies generated with the longest reads (Figure S7c), while the coverage had little influence on the quality of the assembled piRNA clusters (average CQ and ScQ; Figure S8). The three TE landscape metrics (abundance, SNPs, IDs) revealed little differences between assemblies generated with random reads and the longest reads (Figure S9).

Finally, we were interested in whether CUSCO values could be further improved by using de-novo scaffolding with Hi-C data (Figure S7d). Scaffolding algorithms link contigs into longer sequences based on diverse information such as genetic maps, optical maps or the conformation of chromosomes (Rice & Green, 2018). One widely used approach for scaffolding, Hi-C, relies on the three-dimensional organization of chromosomes (Lieberman-Aiden et al., 2009). With Hi-C, chromatin interactions may be identified by

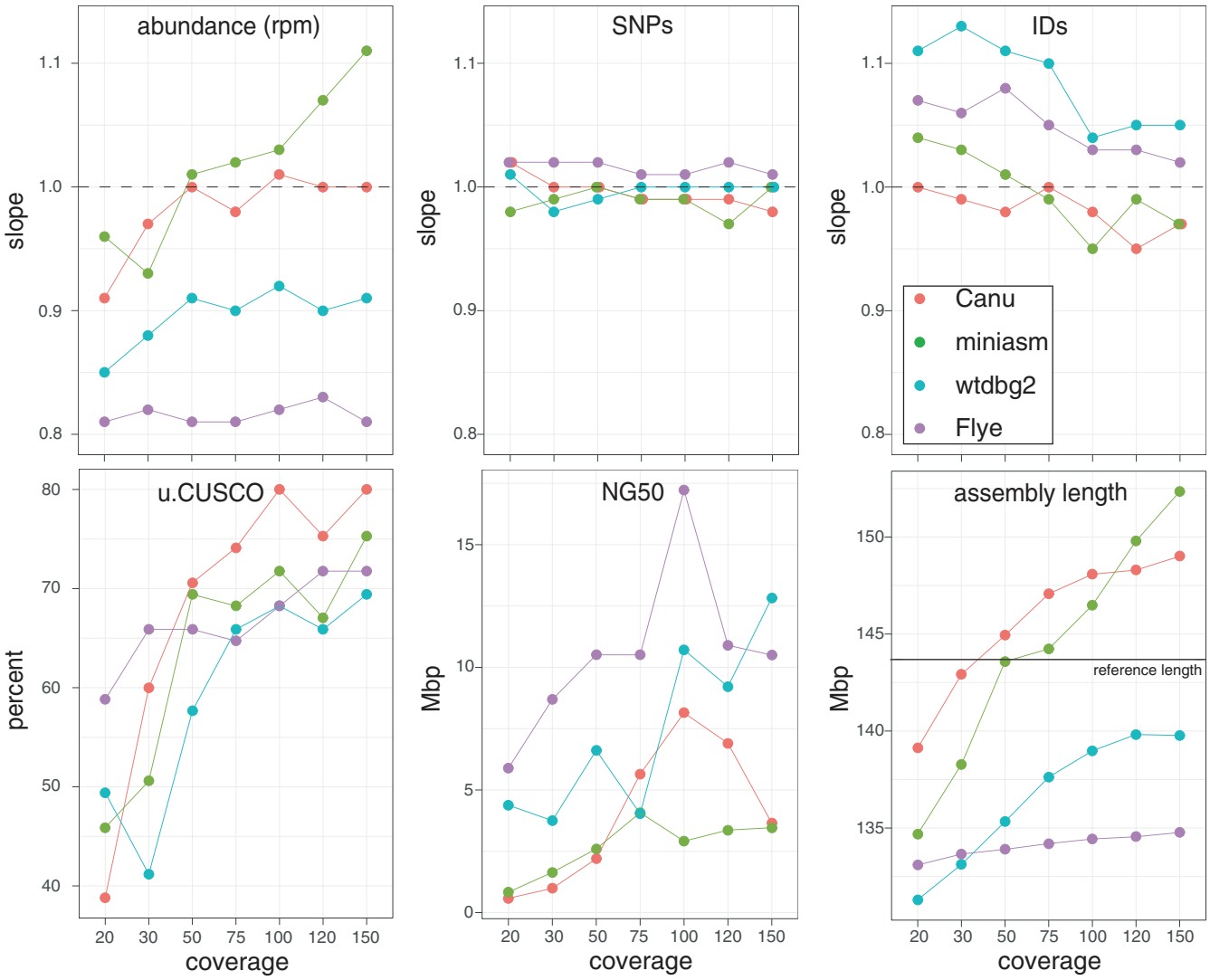

FIGURE 3 Influence of the assembly algorithm (Canu, miniasm, wtdbg2, Flye) and the coverage on the quality of assemblies. Results are shown for our novel TE-centred quality metrics (CUSCO, abundance, SNPs, IDs) as well as classic quality metrics (NG50, BUSCO). Dashed lines indicate optimal performance

sequencing fragments that were physically in close proximity (Rice & Green, 2018; Sedlazeck et al., 2018). Since chromatin interactions are most often observed among neighbouring sites within chromosomes, Hi-C data can also be used for scaffolding (Rice & Green, 2018; Sedlazeck, Lee, et al., 2018).

As scaffolds usually contain gaps of unknown size between the contigs (mostly indicated by 100 'N' characters), we calculated the gapped-CUSCO (Figure S7d).

Despite a substantial increase in NG50 values (145–1033%; Figure S10), scaffolding with Hi-C data only moderately improved the CUSCO values (3.5–20%; Figure S7d). This improvement was most pronounced at low coverages, where CUSCO values were quite low before scaffolding. We note that the clusters scaffolded with Hi-C contained gaps, that is missing sequences, mostly of unknown size (see below). Thus, it is crucial to distinguish between gapped- and ungapped-CUSCO to assess the quality of an assembly. As expected, other quality metrics, such as BUSCO and the three TE landscape metrics, were not influenced by Hi-C-based scaffolding (Table S5).

In summary, we found that our novel metrics are useful for assessing the quality of assemblies. Depending on the choice of the investigated regions (number and complexity), CUSCO may be a sensitive metric that identifies quality differences among assemblies not found by other metrics. With long reads and an optimized assembly strategy, up to 81% of the piRNA clusters may be contiguously assembled in *D. melanogaster*. Especially assemblies based on Canu and a subset of the longest reads (100× coverage) had a high quality. Finally, we found that Hi-C data were of limited use for assembling piRNA clusters.

3.3 | Influence of segregating polymorphisms on assembly quality

Based on Canton-S, we showed that long reads enabled us to generate high-quality assemblies of piRNA clusters. However, Canton-S is highly isogenic, having few segregating polymorphisms

(Figure S11a). We were interested whether piRNA clusters may also be reliably assembled for a less isogenic strain. We relied on the *D. melanogaster* strain Pi2, which is frequently used in TE research, for example, to assess the extent of P-element (a DNA transposon)-induced infertility in females (O'Hare et al., 1992; O'Hare & Rubin, 1983; Srivastav et al., 2019). Pi2 has substantial numbers of segregating SNPs on several chromosomes (Figure S11a). We first generated a high-quality data set for Pi2: 199x ONT long reads (mean read length ≈ 8 kb), 40× of Illumina PE data and 260× coverage Hi-C data (Table S1). An assembly of Pi2 was generated with our previously established strategy: 100x of the longest ONT reads (mean read length 19,219 bp) were assembled with Canu, the assemblies were subject to multiple rounds of polishing, and Hi-C data were used for scaffolding (Table S5). Based on our novel quality metrics, the assemblies of Canton-S and Pi2 are mostly of similar quality (apart from CQ and ScQ values, which have slightly lower values in Pi2; Table S5).

We noticed that the Pi2 assembly is substantially larger than the Canton-S assembly (12.5% larger, Table S5). This difference in assembly size might be due to the polymorphisms segregating in Pi2, where the assembly algorithm could have generated several contigs (e.g. a contig for each homologous chromosome) for polymorphic regions (Pryszcz & Gabaldón, 2016). To test this hypothesis, we sliced assemblies into nonoverlapping fragments of 1kb, aligned them to the reference sequence and calculated the average coverage for 100 kb windows (Figure S11a). Uniquely assembled regions will have a coverage of 1, whereas regions assembled multiple times will have a coverage >1. We observed many multiple-assembled regions for the Pi2 assembly (Figure S11b) that largely overlap with polymorphic regions (Figure S11c). Pi2 had more multiple-assembled regions than Canton-S (paired Wilcoxon rank-sum test, $V = 370610$, $p \leq .0001$). Segregating polymorphism in Pi2 thus led to redundantly assembled contigs which likely account for the large assembly size of Pi2 (Table S5). Interestingly, we did not observe any redundant assemblies of piRNA clusters for Pi2 (we tested if both sequences flanking piRNA clusters map to multiple contigs). This absence of redundant clusters is likely due to the fact that the vast majority of the piRNA clusters lie in regions with few segregating polymorphisms in Pi2 (Figure S12). This may however not necessarily hold for other strains.

Polymorphic regions are also problematic as it is unclear on how to deal with heterozygous TE insertions. We therefore searched for heterozygous (0.25 ≤ frequency ≤ 0.75) structural variants (SVs) in our assemblies, using Sniffles (Sedlazeck, Rescheneder, et al., 2018). In total, we identified 9 heterozygous indels with a minimum size of 1 kb in our Canton-S assembly and 108 in our Pi2 assembly (Figure S11d). A blast search revealed that 66.67% and 84.26% of these SVs in Canton-S and Pi2, respectively, were due to TEs. Two of these heterozygous TE SVs were found in piRNA clusters of Pi2 and none in piRNA clusters of Canton-S. Due to these difficulties, we recommend to use highly isogenic strains for assembling piRNA clusters.

3.4 | Finalizing assemblies

To provide chromosome-scale assemblies of Pi2 and Canton-S to the community, we manually broke up misassemblies (Figure S13) and performed reference-based scaffolding with RaGOO (Alonge et al., 2019). Reference-based scaffolding raised the gapped-CUSCO to 95.3 for Canton-S and to 97.7 for Pi2 but had little effect on other quality metrics (Table S5). An overview of the quality of the final assembly, including the quality at the different assembly steps, can be found in Table S5. The assemblies of Canton-S and Pi2 are available at NCBI (PRJNA618654).

3.5 | Composition of piRNA clusters

Next, we investigated the quality and composition of the assembled piRNA clusters in more detail. We compared piRNA clusters between our chromosome-scale assemblies of Canton-S and Pi2 to the reference genome. Assembly errors, but also presence/absence polymorphism of TE insertions in piRNA clusters, could lead to vast size differences of clusters among assemblies. Thus, we first investigated the length of the piRNA clusters (i.e. the distance between the two sequences flanking each cluster). The length of ungapped clusters in both assemblies is very similar to the length in the reference genome (release 6; paired Wilcoxon rank-sum test; CS: $V = 951$, $p = .55$; Pi2: $V = 1259.5$, $p = .45$; Figure 4a). Solely 19 clusters in Pi2 and 25 clusters in Canton-S deviated in length by more than 20% from the length of the clusters in the reference genome. Some of this size variation (11 in Pi2 and 11 in Canton-S) was due to clusters with a gap in the assembly (recognized by several 'N' characters; Figure 4a, coloured dots). An analysis of gapped clusters revealed a significant length difference in Canton-S, indicating that length estimates of clusters with gaps might not be reliable (paired Wilcoxon rank-sum test; CS: $V = 995$, $p = .007$; Pi2: $V = 1520.5$, $p = .51$).

When we estimated the quality of the assembled piRNA clusters using CQ and ScQ, we observed considerable differences among the clusters in both assemblies (Figure 4b,c). As expected, piRNA clusters with assembly gaps have low CQ values (Figure 4b). By contrast, assembly gaps had little impact on the ScQ values (Figure 4c). Investigating the base coverage and the soft-clip coverage along each position of some clusters with low and high ScQ values revealed potential assembly issues at some positions of clusters with a low ScQ but not in the clusters with a high ScQ (Figure S14). Taken together, this illustrates that both CQ and ScQ values help to identify clusters with potential assembly issues. However, solely an analysis of the base coverage and the soft-clip coverage along clusters will provide detailed information about the abundance and position of potential assembly problems. For comparing the composition of piRNA clusters, it is therefore necessary to consider the annotations of the clusters as well as the quality along clusters. We illustrate this approach with 42AB, one of the largest contiguously assembled clusters in *D. melanogaster* (Figure 5). We computed the

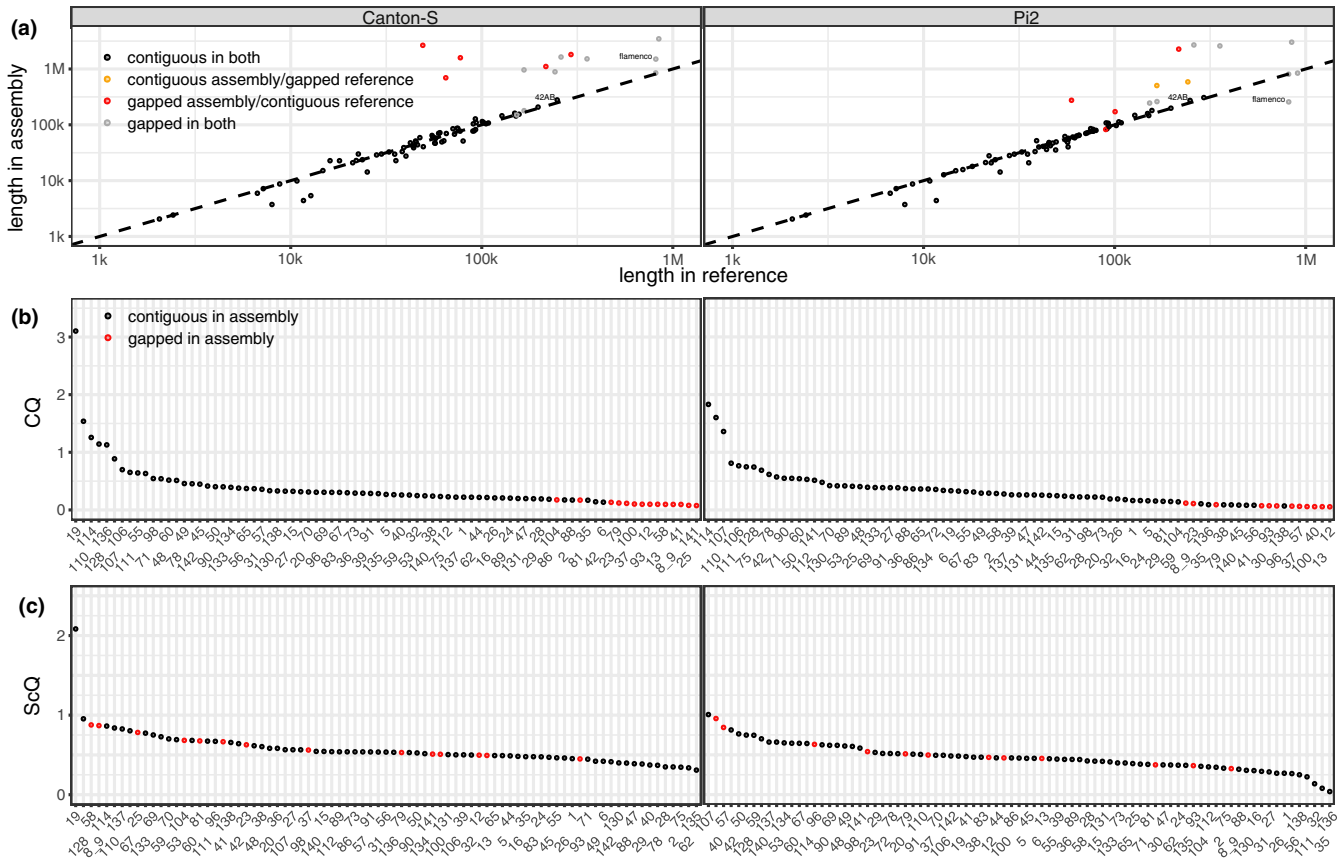

FIGURE 4 Overview of the length and quality of the piRNA clusters in our assemblies of Canton-S and Pi2. (a) The length of piRNA clusters in our assemblies is similar to the length of the clusters in the reference genome (*x*-axis). Clusters with assembly gaps (i.e. 'N' characters in assembly) are indicated in colour. (b) Base-coverage quality (CQ) of the piRNA clusters in our assemblies. The *x*-axis shows the ID of each piRNA cluster based on Brennecke et al. (2007). (c) Soft-clip quality (ScQ) of the piRNA clusters in our assemblies

base coverage and the soft-clip coverage for 42AB in Canton-S and Pi2 (Figure 5). We also annotated TEs with RepeatMasker (Smit et al., 2015) and identified sequence similarity between our assemblies and the reference genome with Blast (Figure 5; Altschul et al., 1990). The base coverage and the soft-clip coverage of 42AB in Canton-S are mostly within the 99% quantiles of BUSCO genes, which suggests that this assembly is of high-quality (Figure 5). However, the soft-clip coverage and to a lesser extent also the base coverage of 42AB in Pi2 is elevated at the end and in the central simple-repeat region, indicating potential assembly problems in these regions (Figure 5). When searching for causes of these potential assembly problems with Sniffles, we found a P-element insertion with a frequency of 100% at a site of the elevated soft-clip coverage in Pi2 (Figure 5), demonstrating the utility of our novel quality metrics. We did not find a cause for the elevated soft-clip coverage in the central regions of 42AB in Pi2. Most TE insertions are shared between the three strains, and large synteny blocks, frequently involving several TE insertions, can be found (Figure 5). Nevertheless, we also found differences among the three strains (Figure 5). Most notably, a 26-kb region – involving the X-element, GATE, Max-element and rover – was duplicated in Pi2 (Figure 5). Relative to the reference genome, we also found several TE presence/absence polymorphism in both strains (7 in Pi2 and 11 in Canton-S; Figure 5). Interestingly, most of these polymorphic TEs show little divergence from the consensus sequence (<1%; Figure 5), which suggests that these polymorphisms are due to recent TE insertions into 42AB. These polymorphisms are largely in regions with inconspicuous base coverage and soft-clip coverage, which suggests that they are not due to assembly mistakes. Apart from Chimpo, which was identified using RepBase (Bao et al., 2015), all TEs identified in the cluster 42AB were present in the consensus sequences of TEs in *D. melanogaster* (version 10.01; Quesneville et al., 2005).

Finally, we validated several of the polymorphic TE insertions in piRNA clusters with PCR. In the cluster 42AB, we confirmed 11 out of the 14 tested polymorphic TE insertions, including the missing P-element insertion and the large duplication in Pi2 (7 present in Pi2; 3 present in Canton-S; 7 present in Iso-1 of which three are shared with Pi2; Figure S15; Table S3). In other piRNA clusters, we confirmed 20 out of the 22 tested polymorphic TE insertions (12 present in Pi2; 10 present in Canton-S; Figure S15; Table S3).

We conclude that our assembly strategy yields contiguous sequences of many piRNA clusters. Furthermore, our novel quality metrics may be used to identify the location of potential assembly problems in piRNA clusters.

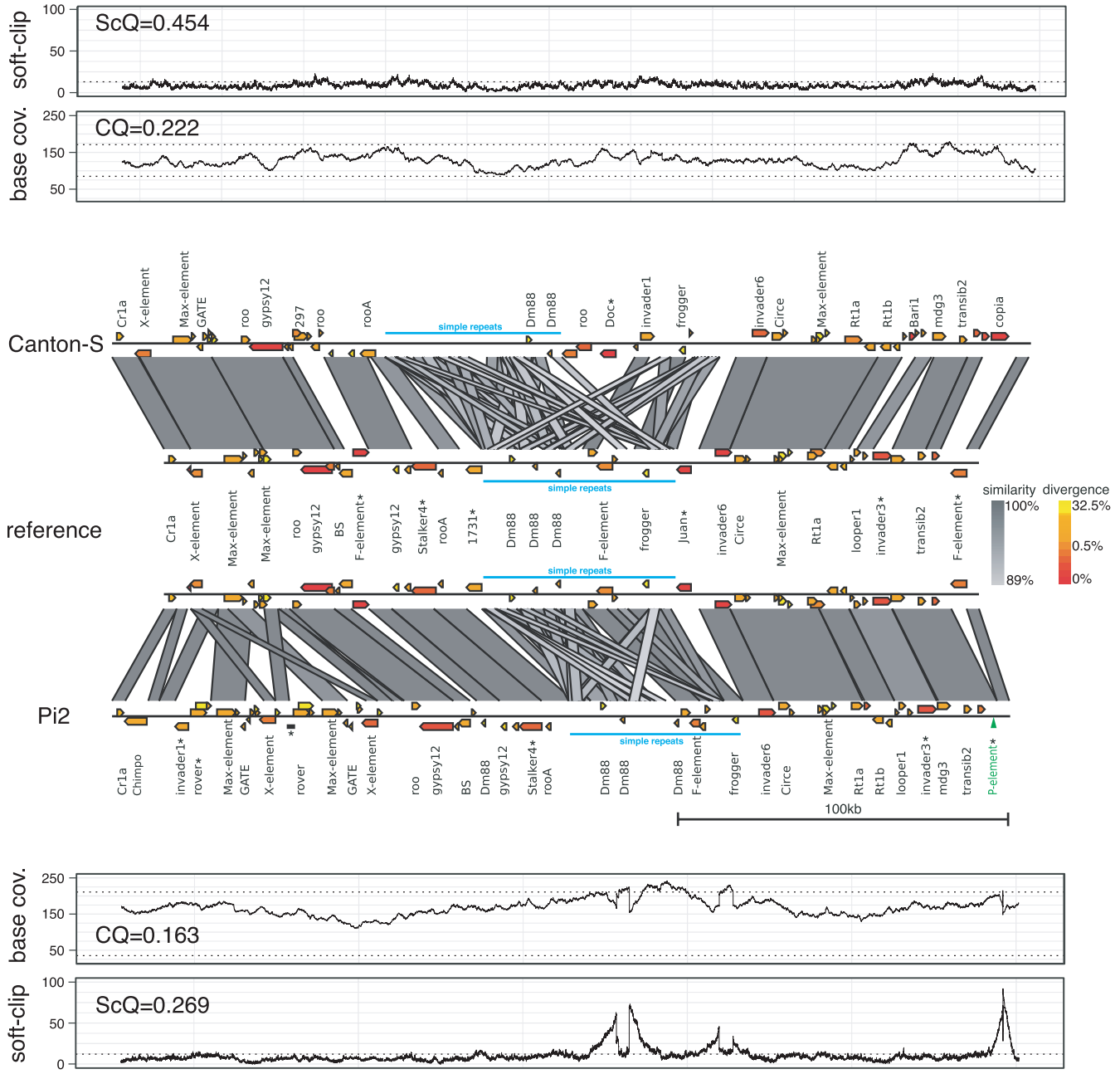

FIGURE 5 The sequence of the cluster 42AB in our assemblies compared to the reference genome. The TE annotation (yellow-red gradient indicates similarity to the consensus sequence of the TE) and sequence similarity to the reference genome (grey gradient indicates the degree of similarity) are shown. PCR validated presence/absence polymorphisms of TEs or SVs are marked with a '*'. A TE insertion missed in the assembly is shown in green. The base coverage and soft-clip coverage are shown for Canton-S (top) and for Pi2 (bottom). The 99% quantiles based on BUSCO genes are shown as dotted lines. Note that the soft-clip coverage and to a lesser extent the coverage is elevated at the site of the missing TE insertion and in the simple-repeat region, indicating possible assembly problems

3.6 | Extending our approach to different species

To demonstrate the generality of our approach, we extended our metrics to different species. We first tested the TE landscape metrics with a short- and a long-read based assembly (observations) of the same human individual (Korean reference genome: KOREF1.0 (Cho et al., 2016) and PB_62x (Kim et al., 2019)). The expected TE abundance and diversity was derived from the short-read data (Cho et al., 2016). The TE landscape metrics are based on 1063 TE families. We did not compute the ID metric as solely 39 TE families possessed IDs in the 'expected' data set. Similarly to *Drosophila*, the long-read assembly of humans captures the abundance and diversity of TEs better than the short-read assembly (abundance: long-read = 0.898, short-read = 0.824; SNPs: long-read = 1.051, short-read = 1.056; Figure S16).

To extend CUSCO to humans, we designed flanking sequences for 168 piRNA clusters (Sarkar et al., 2014) and obtained unique flanking sequences for 136 of them. We applied CUSCO to 11

publicly available human assemblies, where five are different versions of the same Korean individual (Cho et al., 2016; Kim et al., 2019). Although BUSCO values were nearly identical among the assemblies the CUSCO values showed more variation, where especially the ungapped-CUSCO revealed marked differences among the assemblies (Kolmogorov-Smirnoff test; u.CUSCO vs. BUSCO $p = .006$; Figure 6a). The lowest ungapped-CUSCO value was obtained with the short-read assembly (Figure 6a KOREF1.0; u.CUSCO = 10.29), while all long-read assemblies had markedly higher ungapped-CUSCO values (between 38.24 for HG00733 and 98.53 for T2T). This shows that CUSCO is a sensitive metric in humans, and can be used to identify assemblies with a high fraction of assembled piRNA clusters.

Next, we asked if the base-coverage heterogeneity and the soft-clip coverage can be used in humans to identify clusters with potential assembly errors. We investigated the cluster chr4.117 in different versions of the Korean reference genome (KOREF). This cluster has an apparent polymorphism in the short-read assembly (Figure 6c KOREF1.0). However, the base coverage and soft-clip coverage reveal that this polymorphism is likely an assembly error (Figure 6c). Accordingly, this cluster has high ScQ and CQ values in the long-read assembly (KOREF PB_62x) but low values in the short-read assembly (Figure 6c). We thus argue that the base-coverage heterogeneity and the soft-clip coverage will be useful to identify potential assembly problems in human piRNA clusters.

So far we used the CUSCO solely with piRNA clusters. However, our approach where sequences flanking piRNA clusters are aligned to assemblies can be extended to any regions of interest, such as heterochromatic regions or rDNA clusters. This would also enable extending the CUSCO approach to species not having piRNA clusters such as plants. To test whether our CUSCO approach can be used with such alternative regions, we designed flanking sequences for the 10 *KEE* regions forming the KNOT region in *A. thaliana* (Grob et al., 2014). These *KEE* regions are thought to be involved in control of TEs (Grob et al., 2014). Although the assemblies had similarly high BUSCO values, the CUSCO (i.e. the fraction number of complete *KEE* regions) differed significantly (Kolmogorov–Smirnoff test; u.CUSCO vs. BUSCO $p = .004$) among the assemblies (Figure 6c). The short-read assembly (AthNd1_v1.0) again had the lowest CUSCO value (10.0) (Figure 6c). However, the resolution with solely 10 *KEE* regions is rather coarse as compared to humans (136 clusters) and *D. melanogaster* (85 clusters). CUSCO values will likely be most informative if they are based on many regions.

In summary, we argue that our quality metrics can be readily extended to diverse species and that CUSCO in particular is a sensitive metric detecting differences in assembly quality that are not easily detected by classic metrics such as BUSCO.

4 | DISCUSSION

Here, we showed that long-read sequencing technologies enable us to generate high-quality assemblies of piRNA clusters. With an optimized assembly strategy, more than 80% of the piRNA clusters in *D. melanogaster* may be assembled, which can be increased up to 98% with scaffolding approaches.

4.1 | Novel quality metrics

Since current metrics of assembly quality largely ignore TEs and piRNA clusters, we introduced several novel quality metrics.

With three metrics, we first estimate whether an assembly accurately captures the TE landscape (abundance, SNPs and IDs) of an organism. These metrics may be viewed as a general control, since it is unlikely that repeat rich regions, like piRNA clusters, have been accurately assembled when TEs are poorly represented in the assembly. Unfortunately, the real TE landscape is not known for any organism. However, we argue that Illumina raw reads may be used to derive a useful approximation of the expected TE landscape. Assuming that reads are more or less randomly distributed over the genome, and that sequencing errors are largely random within reads, this assumption should mostly be valid. Sequencing errors can be largely eliminated from the analysis of the abundance of SNPs and IDs by using a minimum allele frequency (Kofler, Orozco-terWengel, et al., 2011). Here, we used a minimum allele frequency of 2% for SNPs and IDs. In case more stringent criteria are required, a higher threshold may be used. The coverage will fluctuate over the genome, which could affect estimates of TE abundance. Especially, the GC-bias, where regions with a high GC content have an elevated coverage (Minoche et al., 2011), could lead to overestimating the expected abundance of TEs with a high GC content. However, since we sum the average coverage over many different insertions of a TE family, with insertion sites in diverse genomic backgrounds (with varying GC contents), the influence of the GC-bias and of stochastic coverage fluctuations should be minimized by our approach. Furthermore, since we rely on the slope between the expected and the observed TE abundance, which is based on many TE families with different GC contents, the influence of the GC bias should be further reduced. In agreement with this, we did not find any correlation between GC content and TE abundance in the raw reads (Figure S17). Moreover, we solely found a small but nonsignificant difference in GC content among TEs that are well-represented in genomes compared to TEs that are not well-represented (Figure S17). Finally, it is reassuring that an assembly based on long reads captures the expected TE landscape more accurately than an assembly based on the short reads, which have been used for estimating the expected TE landscape (Figure 1).

Apart from sequencing biases, also biases occurring during data analysis, such as mapping and quantifying of reads may occur. Since we use the same pipeline for the raw reads (expectations) and the artificial reads derived from an assembly (observations), these biases should largely be eliminated. Taken together, we think that Illumina raw reads provide a useful approximation of the expected TE landscape. Since computing the TE landscape metrics only requires Illumina short reads for an organism (which are often generated

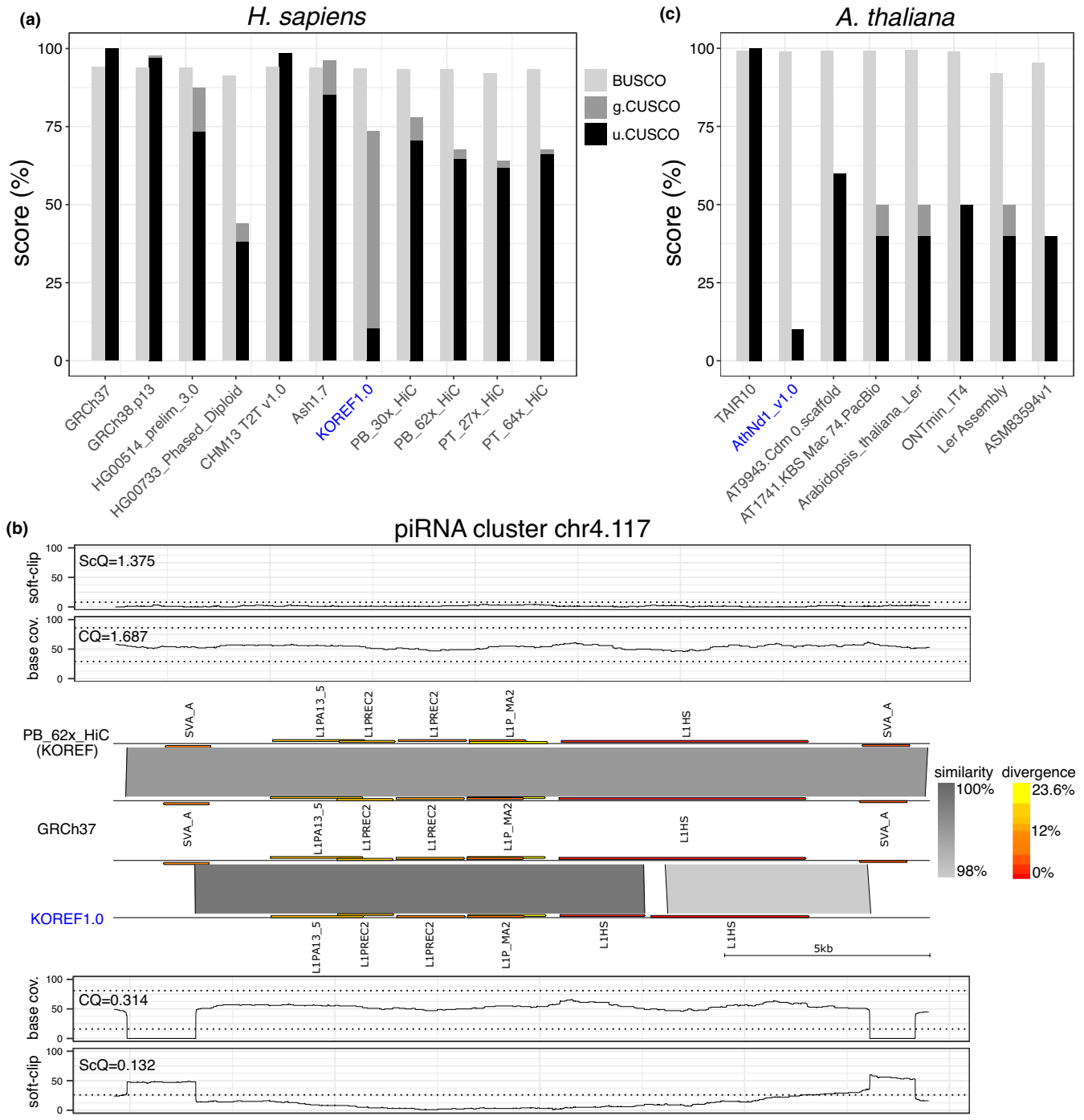

FIGURE 6 Extending the quality metrics to different species. (a) CUSCO and BUSCO values of different human assemblies. (b) Coverage heterogeneity and soft-clip coverage for a short- and a long-read assembly of the KOREF individual. Note that our metrics reveal misassemblies at both ends of the short-read assembly. (c) CUSCO and BUSCO values for different *A. thaliana* assemblies. CUSCO values are based on flanking sequences of the 10 *KEE* regions. Short-read assemblies are labeled in blue

for the polishing of assemblies anyway) and consensus sequences of TEs, these metrics may thus be used for model and nonmodel organisms (assuming some TE sequences are available or identified *de novo*).

The CUSCO value estimates the fraction of contiguously assembled piRNA clusters based on an alignment of unique sequences flanking the clusters. piRNA clusters are of central importance for TE biology as they are thought to act as genomic traps that stop TE invasions (Bergman et al., 2006; Duc et al., 2019; Goriaux et al., 2014; Malone & Hannon, 2009; Ozata et al., 2019; Yamanaka et al., 2014; Zanni et al., 2013). However, the CUSCO may generally be a useful metric for assessing the quality of assemblies. An increased CUSCO indicates a more contiguous and thus generally more complete assembly. Furthermore, CUSCO allows us to differentiate

between assemblies of very different qualities, as the difficulty of assembling a piRNA cluster varies substantially among the clusters. Long clusters may, for example, be much more challenging to assemble than short ones. This broad range of CUSCO values is demonstrated by our assemblies of Canton-S, where the CUSCO ranges from 5.88% (short reads, ungapped CUSCO), over 81.18% (long reads, ungapped CUSCO) to 95.29% (scaffolding, gapped CUSCO). Also, results in humans and *A. thaliana* support the broad range and general applicability of CUSCO (Figure 6). Depending on the choice of the repetitive region (number and complexity), CUSCO may thus be a sensitive quality metric capable of differentiating among assemblies of diverse qualities, even when assemblies have a similar quality according to other metrics such as BUSCO (Figure 3; Figure 6; Table S6).

It is important to distinguish between ungapped- and gapped-CUSCO values. Clusters containing gaps likely miss some sequences, including TE insertions, which prevents a comprehensive analysis of the composition of clusters. It is thus most important to maximize ungapped-CUSCO values. However, scaffolding algorithms, which introduce gaps between adjacent contigs, have been used to generate most publicly available assemblies (Figure S18). In these assemblies, many piRNA clusters may contain gaps. To gain a complete picture of piRNA clusters in an assembly, we thus recommend evaluating both CUSCO values (our script computes both).

Identification of the sequences flanking piRNA clusters requires a reference genome. Hence, CUSCO can only be used with species with a reference genome and an annotation of piRNA clusters. But even for species with a reference genome, it will not be feasible to identify suitable flanking sequences for all piRNA clusters (e.g. clusters at terminal ends of contigs/chromosomes).

One limitation of CUSCO is that the sequences flanking piRNA clusters need to be identified for each species separately. However, once sequences flanking piRNA clusters are identified, CUSCO values can be readily computed for many different assemblies (Figure 6; Figure S18). Although we primarily designed CUSCO for species with piRNA clusters, we showed that the CUSCO approach can be extended to any regions of interest such as *KEE* regions in *A. thaliana* (Figure 6c).

Since CUSCO ignores the actual sequence within the piRNA clusters, complete clusters may yet contain assembly errors, for example if internal regions are missing in the assembly. Therefore, we suggested that the base-coverage heterogeneity and the soft-clip coverage are useful metrics to identify potential assembly problems in piRNA clusters (Figures 5 and 6b). To derive the null expectations for these two metrics, we relied on complete BUSCO genes. Complete BUSCO genes are ideal for this task: first, BUSCO genes are conserved single copy genes, which makes them relatively easy to assemble, even with short reads and a low coverage (Figure 2b). Second, complete BUSCO genes provide a high-confidence set of genes that contain no or few assembly errors (since the ORFs are mostly complete). Third, BUSCO values are usually computed as a standard metric to assess the quality of novel assemblies. The list of complete BUSCO genes is provided as an output of the BUSCO pipeline. Based on the base coverage and the soft-clip coverage, potential assembly errors in a cluster can be identified by coverage values transgressing the quantiles computed from the BUSCO genes (Figure 5; Figure S14). To roughly summarize the assembly quality of each piRNA cluster with representative numbers, we introduced the ScQ and CQ values.

Although the soft-clip coverage of many piRNA clusters approaches the soft-clip coverage of BUSCO genes, the base-coverage heterogeneity of piRNA clusters is always higher than of BUSCO genes, which explains why the CQ values are usually smaller than ScQ values and rarely approach optimal values (i.e. ≥1.0; Figure 4; Figure S19). Repetitive regions, such as found within piRNA clusters, usually lead to alignment problems that may be responsible for the high base-coverage heterogeneity of piRNA clusters. Computing the base-coverage heterogeneity and the soft-clip coverage along piRNA clusters requires long reads (that are then mapped to the assembly), which are usually available anyway when assembling repetitive regions such as piRNA clusters. Furthermore, the CQ and ScQ values depend on complete BUSCO genes to derive the null expectations. In case few BUSCO genes are assembled (i.e. low BUSCO values), the CQ and ScQ values should be interpreted with caution, that is an assembly with high CQ/ScQ values but a low BUSCO is likely of low quality. This emphasizes that our metrics should not be interpreted in isolation but rather be used in combination with classic metrics such as BUSCO and NG50. However, we think the main use of CQ and ScQ values is to identify clusters with potential assembly problems within a given assembly. Finding such outlier clusters is robust to varying numbers of BUSCO genes as the null expectation for computing CQ and ScQ is identical for all clusters within an assembly.

Our novel quality metrics may not only be used to compare the quality of available assemblies but may also serve as a guide during the assembly procedure, for example, to identify the most suitable assembly algorithm. Our metrics should thus help to generate and to identify assemblies having a high fraction of correctly assembled piRNA clusters (or other regions of interest). In unison with standard assembly metrics such as NG50, BUSCO and the total size of assemblies, our metrics should help to generate and identify assemblies with high contiguity and reliability.

4.2 | Assembly strategy

We showed that high-quality assemblies of piRNA clusters can be obtained if: (a) the sequenced strains are isogenic; (b) long reads are available; (c) suitable assemblers, such as Canu are used with an optimized coverage and read length; (d) assemblies are polished using short and long reads; and (e) a scaffolding approach is used. Isogenic strains are necessary to avoid redundant contigs and error-prone assemblies of piRNA clusters (Figure S11; Figure 5). However, it is possible that future tools generate high-quality assemblies of nonisogenic strains. For example, phased assemblers, such as FALCON-phase (Kronenberg et al., 2018), may yield a

separate contig for each homologous chromosome. For these algorithms, segregating polymorphism could even be an advantage as polymorphisms may help to distinguish between homologous chromosomes. We also found that long reads allow us to generate high-quality assemblies of piRNA clusters. Assemblies generated by any of the two major long-read technologies, ONT and PacBio, have a high quality (Figure S18).

We found that Canu yields high-quality assemblies of piRNA clusters and that the TE landscape is most accurately reproduced. The high quality of assemblies generated by Canu was also noticed in several previous works (Jayakumar & Sakakibara, 2017; de Lannoy et al., 2017; Solares et al., 2018; Wick & Holt, 2019). The best assemblies were obtained when solely a subset of the long reads was used for an assembly with Canu, that is 100x coverage with the longest reads. We suspect that this may be related to an algorithmic assumptions about the corrected error rate, which is coverage-dependent and governs the overlap among reads (see Canu manual https://canu.readthedocs.io/en/latest/parameter-reference.html).

Since long reads have a high error rate, polishing of assemblies using long or short reads is usually recommended (Rice & Green, 2018; Sedlazeck, Lee, et al., 2018). Interestingly, polishing also increased the fraction of contiguously assembled piRNA clusters as well as the representation of the TE abundance and diversity (Tables S4 and S5).

Scaffolding with Hi-C slightly increased the number of assembled piRNA clusters (using gapped-CUSCO) but had little influence on the representation of the TE landscape (Figure S7 Table S5). Nevertheless, scaffolding approaches may still be useful for TE research, since scaffolding enables generating chromosome-sized sequences, which could be important when the genomic context of a TE insertion is relevant (e.g. whether a TE or piRNA cluster is close to a telomere).

Despite our optimized assembly strategy, about 19% of the piRNA clusters were not contiguously assembled (Table S5, after polishing). Additionally, manual curation of the final assemblies was necessary to avoid misassemblies (Figure S13). This demonstrates that assembly strategies may still be improved. Especially, promising may be further advances in the length of reads (e.g. by improvements in library preparation protocols), their accuracy (e.g. long high-fidelity reads (Wenger et al., 2019)) and in algorithms generating phased assemblies, which could yield a separate contig for each homologous chromosome. Phased assembly algorithms may even allow us to use outbred strains. Furthermore, such phase assemblers avoid the central problem of assemblies of diploid organisms; that is, that two potentially distinct sequences (i.e. the homologous chromosomes) need to be represented as a single one.

Our novel quality metrics may be used to generate high-quality assemblies of piRNA clusters and thus allow us to address some of the central open questions in TE biology, such as the evolutionary dynamics of piRNA clusters.

ACKNOWLEDGEMENTS

We thank Kirsten-André Senti for advice and providing the *D. melanogaster* strain Iso-1, Christos Vlachos for sharing scripts, Elisabeth Salbaba for technical support and all members of the Institute of Population Genetics for feedback and support. This work was supported by the Austrian Science Foundation (FWF) grants P30036-B25 to RK and W1225.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

RK, FS and FW conceived this work. FS and OC generated the data. FW performed PCR. FS and FW analysed the data. RK and FW provided software. RK, FS and FW wrote the manuscript.

DATA AVAILABILITY STATEMENT

Scripts for computing our quality metrics, including a manual and a walkthrough, are available at https://sourceforge.net/projects/cuscoquality/. We recommend to obtain the scripts via subversion (using the command 'svn checkout https://svn.code.sf.net/p/cuscoquality/code/cuscoquality'). The assemblies of Canton-S and Pi2 and the reads are available at NCBI (PRJNA618654). Tables showing the positions of piRNA clusters and the flanking sequences are available at https://sourceforge.net/projects/cuscoquality/files/CUSCO-data/. The positions of piRNA clusters in our assemblies of Canton-S and Pi2 are available at https://sourceforge.net/projects/cuscoquality/files/publicationdata/piRNA-cluster/. All other scripts used in this work are available at https://sourceforge.net/projects/cuscoquality/files/publicationdata/scripts/.

ORCID

Filip Wierzbicki https://orcid.org/0000-0002-6171-2461
Florian Schwarz https://orcid.org/0000-0002-3683-3974
Robert Kofler https://orcid.org/0000-0001-9960-7248

REFERENCES

Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., Lippman, Z. B., & Schatz, M. C. (2019). RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology*, *20*(1), 224.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Anreiter, I., Kramer, J. M., & Sokolowski, M. B. (2017). Epigenetic mechanisms modulate differences in *Drosophila* foraging behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(47), 12518–12523.

Asif-Laidin, A., Delmarre, V., Laurentie, J., Miller, W. J., Ronsseray, S., & Teysset, L. (2017). Short and long-term evolutionary dynamics of subtelomeric piRNA clusters in *Drosophila*. *DNA Research*, *24*(5), 1–14.

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*(1), 11.

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., & Huala, E. (2015). The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis*, *53*(8), 474–485.

Bergman, C. M., Quesneville, H., Anxolabéhère, D., & Ashburner, M. (2006). Recurrent insertion and duplication generate networks of

transposable element sequences in the *Drosophila melanogaster* genome. *Genome Biology*, *7*(11), R112.

Berlin, K., Koren, S., Chin, C. S., Drake, J. P., Landolin, J. M., & Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, *33*(6), 623–630.

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., & Hannon, G. J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell*, *128*(6), 1089–1103.

Brookfield, J. F., & Badge, R. M. (1997). Population genetics models of transposable elements. *Genetica*, *100*(1–3), 281–294.

Chakraborty, M., Emerson, J. J., Macdonald, S. J., & Long, A. D. (2019). Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nature Communications*, *10*(1), 4872.

Cho, Y. S., Kim, H., Kim, H.-M., Jho, S., Jun, J., Lee, Y. J., Chae, K. S., Kim, C. G., Kim, S., Eriksson, A., Edwards, J. S., Lee, S., Kim, B. C., Manica, A., Oh, T. K., Church, G. M., & Bhak, J. (2016). An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. *Nature Communications*, *7*(1), 1–13.

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W. M., Ritchie, G. R. S., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., … Hubbard, T. (2011). Modernizing reference genome assemblies. *PLoS Biology*, *9*(7), e1001091.

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics*, *34*(15), 2666–2669.

de Lannoy, C., de Ridder, D., & Risse, J. (2017). The long reads ahead: de novo genome assembly using the MinION. *F1000Research*, *6*(1), 1083.

Duc, C., Yoth, M., Jensen, S., Mouniée, N., Bergman, C. M., Vaury, C., & Brasset, E. (2019). Trapping a somatic endogenous retrovirus into a germline piRNA cluster immunizes the germline against further invasion. *Genome Biology*, *20*, 127.

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems*, *3*(1), 95–98.

Earl, D., Bradnam, K., St. John, J., Darling, A., Lin, D., Fass, J., Yu, H. O. K., Buffalo, V., Zerbino, D. R., Diekhans, M., Nguyen, N., Ariyaratne, P. N., Sung, W.-K., Ning, Z., Haimel, M., Simpson, J. T., Fonseca, N. A., Birol, I., Docking, T. R., … Paten, B. (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, *21*(12), 2224–2241.

Ellison, C. E., & Cao, W. (2020). Nanopore sequencing and Hi-C scaffolding provide insight into the evolutionary dynamics of transposable elements and piRNA production in wild strains of *Drosophila melanogaster*. *Nucleic Acids Research*, *48*(1), 1–14.

Ghurye, J., Rhie, A., Walenz, B. P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A. M., & Koren, S. (2019). Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Computational Biology*, *15*(8), 1–19.

Goriaux, C., Théron, E., Brasset, E., & Vaury, C. (2014). History of the discovery of a master locus producing piRNAs: The flamenco/COM locus in *Drosophila melanogaster*. *Frontiers in Genetics*, *5*, 257.

Grob, S., Schmid, M., & Grossniklaus, U. (2014). Hi-C analysis in *Arabidopsis* identifies the *KNOT*, a structure with similarities to the *flamenco* locus of *Drosophila*. *Molecular Cell*, *55*(5), 678–693.

Gunawardane, L. S., Saito, K., Nishida, K. M., Miyoshi, K., Kawamura, Y., Nagami, T., Siomi, H., & Siomi, M. C. (2007). A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*. *Science*, *315*(5818), 1587–1590.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075.

Hickey, D. A. (1982). Selfish DNA: A sexually-transmitted nuclear parasite. *Genetics*, *101*(3–4), 519–531.

Hoskins, R. A., Carlson, J. W., Kennedy, C., Acevedo, D., Evans-Holm, M., Frise, E., Wan, K. H., Park, S., Mendez-Lago, M., Rossi, F., Villasante, A., Dimitri, P., Karpen, G. H., & Celniker, S. E. (2007). Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. *Science*, *316*(5831), 1625–1628.

Hoskins, R. A., Carlson, J. W., Wan, K. H., Park, S., Mendez, I., Galle, S. E., Booth, B. W., Pfeiffer, B. D., George, R. A., Svirskas, R., Krzywinski, M., Schein, J., Accardo, M. C., Damia, E., Messina, G., Méndez-Lago, M., de Pablos, B., Demakova, O. V., Andreyeva, E. N., … Celniker, S. E. (2015). The Release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome Research*, *25*(3), 445–458.

Jayakumar, V., & Sakakibara, Y. (2017). Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. *Briefings in Bioinformatics*, *20*(3), 866–876.

Kelleher, E. S., Azevedo, R. B. R., & Zheng, Y. (2018). The evolution of small-RNA-mediated silencing of an invading transposable element. *Genome Biology and Evolution*, *10*(11), 3038–3057.

Kim, H.-S., Jeon, S., Kim, C., Kim, Y. K., Cho, Y. S., Kim, J., Blazyte, A., Manica, A., Lee, S., & Bhak, J. (2019). Chromosome-scale assembly comparison of the Korean Reference Genome KOREF from PromethION and PacBio with Hi-C mapping information. *GigaScience*, *8*(12), giz125.

Kofler, R. (2019). Dynamics of transposable element invasions with piRNA clusters. *Molecular Biology and Evolution*, *36*(7), 1457–1472.

Kofler, R. (2020). piRNA clusters need a minimum size to control transposable element invasions. *Genome Biology and Evolution*, *12*(5), 736–749.

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., Kosiol, C., & Schlötterer, C. (2011). PoPoolation: A toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One*, *6*(1), e15925. https://doi.org/10.1371/journal.pone.0015925

Kofler, R., Pandey, R. V., & Schlötterer, C. (2011). PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, *27*(24), 3435–3436. https://doi.org/10.1093/bioinformatics/btr589

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736.

Kronenberg, Z. N., Hall, R. J., Hiendleder, S., Smith, T. P. L., Sullivan, S. T., Williams, J. L., & Kingan, S. B. (2018). FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes. *bioRxiv*. https://doi.org/10.1101/327064

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, *5*(2), R12.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A., & Huala, E. (2012). The *Arabidopsis* Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Research*, *40*, D1202–D1210.

Le Thomas, A., Rogers, A. K., Webster, A., Marinov, G. K., Liao, S. E., Perkins, E. M., Hur, J. K., Aravin, A. A., & Tóth, K. F. (2013). Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes and Development*, *27*(4), 390–399.

Lewis, S. H., Quarles, K. A., Yang, Y., Tanguy, M., Frézal, L., Smith, S. A., Sharma, P. P., Cordaux, R., Gilbert, C., Giraud, I., Collins, D. H., Zamore, P. D., Miska, E. A., Sarkies, P., & Jiggins, F. M. (2018). Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence

against transposable elements. *Nature Ecology and Evolution*, *2*(1), 174–181.

Li, H. (2016). Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, *32*(14), 2103–2110.

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

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760.

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

Lieberman-Aiden, E., Berkum, N. L. V., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, B. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, *326*(5950), 289–293.

Malone, C. D., & Hannon, G. J. (2009). Small RNAs as guardians of the genome. *Cell*, *136*(4), 656–668.

Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982). *Molecular cloning: A laboratory manual* (Vol. *545*). Cold Spring Harbor Laboratory.

McCoy, R. C., Taylor, R. W., Blauwkamp, T. A., Kelley, J. L., Kertesz, M., Pushkarev, D., Petrov, D. A., & Fiston-Lavier, A.-S. (2014). Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PLoS One*, *9*(9), e106689.

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G., Schneider, V. A., Potapova, T., Wood, J., Chow, W., Armstrong, J., Fredrickson, J., Pak, E., Tigyi, K., Kremitzki, M., … Phillippy, A. M. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, *585*(7823), 79–84.

Minoche, A. E., Dohm, J. C., & Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biology*, *12*(11), R112.

Mohamed, M., Dang, N.-T.-M., Ogyama, Y., Burlet, N., Mugat, B., Boulesteix, M., Vincent, M., Veber, P., Salces-ortiz, J., Severac, D., Pélisson, A., Vieira, C., Sabot, F., Fablet, M., & Chambeyron, S. (2020). A transposon story: From TE content to TE dynamic Invasion of *Drosophila* genomes using the single-molecule sequencing technology from Oxford Nanopore. *Cells*, *9*(8), 1776.

O'Hare, K., Driver, A., McGrath, S., & Johnson-Schiltz, D. M. (1992). Distribution and structure of cloned P elements from the *Drosophila melanogaster* P strainπ2. *Genetical Research*, *60*(1), 33–41.

O'Hare, K., & Rubin, G. M. (1983). Structures of P transposable elements and their sites of insertion and excision in the *Drosophila melanogaster* genome. *Cell*, *34*(1), 25–35.

Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D., & Zamore, P. D. (2019). PIWI-interacting RNAs: Small RNAs with big functions. *Nature Reviews Genetics*, *20*(2), 89–108.

Pryszcz, L. P., & Gabaldón, T. (2016). Redundans: An assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research*, *44*(12), e113.

Pucker, B., Holtgrawe, D., Sorensen, T. R., Stracke, R., Viehover, P., & Weisshaar, B. (2016). A de novo genome sequence assembly of the *Arabidopsis thaliana* accession niederzenz-1 displays presence/absence variation and strong synteny. *PLoS One*, *11*(10), e0164321.

Quesneville, H., Bergman, C. M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., & Anxolabéhère, D. (2005). Combined evidence annotation of transposable elements in genome sequences. *PLoS Computational Biology*, *1*(2), 166–175.

R Core Team. (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Rice, E. S., & Green, R. E. (2018). New approaches for genome assembly and scaffolding. *Annual Review of Animal Biosciences*, *7*(1), 17–40.

Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, *17*(2), 155–158.

Sarkar, A., Maji, R. K., Saha, S., & Ghosh, Z. (2014). piRNAQuest: searching the piRNAome for silencers. *BMC Genomics*, *15*(1), 555.

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, W., Collins, J., … Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, *27*(5), 849–864.

Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018). Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, *19*(6), 329–346. https://doi.org/10.1038/s41576-018-0003-4

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*(6), 461–468. https://doi.org/10.1038/s41592-018-0001-7

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing genome assembly and annotation completeness. *Methods in Molecular Biology*, *1962*, 227–245.

Shumate, A., Zimin, A. V., Sherman, R. M., Puiu, D., Wagner, J. M., Olson, N. D., Pertea, M., Salit, M. L., Zook, J. M., & Salzberg, S. L. (2020). Assembly and annotation of an Ashkenazi human reference genome. *Genome Biology*, *21*(1), 1–18.

Sienski, G., Dönertas, D., & Brennecke, J. (2012). Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell*, *151*(5), 964–980.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117–1123.

Singhal, K., Khanna, R., & Mohanty, S. (2017). Is *Drosophila*-microbe association species-specific or region specific? A study undertaken involving six Indian *Drosophila* species. *World Journal of Microbiology and Biotechnology*, *33*(6), 103.

Smit, A. F. A., Hubley, R., & Green, P. (2013–2015). *RepeatMasker Open-4.0*.

Solares, E. A., Chakraborty, M., Miller, D. E., Kalsow, S., Hall, K., Perera, A. G., Emerson, J. J., & Hawley, R. S. (2018). Rapid low-cost assembly of the *Drosophila melanogaster* reference genome using low-coverage, long-read sequencing. *G3: Genes, Genomes, Genetics*, *8*(10), 3143–3154.

Sović, I., Križanović, K., Skala, K., & Šikić, M. (2016). Evaluation of hybrid and non-hybrid methods for de novo assembly of nanopore reads. *Bioinformatics*, *32*(17), 2582–2589.

Srivastav, S. P., Rahman, R., Ma, Q., Pierre, J., Bandyopadhyay, S., & Lau, N. C. (2019). Har-P, a short P-element variant, weaponizes p-transposase to severely impair *Drosophila* development. *eLife*, *8*, e49948.

Sullivan, M. J., Petty, N. K., & Beatson, S. A. (2011). Easyfig: A genome comparison visualizer. *Bioinformatics*, *27*(7), 1009–1010.

Vaser, R., Sovic, I., Nagarajan, N., & Sikic, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746.

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

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial

variant detection and genome assembly improvement. *PLoS One*, *9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, *35*(3), 543–548.

Weilguny, L., & Kofler, R. (2019). DeviaTE: Assembly-free analysis and visualization of mobile genetic element composition. *Molecular Ecology Resources*, *19*(5), 1346–1354.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C. S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, J. R., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162.

Wick, R. R., & Holt, K. E. (2019). Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Research*, *8*, 2138.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer Nature.

Yamanaka, S., Siomi, M. C., & Siomi, H. (2014). piRNA clusters and open chromatin structure. *Mobile DNA*, *5*(1), 22.

Yang, P., Wang, Y., & Macfarlan, T. S. (2017). The role of KRAB-ZFPs in transposable element repression and mammalian evolution. *Trends in Genetics*, *33*(11), 871–881.

Zanni, V., Eymery, A., Coiffet, M., Zytnicki, M., Luyten, I., Quesneville, H., Vaury, C., & Jensen, S. (2013). Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(49), 19842–19847.

Zapata, L., Ding, J., Willing, E. M., Hartwig, B., Bezdan, D., Jiao, W. B., Patel, V., James, G. V., Koornneef, M., Ossowski, S., & Schneeberger, K. (2016). Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(28), E4052–E4060.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Wierzbicki, F., Schwarz, F., Cannalonga, O., & Kofler, R. (2021). Novel quality metrics allow identifying and generating high-quality assemblies of piRNA clusters. *Molecular Ecology Resources*, 00, 1–20. https://doi.org/10.1111/1755-0998.13455

Chapter 2

Tirant Stealthily Invaded Natural *Drosophila melanogaster* Populations during the Last Century

Florian Schwarz [1,2] Filip Wierzbicki,[1,2] Kirsten-André Senti,[†,1] and Robert Kofler [*,1]

[1]Institut für Populationsgenetik, Vetmeduni Vienna, Vienna, Austria

[2]Vienna Graduate School of Population Genetics, Vetmeduni Vienna, Vienna, Austria

[†]Present address: Institute of Molecular Biotechnology of the Austrian Academy of Sciences, Vienna Biocenter, Vienna, Austria

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

Associate editor: John True

Abstract

It was long thought that solely three different transposable elements (TEs)—the I-element, the P-element, and hobo—invaded natural *Drosophila melanogaster* populations within the last century. By sequencing the "living fossils" of *Drosophila* research, that is, *D. melanogaster* strains sampled from natural populations at different time points, we show that a fourth TE, Tirant, invaded *D. melanogaster* populations during the past century. Tirant likely spread in *D. melanogaster* populations around 1938, followed by the I-element, hobo, and, lastly, the P-element. In addition to the recent insertions of the canonical Tirant, *D. melanogaster* strains harbor degraded Tirant sequences in the heterochromatin which are likely due to an ancient invasion, likely predating the split of *D. melanogaster* and *D. simulans*. These degraded insertions produce distinct piRNAs that were unable to prevent the novel Tirant invasion. In contrast to the I-element, P-element, and hobo, we did not find that Tirant induces any hybrid dysgenesis symptoms. This absence of apparent phenotypic effects may explain the late discovery of the Tirant invasion. Recent Tirant insertions were found in all investigated natural populations. Populations from Tasmania carry distinct Tirant sequences, likely due to a founder effect. By investigating the TE composition of natural populations and strains sampled at different time points, insertion site polymorphisms, piRNAs, and phenotypic effects, we provide a comprehensive study of a natural TE invasion.

Key words: **transposable elements, *Drosophila melanogaster*, transposon invasions, next-generation sequencing, Tirant, P-element, I-element, hobo.**

Introduction

Transposable elements (TEs) are DNA sequences that multiply within host genomes, even if this activity is deleterious to hosts (Doolittle and Sapienza 1980; Orgel and Crick 1980; Hickey 1982; Wicker et al. 2007). To enhance their rate of transmission into the next generation, TEs need to infect the germ cells. Although most TEs achieve this by being active in the germline, some LTR retrotransposons generate virus-like particles in the somatic follicle cells surrounding the germline, which may infect the germ cells (Song et al. 1997; Blumenstiel 2011; Goodier 2016; Moon et al. 2018; Wang et al. 2018). Since many TE insertions are deleterious, host organisms evolved elaborate defense mechanisms against TEs (Brennecke et al. 2007; Marí-Ordóñez et al. 2013; Yang et al. 2017). In *Drosophila melanogaster*, the defense against TEs is based on piRNAs (PIWI-interacting RNAs), that is, small RNAs with a size between 23–29nt, that repress TE activity at the transcriptional and the posttranscriptional level (Brennecke et al. 2007; Gunawardane et al. 2007; Sienski et al. 2012; Le Thomas et al. 2013). piRNAs are derived from distinct genomic loci termed piRNA clusters (Brennecke et al. 2007). Different piRNA pathways are active in the germline and in the follicle cells surrounding the germline (Li, Vagin, et al. 2009; Malone et al. 2009), where solely the germline pathway depends on maternally transmitted piRNAs for efficient silencing of TEs (Le Thomas et al. 2014).

One option to escape the host defense is to infect a novel species. Many TEs cross species boundaries, for example, due to horizontal transfer (HT) from one host species to another, and trigger invasions in naive species not having the TE (Mizrokhi and Mazo 1990; Maruyama and Hartl 1991; Lohe et al. 1995; Terzian et al. 2000; Sánchez-Gracia et al. 2005; Loreto et al. 2008; Kofler, Hill, et al. 2015; Peccoud et al. 2017). A striking example for a high frequency of TE invasions can be seen in *D. melanogaster*, which was invaded by at least three different TE families within the last century: the I-element, hobo, and the P-element (Kidwell 1983; Anxolabéhère et al. 1988; Periquet et al. 1989; Daniels, Chovnick, et al. 1990; Daniels, Peterson, et al. 1990; Bucheton et al. 1992; Bonnivard et al. 2000). All of these three TEs actively replicate only in the germline and induce some phenotypic effects, the hybrid dysgenesis (HD) symptoms, which historically led to the discovery of the recent TE invasions in *D. melanogaster* (Bingham et al. 1982; Calvi and Gelbart 1994; Biémont 2010; Moon et al. 2018; Wang et al. 2018). An important hallmark of these HD symptoms is that the direction of

crosses between two strains is important. The offspring of crosses between males carrying a genomic factor (the TE) and females not carrying this factor frequently show various symptoms (e.g., atrophic ovaries) whereas the offspring of the reciprocal crosses is usually free of symptoms (Bucheton et al. 1976; Kidwell et al. 1977; Blackman et al. 1987; Yannopoulos et al. 1987). Hence, hybrid dysgenesis has a cytoplasmic as well as a genomic component.

Although TEs were quickly identified as the responsible genomic factor, the cytoplasmic component, the maternally transmitted piRNAs, was discovered much later (Bingham et al. 1982; Brennecke et al. 2008). It was realized that the presence of an HD-inducing TE in a strain mostly depends on the sampling date of a strain, where more recently sampled strains frequently carry the TE while old strains, sampled before the invasion, do not. It was thus suggested that the HD-inducing TEs recently invaded *D. melanogaster* populations (Kidwell 1983; Periquet et al. 1994). These invasions were probably triggered by HT events, where the P-element was likely acquired from *D. willistoni* and the I-element as well as hobo possibly from *D. simulans* (or another species from the *simulans* clade) (Daniels, Chovnick, et al. 1990; Daniels, Peterson, et al. 1990; Simmons 1992; Loreto et al. 2008; Blumenstiel 2019). However, even the old strains carried short and highly degraded (probably inactive) fragments of the I-element and hobo, mostly in the heterochromatin (Bucheton et al. 1984, 1986, 1992; Daniels, Chovnick, et al. 1990). Hence, the I-element and hobo likely invaded *D. melanogaster* populations at least twice. Solely the P-element does not have substantial similarity to sequences in the *D. melanogaster* genome, which suggests that the P-element invaded *D. melanogaster* populations for the first time. *Drosophila melanogaster* strains sampled at different time points, previously labeled as the "living fossils" of *Drosophila* research (Bucheton et al. 1992), were not only used to discover the three recent TE invasions but also to estimate the timing of the invasions: the I-element invasion occurred presumably between 1930 and 1950, the hobo invasion around 1955 and the P-element invasion between 1950 and 1980 (Kidwell 1983; Anxolabéhère et al. 1988; Periquet et al. 1989).

By sequencing these "living fossils," we discovered that an additional transposon, Tirant, invaded *D. melanogaster* populations within the last century. Previous work showed that Tirant is an LTR retrotransposon and a member of the *Ty3/Gypsy* superfamily (Moltó et al. 1996; Viggiano et al. 1997; Cañizares et al. 2000; Terzian et al. 2001). It encodes an envelope protein and completes the retroviral cycle in the closely related *D. simulans* (Lemeunier et al. 1976; Marsano et al. 2000; Akkouche et al. 2012). In contrast to the P-element, hobo, and the I-element, which are active in the germline, Tirant was classified as an intermediate TE based on the amount of maternally transmitted piRNAs, that is, Tirant is likely expressed and targeted in both the germline and in somatic follicle cells (Malone et al. 2009). In agreement with this, Tirant activity was reported in both tissues (Akkouche et al. 2012). Furthermore, knockdowns of components of the germline as well as the somatic piRNA pathway, result in a reduction of Tirant piRNAs (Nefedova et al. 2012; Czech et al. 2013; Rozhkov et al. 2013; Barckmann et al. 2018). Generally, intermediate TEs are little understood. However, for Tirant in particular, peculiarities in the regulation were noted (Akkouche et al. 2013; Parhad et al. 2017; Wang et al. 2020). For example, in some backgrounds Tirant may be upregulated independent of piRNAs (Parhad et al. 2017).

Fablet et al. (2007) suggested that Tirant is an ancient TE that is largely vertically transmitted in the *D. melanogaster* species subgroup. Analyses of the reference genome of *D. melanogaster* revealed the presence of degraded Tirant insertions in the heterochromatin and full-length insertions in the euchromatin (Bowen and McDonald 2001; Mugnier et al. 2008). The heterochromatic insertions are likely ancient, possibly predating the split of *D. melanogaster* and *D. simulans*, whereas the euchromatic insertions are likely more recent (<16,000–200,000 years) (Bowen and McDonald 2001; Bergman and Bensasson 2007; Mugnier et al. 2008). This raises the question on how this uneven age distribution of Tirant insertions evolved.

Here, we show that full-length (canonical) Tirant sequences are absent from laboratory strains sampled before 1938 but present in strains sampled after 1938. We thus suggest that the canonical Tirant invaded natural *D. melanogaster* populations between 1930 and 1950, possibly following an HT from *D. simulans*. This invasion constitutes a second wave of activity, with degraded heterochromatic fragments being the remnants of an ancient Tirant invasion, possibly in the ancestor of the *D. melanogaster* species subgroup. Tirant is thus the fourth TE to invade *D. melanogaster* populations within the last century. Based on a consistent approach (i.e., the same method and strains) for all four TEs, we estimate that Tirant invaded *D. melanogaster* populations first, followed by the I-element, hobo and, finally, the P-element. Recent Tirant insertions were found in all investigated natural populations, where populations from Tasmania carry distinct Tirant sequences, likely due to a founder effect.

Although all strains carry piRNAs complementary to the degraded Tirant insertions solely recently invaded strains carry piRNAs complementary to the canonical Tirant. We thus suggest that piRNAs complementary to heterochromatic insertions were too diverged to prevent the spread of the canonical Tirant. Finally, we did not find apparent HD symptoms induced by Tirant, which may account for the late discovery of the Tirant invasion. By investigating the TE composition (i.e., abundance of TEs and frequency of internal deletions and SNPs) of natural populations and strains sampled at different time points, insertion site polymorphisms, piRNAs, and phenotypic effects, we provide a comprehensive study of a natural TE invasion.

Results

Canonical Tirant Insertions Are Present in Iso-1 but Not in Canton-S

Given the striking accumulation of TE invasions within the last century (Kidwell 1983; Anxolabéhère et al. 1988; Periquet et al. 1989; Daniels, Chovnick, et al. 1990; Daniels, Peterson, et al. 1990; Bucheton et al. 1992; Bonnivard et al. 2000), we

speculated that additional, hitherto undetected TEs, may have recently invaded *D. melanogaster* populations.

To test this hypothesis, we compared the abundance of TEs between one of the oldest available *D. melanogaster* laboratory strains, Canton-S (collected by C. Bridges in 1935; Lindsley and Grell 1968) and the reference strain, Iso-1 (fig. 1*A*; Brizuela et al. 1994). We aligned publicly available short-read data from these strains to the consensus sequences of TEs in *D. melanogaster* (Quesneville et al. 2005) and estimated the normalized abundance (reads per million) of the TEs in these two strains with our novel tool DeviaTE (Weilguny and Kofler 2019). Apart from the telomeric TEs (TART-A, TART-B, and TAHRE) which show distinct evolutionary dynamics (Pardue and DeBaryshe 2011; Saint-Leandre and Levine 2020), the most striking difference between the two strains was due to the LTR retrotransposon Tirant (fig. 1*A*). As expected, hobo and the I-element, two TEs that invaded *D. melanogaster* recently, are more abundant in the Iso-1 strain than in the older Canton-S strain (fig. 1*A*). The P-element is not present in both strains. To further investigate the abundance of Tirant in the two strains, we calculated the coverage of reads along the Tirant sequence with DeviaTE (fig. 1*B*; Weilguny and Kofler 2019). We observed striking coverage differences between Canton-S and Iso-1 over the entire sequence of Tirant (fig. 1*B*; average normalized coverage; Iso-1 = 20.9, Canton-S = 0.86). Only few highly diverged reads aligned to Tirant in Canton-S (fig. 1*B*). In addition to these diverged reads, many reads with a high similarity to the consensus sequence of Tirant aligned in Iso-1 (fig. 1*B*). We refer to Tirant sequences with a high similarity to the consensus sequence as "canonical" Tirant. To identify the genomic location of the canonical and the diverged Tirant sequences, we annotated TEs in publicly available assemblies of Canton-S (based on Oxford Nanopore long-read data) and Iso-1 (i.e., the reference genome) with RepeatMasker (fig. 1*C*; Hoskins et al. 2015; Wierzbicki et al. 2020). Both assemblies are of high quality and suitable for genomic analysis of TEs (Wierzbicki et al. 2020). In Canton-S, only highly fragmented and diverged Tirant sequences were found close to the centromeres (fig. 1*C* and supplementary fig. 1, Supplementary Material online). In addition to these diverged Tirant sequences, Iso-1 carries several canonical Tirant insertions on each chromosome arm (fig. 1*C*). This genomic distribution of Tirant, that is, degraded Tirant fragments in the heterochromatin and canonical insertions in the euchromatin of *D. melanogaster*, was also noted in previous studies (Marsano et al. 2000; Mugnier et al. 2008). The absence of canonical Tirant insertions in euchromatin is also found in an independent assembly of Canton-S which is based on PacBio reads (supplementary fig. 2, Supplementary Material online; Chakraborty et al. 2019). It was proposed that the degraded Tirant insertions located in heterochromatin are ancient and likely vertically inherited from the ancestor of the *D. melanogaster* species subgroup (Moltó et al. 1996; Fablet et al. 2007; Mugnier et al. 2008). It was further proposed that canonical insertions in Iso-1 are of more recent origin (i.e., <16,000–200,000 years (Bowen and McDonald 2001; Bergman and Bensasson 2007; Lerat et al. 2011; Rahman et al. 2015). We thus speculated that the canonical insertions of Tirant may have recently been active, whereas the degraded insertions in the heterochromatic may be inactive for some time (see also, Mugnier et al. 2008; Fablet et al. 2009). If this is true, canonical insertions ought to segregate at low frequency in natural populations, whereas the degraded insertions should mostly be fixed. To test this hypothesis, we estimated the population frequencies of the canonical and the degraded Tirant insertions in a natural *D. melanogaster* population from France (Viltain) (Kapun et al. 2020) with PoPoolationTE2 (Kofler et al. 2016). Indeed, most canonical Tirant insertions segregate at a low population frequency ($f = 0.063$) in the euchromatin, whereas most degraded insertions are in the heterochromatin and segregate at significantly higher frequencies ($f = 0.73$; Wilcoxon rank sum test $P < 2.2e{-}16$; supplementary fig. 3, Supplementary Material online). Due to relaxed purifying selection in low-recombining regions (Eanes et al. 1992; Sniegowski and Charlesworth 1994; Bartolomé et al. 2002; Petrov et al. 2011; Kofler et al. 2012), degraded Tirant insertions may have accumulated in the heterochromatin. Taken together, we hypothesize that Tirant invaded natural *D. melanogaster* populations in at least two waves of activity: an ancient wave, possibly predating the split of *D. melanogaster* and *D. simulans*, and a recent wave after Canton-S was sampled.

Canonical Tirant Invaded *D. melanogaster* Populations between 1930 and 1950

If Tirant invaded natural *D. melanogaster* populations recently, old strains should only have a few highly degraded Tirant sequences (similar to Canton-S), whereas more recently collected strains should have many insertions with a high similarity to the consensus sequence of Tirant (i.e., canonical Tirant insertions). To test this, we sequenced 12 of the oldest available *D. melanogaster* strains (sampled between 1920 and 1970; fig. 2; supplementary table 1, Supplementary Material online). Additionally, we included publicly available data of 15 different *D. melanogaster* strains into the analyses (fig. 2*A* and supplementary table 1, Supplementary Material online). The reads were mapped to the consensus sequences of TEs in *Drosophila* and the TE abundance was assessed with DeviaTE (supplementary fig. 4, Supplementary Material online; Weilguny and Kofler 2019).

Strikingly, six out of seven strains sampled before or in 1938 solely contained degraded Tirant sequences (supplementary table 1 and fig. 4, Supplementary Material online). The first strain carrying canonical Tirant sequences (Urbana-S) was collected around 1938. All 16 strains collected around or after 1950 carried canonical Tirant sequences (supplementary table 1, Supplementary Material online). Estimates of the TE copy numbers support these observations (fig. 2*A*). To obtain estimates of the TE abundance independent of DeviaTE, we also computed the normalized number of reads mapping to each TE (rpm; reads per million). These data also support the sudden increase in reads mapping to Tirant in strains sampled after 1938 (supplementary table 2, Supplementary Material online). We note that the raw abundance of reads mapping

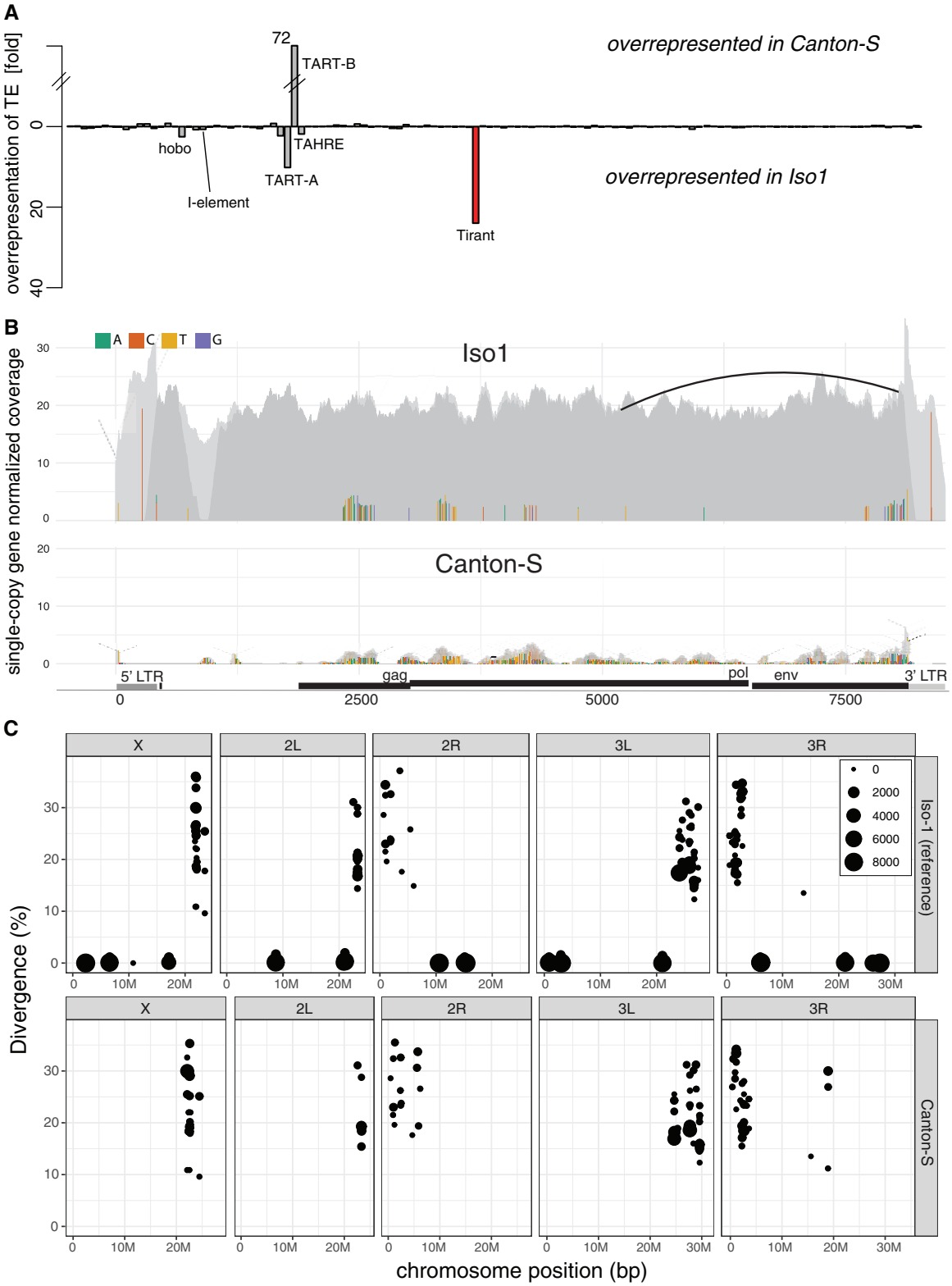

Fig. 1. Canonical Tirant insertions are present in Iso-1 but not Canton-S. (*A*) Differences in TE content between Canton-S and Iso-1. For each TE family (*x* axis), we show the fold-difference in the number of reads mapping to a TE (*y* axis) between the two strains. Note that reads mapping to Tirant (red) are overrepresented in Iso-1. (*B*) Abundance and diversity of Tirant in Iso-1 and Canton-S. Short reads were aligned to the consensus sequence of Tirant and visualized with DeviaTE. The coverage of Tirant was normalized to the coverage of single-copy genes. Single-nucleotide polymorphisms (SNPs) and small internal deletions (indels) are shown as colored lines. Large internal deletions are shown as black arcs (the frequency of the shown deletion is ≈5%). Coverage based on uniquely and ambiguously mapped reads is shown in dark and light gray, respectively. Note that solely a few, highly degraded copies of Tirant are present in Canton-S. (*C*) Overview of Tirant insertions in the genomes of Iso-1 and Canton-S. For each Tirant insertion, we show the position in the assembly, the length (size of dot), and the similarity to the consensus sequence (divergence).

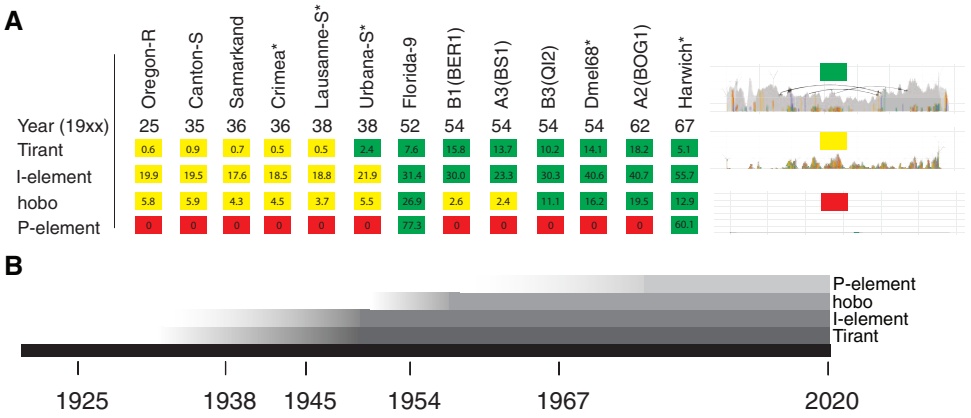

FIG. 2. History of recent TE invasions in natural *Drosophila melanogaster* populations. (*A*) Overview of Tirant, I-element, hobo, and P-element sequences in some *D. melanogaster* strains. For an overview of these TEs in all investigated strains, see supplementary table 1, Supplementary Material online. The year corresponds to the sampling date of the strain. Strains sequenced in this work are marked by a star (*). For each family, we classified the TE content into three distinct categories: red, absence of any TE sequence; yellow, solely degraded TE sequences are present; green, nondegraded sequences, with a high similarity to the consensus sequence are present; Numbers represent estimates of TE copy numbers per haploid genome obtained with DeviaTE. The abundance of degraded copies may be underestimated as copy-number estimates are based on the average coverage of the consensus sequence. The right panel shows an example for each of the three categories (similar to fig. 1*B*). (*B*) Timeline showing the estimated invasion history of Tirant, the I-element, hobo, and the P-element.

to a TE is highly correlated with the estimates of TE abundance obtained with DeviaTE (supplementary fig. 5, Supplementary Material online). Our results thus suggest that the canonical Tirant invaded *D. melanogaster* populations between 1938 and 1950 (fig. 2). Since we were interested in the timing of the Tirant invasion relative to the other three TEs that recently invaded *D. melanogaster* populations, we also investigated the abundance and diversity of the I-element, hobo, and the P-element in these strains (supplementary table 1 and figs. 6–8, Supplementary Material online). Our data suggest that Tirant invaded natural *D. melanogaster* populations just before the I-element, followed by hobo and, lastly, by the P-element (supplementary tables 1 and 2, Supplementary Material online and fig. 2*B*).

Canonical Tirant Insertions Are Found in Worldwide Populations of *D. melanogaster* and Populations from Tasmania Carry Distinct Tirant Variants

To further investigate the Tirant composition among strains, we performed a PCA based on the allele frequencies of Tirant single-nucleotide polymorphism (SNPs) (fig. 3). Note that our usage of the term SNP is not strictly identical to the common usage describing allelic variants at a single locus. Here, a SNP describes a variant among dispersed Tirant copies. Our allele frequency estimates thus reflect the Tirant composition within a particular strain (e.g., if 14 Tirant insertions in a given strain carry an "A" at some site and 6 a "T," the frequency of "A" at this site is 0.7). In addition to the above-mentioned strains (supplementary table 1, Supplementary Material online), we also analyzed the Tirant content of natural populations. To do this, we relied on the global diversity lines (GDL), that is, several *D. melanogaster* strains sampled after 1988 (Begun and Aquadro 1995) from five different continents (Africa—Zimbabwe, Asia—Beijing, Australia—Tasmania, Europe—Netherlands, America—Ithaca; Grenier et al. 2015).

Old strains, collected before 1938, formed a distinct group (fig. 3), supporting our view that they carry distinct Tirant sequences. By contrast, most strains collected after 1938 and the majority of the GDLs group into one large cluster (fig. 3). All GDL strains thus carry nondegraded Tirant sequences. This observation also holds when additional, recently collected *D. melanogaster* strains are analyzed (e.g., DGRP, DrosEU, DrosRTEC; supplementary fig. 9, Supplementary Material online; Mackay et al. 2012; Bergland et al. 2014; Lack et al. 2015; Machado et al. 2019; Kapun et al. 2020). Old strains also form a distinct group in an unrooted tree computed from pairwise F_{ST} values based on the frequency of Tirant SNPs (supplementary fig. 10, Supplementary Material online). Our data thus suggests that Tirant invaded most worldwide *D. melanogaster* populations. The reference strain Iso-1 is distant to the large cluster (fig. 3). Closer inspection revealed that Tirant insertions from natural populations carry eight SNPs that are not found in the reference strain (supplementary fig. 11 and table 3, Supplementary Material online). Interestingly, also strains collected from Tasmania (Australia) formed a distinct group (fig. 3 and supplementary fig. 10, Supplementary Material online). We hypothesized that this is due to multiple SNPs having markedly different allele frequencies in Tasmanian populations than in populations from other geographic locations (supplementary fig. 12 and table 4, Supplementary Material online). Indeed, when excluding those SNPs from the PCA, strains from Tasmania clustered with strains sampled from the other locations (supplementary fig. 13, Supplementary Material online). For hobo and the I-element, Tasmanian populations did not form a separate cluster (supplementary fig. 14, Supplementary Material online; the P-element is absent in many samples, hence allele frequencies could not be calculated). This raises the question of what processes could be responsible for such striking differences in the Tirant composition among natural

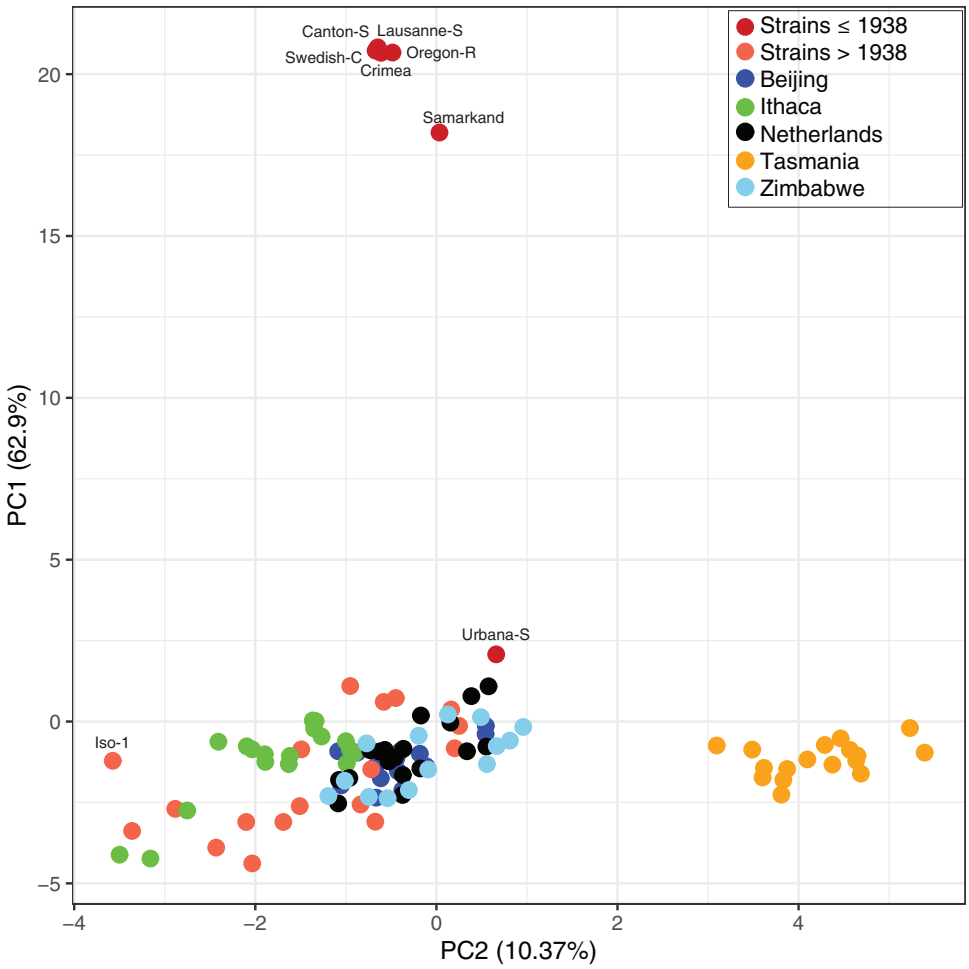

Fig. 3. PCA based on the allele frequencies of Tirant SNPs in different *Drosophila melanogaster* strains. In addition to previously described *D. melanogaster* strains, the Global Diversity Lines (GDL) were analyzed. Note that the strains without canonical Tirant insertions as well as populations from Tasmania form distinct groups.

populations. We suggest that the Tirant invasion in Tasmania was subject to a founder effect, where flies carrying some rare variants of Tirant migrated to Tasmania, thereby triggering the spread of these rare Tirant variants in Tasmanian populations. Similarly, the strains used for generating Iso-1 may have carried rare Tirant variants that multiplied in these lines after they were sampled. In agreement with this, most Iso-1 specific SNPs segregate at low frequency in some *D. melanogaster* populations from Europe and North America (supplementary fig. 11, Supplementary Material online).

In summary, we conclude that Tirant invaded all investigated worldwide populations of *D. melanogaster* during the past century. Furthermore, founder effects may be important components of TE invasions, since they may lead to a geographically heterogeneous TE composition.

The Canonical Tirant Is Silenced by a piRNA-Based Defense Mechanism

If Tirant recently invaded *D. melanogaster* populations, we expect to see differences in the composition of piRNAs between strains sampled before and after the invasion. Strains invaded by Tirant, such as Iso-1, should have established a functional defense against the TE and thus generate large amounts of piRNAs complementary to canonical Tirant. By contrast, naive strains, such as Canton-S, should have few canonical Tirant piRNAs. To test this, we sequenced piRNAs from the ovaries of both strains. Indeed, piRNAs against canonical Tirant were highly abundant in Iso-1 but not in Canton-S (fig. 4*A* and *D*). Compared with the piRNA abundance of other TE families in *D. melanogaster*, Tirant piRNAs rank among the most abundant in Iso-1 but the least abundant in Canton-S (fig. 4*A*). Both sense and antisense piRNAs are distributed over the entire sequence of Tirant in Iso-1 (fig. 4*B*). TEs that are silenced in the germline by dual-strand clusters show a characteristic 10 nt overlap between sense and antisense piRNAs, that is, the ping-pong signature (Brennecke et al. 2007; Malone et al. 2009). Tirant has a pronounced ping-pong signature in Iso-1 but not in Canton-S (fig. 4*C*), consistent with Tirant being silenced in the germline (likely in addition to the soma) (Malone et al. 2009). Finally, we wondered whether the ancient Tirant invasion, responsible for the degraded Tirant fragments in the heterochromatin, led to piRNAs against Tirant. Both Iso-1 and Canton-S, carry piRNAs complementary to the degraded Tirant fragments (6252.0 ppm in Canton-S and

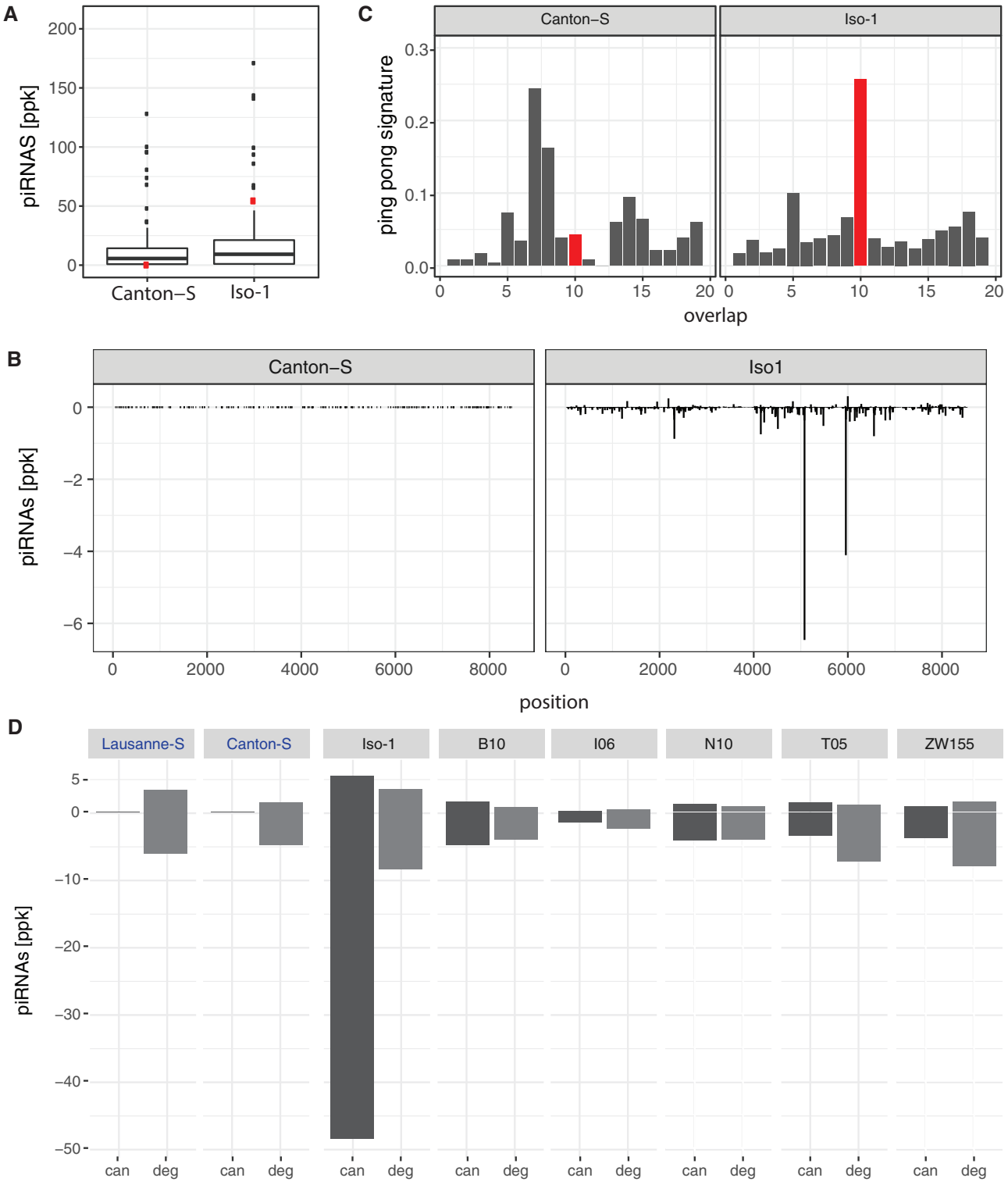

FIG. 4. Tirant piRNAs in strains with (e.g., Iso-1) and without (e.g., Canton-S) canonical Tirant insertions. (*A*) Abundance of canonical Tirant piRNAs (red) compared with piRNA of the other TEs of *Drosophila melanogaster*. (*B*) Abundance of piRNAs along the canonical sequence of Tirant. (*C*) Ping-pong signature for canonical Tirant piRNAs. A pronounced peak at position 10 (red) suggests secondary amplification of piRNAs by the ping-pong cycle. (*D*) Abundance of piRNAs complementary to canonical (can; dark gray) and degraded (deg; light gray) Tirant sequences in laboratory strains (Lausanne-S, Canton-S, and Iso-1) and GDL lines (B10, Beijing; I06, Ithaca; N10, Netherlands; T05, Tasmania; ZW155, Zimbabwe; Grenier et al. 2015). The names of two strains not having canonical Tirant insertions are shown in blue. Sense piRNAs are on the positive *y* axis and antisense piRNAs on the negative *y* axis (*B* and *D*). ppk, piRNAs per 1000 miRNAs.

11886.0 ppm in Iso-1; fig. 4*D*). An analysis of the piRNA content of additional strains (Lausanne-S and GDL lines; Luo et al. 2020) confirms that all investigated strains carry piRNAs complementary to the degraded Tirant whereas only strains with canonical Tirant insertions carry piRNAs complementary to the canonical Tirant (fig. 4*D*). This raises the question why these piRNAs of the degraded Tirant were unable to prevent the invasion of the canonical Tirant. Previous works suggest that piRNAs need to match over the bulk of a sequence with a sequence divergence of less than 10% for efficient silencing of the target sequence (Post et al. 2014; Kotov et al. 2019). Heterochromatic Tirant sequences, however, are about 10–30% diverged from the canonical Tirant (supplementary fig. 1, Supplementary Material online). The high divergence can be found over the entire sequence of these Tirant fragments (supplementary fig. 15, Supplementary Material online). Consequently very few of the degraded piRNAs match to the canonical Tirant with a sequence divergence of less than 10% (supplementary fig. 16, Supplementary Material online).

We conclude that a piRNA-based defense mechanism against the canonical Tirant is present in all strains carrying canonical Tirant insertions but absent in strains solely having heterochromatic Tirant insertions. Although piRNAs derived from these heterochromatic Tirant fragments are present in all strains, these piRNAs were likely too diverged to silence the canonical Tirant and therefore could not prevent its recent invasion.

No Apparent Hybrid Dysgenesis Symptoms Can Be Found for Tirant

The other three TEs that invaded *D. melanogaster* populations within the last 100 years (I-element, hobo, P-element) caused some hybrid dysgenesis (HD) symptoms. To test whether Tirant also induces HD symptoms, we performed crosses between strains having recent Tirant insertions (Urbana-S and Hikone-R) and strains not having such insertions (Lausanne-S and Canton-S). All strains do not have recent P-element, I-element, and hobo insertions, which rules out interference by the other HD systems (fig. 2*A* and supplementary table 1, Supplementary Material online). We investigated the fraction of dysgenic ovaries in the F1 generation, a trait influenced by P-element and hobo mobilization (Kidwell et al. 1977; Blackman et al. 1987; Yannopoulos et al. 1987), and the fraction of hatched F2 embryos, a trait influenced by I-element mobilization (Bucheton et al. 1976). We performed all crosses at several temperatures (supplementary fig. 17*A* and *B*, Supplementary Material online), as temperature frequently has a strong influence on the extent of HD symptoms (Kidwell et al. 1977; Bucheton 1979; Kidwell and Novy 1979; Serrato-Capuchina et al. 2020). We did not find any significant differences in the number of dysgenic ovaries nor in the number of hatched eggs between the reciprocal crosses (supplementary fig. 17A and B and table 5, Supplementary Material online). As the number of paternally inherited TEs may influence the magnitude of HD (Serrato-Capuchina et al. 2020), we performed reciprocal crosses with the strain carrying the largest number of canonical Tirant insertions, that is, Iso-1, and strains not having canonical Tirant insertions (Lausanne-S and Crimea; supplementary fig. 1, Supplementary Material online). However, Iso-1 also carries I-element and hobo insertions (supplementary fig. 1, Supplementary Material online). Therefore, we performed crosses solely at 25 °C, a temperature where I-element HD is usually not observed (Bucheton et al. 1976). As strains inducing hobo HD are very rare (Pascual and Periquet 1991), there is solely a small chance that hobo activity will generate atrophic ovaries in this crosses. We again did not find any significant differences in the number of dysgenic ovaries nor in the number of hatched eggs among these crosses (supplementary fig. 17C and D and table 5, Supplementary Material online; which also rules out hobo HD).

We hypothesize that the absence of apparent HD symptoms may be one reason why the invasion of Tirant in natural *D. melanogaster* populations during the past century was not detected before.

Origin of the Canonical Tirant Invasion

Lastly, we aimed to shed light on the origin of the Tirant invasion. Since canonical Tirant insertions are mostly absent in strains collected before 1938, we reasoned that the recent Tirant invasion was likely triggered by HT (or an introgression). To identify the putative donor species, we investigated Tirant sequences in different *Drosophila* species. We first tested if Tirant sequences can be found in 11 sequenced *Drosophila* genomes (*Drosophila* 12 Genomes Consortium 2007). Solely members of the *D. melanogaster* species subgroup contained reads mapping to Tirant (supplementary fig. 18, Supplementary Material online; *D. melanogaster*, *D. simulans*, *D. erecta*, *D. yakuba*; in agreement with Fablet et al. [2007]). We also found that *D. simulans* is the only species that may carry full-length insertions of Tirant (apart from *D. melanogaster*) and that some Tirant insertions in *D. simulans* may have a high similarity to the consensus sequence of Tirant (supplementary fig. 18, Supplementary Material online). To further investigate the composition of Tirant in the *D. melanogaster* species subgroup, we obtained Illumina short-read data for several individuals from different species of this subgroup. In addition to *D. melanogaster*, *D. simulans*, *D. erecta*, and *D. yakuba*, we also obtained data for *D. sechellia*, *D. mauritiana*, and *D. teisseri* (supplementary table 6, Supplementary Material online). A PCA based on the allele frequencies of Tirant SNPs confirms that the Tirant composition of recently collected *D. melanogaster* strains (>1938) is most similar to *D. simulans* strains (supplementary fig. 19, Supplementary Material online). The high similarity of some Tirant sequences between *D. melanogaster* and *D. simulans* was noted before (Fablet et al. 2006; Lerat et al. 2011; Bargues and Lerat 2017). However, an analysis based on the allele frequencies confounds the two subfamilies of Tirant in these two species, for example, canonical Tirant insertions (Tirant-C in *D. simulans*) and degraded Tirant insertions (Tirant-S in *D. simulans*) (Fablet et al. 2006). Therefore, to further investigate whether some Tirant insertions of *D. simulans* could have triggered the canonical Tirant invasion

in *D. melanogaster*, we analyzed the Tirant content in a recent long-read based assembly of *D. simulans* (strain w^{XD1}; Chakraborty et al. 2020). Indeed, we found that *D. simulans* carries three full-length insertions that have a high similarity to the consensus sequence of Tirant (average divergence: 1.97%, 1.56%, 1.60%; supplementary table 7, Supplementary Material online). We concluded that HT from *D. simulans* may have triggered the invasion of the canonical Tirant in *D. melanogaster* populations.

Discussion

We show that the retrotransposon Tirant invaded most natural *D. melanogaster* populations between 1930 and 1950, possibly following HT from *D. simulans*. Tirant is thus the fourth TE that invaded *D. melanogaster* in the last century. We also provide the first comprehensive timeline of the recent TE invasions in *D. melanogaster* populations that is based on a consistent approach (i.e., the same method and strains). The canonical Tirant invaded natural *D. melanogaster* populations first followed by the I-element, hobo, and finally by the P-element. All investigated strains, including those lacking canonical Tirant insertions, carry highly degraded Tirant fragments, which likely stem from an ancient Tirant invasion predating the split of the *D. melanogaster* species subgroup (Fablet et al. 2007; Lerat et al. 2011). We demonstrate that piRNAs derived from canonical and diverged Tirant insertions can be clearly distinguished and suggest that piRNAs derived from degraded Tirant copies, which were present in all investigated strains, were unable to prevent the invasion of the canonical Tirant. We show that founder effects may be important components of TE invasions that may lead to a heterogeneous TE composition among populations. Finally, we did not find apparent HD symptoms among reciprocal crosses of strains with and without canonical Tirant insertions.

Our conclusion that Tirant recently invaded *D. melanogaster* is mainly based on the absence of canonical Tirant sequences in most strains collected before 1938 and their presence in strains collected after 1938. As an alternative explanation, most strains collected before 1938 could have lost the canonical Tirant sequences. It was, for example, proposed that non-African *D. simulans* populations lost canonical Tirant sequences (Fablet et al. 2006). But this alternative explanation seems unlikely as it requires the independent loss of canonical Tirant sequences in strains collected before 1938 but not in any strain collected after 1938. The low population frequency of euchromatic Tirant insertions (see also Kofler, Nolte, et al. 2015) and the high sequence similarity between the left and the right LTR of Tirant insertions (Bowen and McDonald 2001; Bergman and Bensasson 2007) are also in agreement with our hypothesis of a recent Tirant invasion. Our hypothesis of the recent Tirant invasion is also consistent with the interpretation of the data for the I-element, P-element, and hobo, where the absence of the (canonical) TE in old strains combined with the presence in young strains was taken as evidence for recent invasions of these elements (Kidwell 1983; Daniels, Chovnick, et al. 1990; Daniels, Peterson, et al. 1990; Bucheton et al. 1992).

Our data suggest that Tirant was the first TE that invaded natural *D. melanogaster* populations in the last century. However, these results need to be interpreted with caution as 1) there is some uncertainty about the sampling time of the strains, 2) some strains may have been contaminated (e.g., the presence of the P-element in a strain collected around 1938 [Swedish-C] is likely due to mixing of strains during maintenance of stocks; supplementary table 1, Supplementary Material online), and 3) our strains are from different geographic regions, where some regions might have been invaded earlier than others. Nevertheless, our results are largely in agreement with previous works which suggested that the I-element invasion happened between 1930 and 1950, the hobo invasion around 1955 and the P-element invasion between 1950 and 1980 (Kidwell 1983; Anxolabéhère et al. 1988; Periquet et al. 1989).

We did not find evidence that Tirant induces HD symptoms. Also, a previous work in *D. simulans* did not report HD symptoms for Tirant despite Tirant being activated by reciprocal crosses (Akkouche et al. 2013). However, due to several reasons, more work will be necessary to show whether or not Tirant causes some HD symptoms. First, it is not clear what symptoms to look for. We investigated the fraction of dysgenic ovaries in the F1 and the fraction of hatched eggs (F2), two traits affected by HD from P-element, I-element, or hobo. However, it is feasible that Tirant activity leads to entirely different phenotypic effects, especially given that Tirant may be active in the germline and in the soma (Malone et al. 2009; Akkouche et al. 2013; Czech et al. 2013), and could thus affect both tissues. Second, it is not clear if intermediate TEs, such as Tirant, are able to induce HD. Different phenotypes among reciprocal crosses (i.e., HD) can solely be observed if maternally transmitted piRNAs (i.e., the cytoplasmic component of HD) are necessary to silence a TE (Brennecke et al. 2008). Maternally transmitted piRNAs initiate the ping-pong cycle and recruit silencing chromatin that is then bound by Rhino, which in turn defines the site of dual-strand clusters (Le Thomas et al. 2014). As both ping-pong and dual-strand clusters are solely observed in the germline piRNA pathway (Malone et al. 2009), it is thought that maternally deposited piRNAs are important for the germline pathway but not for the somatic piRNA pathway. Consequently, no HD symptoms are expected for TEs that are solely active in the soma. The I-element, hobo, and the P-element, three TEs that invaded *D. melanogaster* populations recently, were all active in the germline and induced HD symptoms (Bingham et al. 1982; Calvi and Gelbart 1994; Biémont 2010; Moon et al. 2018; Wang et al. 2018). However, the situation is entirely unclear for intermediate elements such as Tirant. Surprisingly, one study even suggested that maternally transmitted piRNAs are necessary to silence Tirant in the soma (Akkouche et al. 2013). The molecular mechanisms behind this influence of maternal piRNAs on the somatic piRNA pathway remain yet unclear. Third, the severity of HD symptoms frequently depends on multiple factors, such as temperature and the age of flies (Kidwell et al. 1977; Bucheton

1979; Kidwell and Novy 1979; Serrato-Capuchina et al. 2020). It is feasible that HD symptoms of Tirant can only be observed under certain conditions, and these conditions could differ substantially from the previously described HD systems. Fourth, previous studies noted marked differences in the ability to induce or repress HD among different strains (Kidwell et al. 1977, 1988; Anxolabéhère et al. 1988; Pascual and Periquet 1991; Srivastav et al. 2019). This could be mediated by differences in the number of paternally transmitted TEs (Srivastav and Kelleher 2017; Serrato-Capuchina et al. 2020), different variants of the TEs (Srivastav et al. 2019), and differences in the tolerance to TE activity among strains (Kelleher et al. 2018). The abundance of strains inducing HD may also vary among the HD systems. For example, strains inducing P-element HD are readily found whereas strains inducing hobo HD are rare (Kidwell 1983; Pascual and Periquet 1991). It is thus feasible that solely crosses of certain strains show HD symptoms of Tirant.

It is currently unclear how canonical Tirant sequences entered *D. melanogaster* populations. Possible explanations are HT or introgression from a related species (Silva et al. 2004; Sánchez-Gracia et al. 2005; Loreto et al. 2008; Bartolomé et al. 2009). In search for a possible donor species, we found that *D. simulans* carries some full-length Tirant insertions with a high similarity to canonical Tirant in *D. melanogaster* (supplementary table 7, Supplementary Material online). Out of the two Tirant subfamilies found in *D. simulans*, Tirant-C (nondegraded insertions) and Tirant-S (degraded insertions), Tirant-C insertions have been previously shown to be closely related to the canonical Tirant in *D. melanogaster* (Fablet et al. 2006; Lerat et al. 2011; Bargues and Lerat 2017). We thus suggest that HT of Tirant-C from *D. simulans* to *D. melanogaster* may have triggered the canonical Tirant invasion in *D. melanogaster*, in agreement with Lerat et al. (2011). Apart from this HT, Tirant is likely mostly vertically transmitted in the *D. melanogaster* species subgroup (Fablet et al. 2007). In agreement with this, a tree based on frequency of Tirant SNPs largely follows the species tree (supplementary fig. 20, Supplementary Material online). HT of TEs between *D. melanogaster* and *D. simulans* is plausible since both species are closely related (Lemeunier et al. 1976) and have largely overlapping habitats (Parsons and Stanley 1981), which generates ample opportunities for HT or introgressions. HT of TEs between these species was observed before in both directions. For example, Kofler, Hill, et al. (2015) suggested that *D. simulans* recently acquired the P-element from *D. melanogaster*. Conversely, hobo and the I-element in *D. melanogaster* were possibly acquired from *D. simulans* (Daniels, Chovnick, et al. 1990; Simmons 1992; Loreto et al. 2008).

We found that Tirant sequences from Tasmania (an island south of Australia) have a different composition than Tirant sequences from other locations (at least five SNPs have distinctly different frequencies; supplementary table 4, Supplementary Material online). We suggest that this may be due to a founder effect during the Tirant invasion, which led to the spread of rare Tirant variants in Tasmanian populations. We wondered whether the observed founder effect could be due to the recent colonization of Australia (Tasmania) by *D. melanogaster* (Bock and Parsons 1981). However, this seems unlikely as the colonization of Australia, and probably also of Tasmania, predates the Tirant invasion. *Drosophila melanogaster* was first spotted in Australia in 1894 and is known to rapidly spread into nearby areas (Bock and Parsons 1981; Keller 2007), whereas the Tirant invasion mostly happened between 1938 and 1950. Moreover, founder effects that occurred during the colonization of Tasmania should affect the entire genomic background of *D. melanogaster* and not just the Tirant sequences. Previous studies did not detect any signatures of bottlenecks for Tasmanian *D. melanogaster* populations (Agis and Schlötterer 2001; Grenier et al. 2015; Bergland et al. 2016; Arguello et al. 2019). We thus argue that the founder effect in Tasmania is specific to Tirant. Founder effects during TE invasions could be important, hitherto little considered, processes that may lead to geographically distinct TE variants.

We suggest that four different TEs invaded *D. melanogaster* populations within 40 years (between the 1930s and 1970s). Why did so many different TEs spread in *D. melanogaster* within such a short time? A possible explanation could be the recent habitat expansion of *D. melanogaster* into the Americas and Australia about 100–200 years ago (Bock and Parsons 1981; Vieira et al. 1999; Kofler, Nolte, et al. 2015). Habitat expansion may bring species into contact that did not coexist before in the same habitat. If these species carry different TE families, HT events between the species may trigger novel TE invasions. A classic example is the P-element in *D. melanogaster* which was likely acquired from *D. willistoni* after *D. melanogaster* entered the habitat of *D. willistoni* in South America (Engels 1992). The lag-time between colonization of the Americas and Australia (~100–200 years ago; Bock and Parsons 1981; Keller 2007) and the four different TE invasions (1930–1970) may be due to the stochasticity of HT events, a strong influence of drift in the early stages of TE invasions and the time required until a TE reaches an appreciable frequency (Ginzburg et al. 1984; Le Rouzic and Capy 2005). It will be interesting to see if such a high rate of novel TE invasions in *D. melanogaster* populations will be maintained over the next century. An absence of novel invasions would support our hypothesis that the habitat expansion triggered the four recent TE invasions in *D. melanogaster*.

Out of the four TEs that invaded *D. melanogaster* populations in the last century, the P-element is unique as it is the only TE that does not show substantial similarity to any sequence of the *D. melanogaster* genome. For the other three TEs—Tirant, the I-element, and hobo—many degraded insertions can be found (mostly in the heterochromatin) (Bucheton et al. 1984, 1986, 1992; Daniels, Chovnick, et al. 1990). Thus, three out of the four TEs probably invaded *D. melanogaster* populations at least twice. This raises the question of how multiple waves of invasions arise. Before a TE can trigger a novel invasion the TE needs to overcome the host defense (or the host defense may break down). For example, in mammals and invertebrates efficient silencing of a TE requires piRNAs that match with less than 10% sequence divergence over the bulk of the TE sequence (Post et al. 2014;

Kotov et al. 2019). A TE that diverged by more than 10% from the piRNA pool of the host (e.g., the canonical Tirant compared with the degraded Tirant sequences) could thus trigger a second wave of an invasion. The same consideration holds for other host defense mechanism that rely on sequence similarity to a TE, like small RNAs in plants or Kruppel-associated box zinc-finger proteins in mammals (Marí-Ordóñez et al. 2013; Yang et al. 2017). It is however an important open question whether sufficient sequence divergence could be acquired within a host species, where host defense mechanisms may coadapt with the TE, or whether HT to an intermediate host (e.g., a closely related species) is necessary to overcome the host defense.

Materials and Methods

Strains and Dating

The sequenced fly strains were obtained from the Bloomington Drosophila Stock Center (BDSC) (Crimea, Lausanne-S, Swedish-C, Urbana-S, Berlin-K, Hikone-R, Florida-9, Pi2, Harwich, Amherst-3) and the National Drosophila Species Stock Center (Dmel68). w1118 and wk were kindly provided by Silke Jensen. We additionally analyzed publicly available sequencing data of different *D. melanogaster* strains (King et al. 2012; Mackay et al. 2012; Bergland et al. 2014; Grenier et al. 2015; Lack et al. 2015; Jakšić et al. 2017; Machado et al. 2019; Kapun et al. 2020; Wierzbicki et al. 2020) (supplementary table 6, Supplementary Material online). The collection dates of the strains were obtained from different sources. If available, we used the collection dates from Lindsley and Grell (1968). Alternatively, we used the collection dates published in previous works (Black et al. 1987; Anxolabéhère et al. 1988; Galindo et al. 1995; Engels 2007; Ruebenbauer et al. 2008) or information from the National Drosophila Species Stock Center (drosophilaspecies.com) and FlyBase (flybase.org/reports/FBrf0222222.html, last accesssed December 15, 2020) (supplementary table 1, Supplementary Material online). For the strains w1118 and Urbana-S, we used the latest possible collection date: for w1118, we used the publication date of the first publication mentioning the strain and for Urbana-S, we used the year of the death of C. Bridges, who collected the strain (Lindsley and Grell 1968) (supplementary table 1, Supplementary Material online). The geographic origin was obtained from the same sources. For an overview of the used strains, the estimated collection date, and the source of the information, see supplementary table 1, Supplementary Material online. The Iso-1 strain was generated by crossing several laboratory strains, with largely unknown sampling dates (Brizuela et al. 1994). Therefore, we did not assign a sampling date to this strain. Additionally, we used publicly available data of different strains from *D. simulans*, *D. sechellia*, *D. mauritiana*, *D. yakuba*, *D. teisseri*, and *D. erecta* (*Drosophila* 12 Genomes Consortium 2007; Garrigan et al. 2012, 2014; Rogers et al. 2014; Turissini et al. 2015; Melvin et al. 2018; Miller et al. 2018; Schrider et al. 2018; Cooper et al. 2019; Kang et al. 2019; Lanno et al. 2019; Meany et al. 2019; Stewart and Rogers 2019). For an overview of all used publicly available data, see supplementary table 6, Supplementary Material online.

DNA Sequencing

DNA for Illumina paired-end sequencing was extracted from whole bodies of 20–30 virgin female flies using a salt-extraction protocol (Maniatis et al. 1982). Libraries were prepared with the NEBNext Ultra II DNA library Prep Kit (New England Bioloabs, Ipswich, MA) using 1 μg DNA. Illumina sequencing was performed by the Vienna Biocenter Core Facilities using the HiSeq2500 platform (2×125 bp; Illumina, San Diego, CA).

Small RNA Sequencing

For small RNA sequencing, we extracted total RNA from ovaries of the strains Canton-S, Iso-1, and Lausanne-S using TRIzol. The small RNA was sequenced by Fasteris (Geneva, Switzerland). After depletion of 2S rRNA, library preparation was performed using the Illumina TruSeq small RNA kit and cDNA was sequenced on an Illumina NextSeq platform (50 bp; Illumina, San Diego, CA). Adapter sequences were trimmed with cutadapt (v2.3) (Martin 2011) (adapter: TGGAATTCTCGGGTGCCAAGGAACTCCAGTCACCATTTTATCTCGTATGC) and filtered for reads with a length between 18 and 36 nt. The reads were mapped to a database consisting of *D. melanogaster* miRNAs, mRNAs, rRNAs, snRNAs, snoRNAs, tRNAs (Thurmond et al. 2019), and the TE sequences (Quesneville et al. 2005) using novoalign (v3.09; http://novocraft.com/, last accesssed December 15, 2020) and allowing for two mismatches (unless mentioned otherwise). Solely piRNAs with a length between 23 and 29 nt were retained and the abundance of piRNAs was normalized to a million miRNAs as described previously (Kofler et al. 2018). For computing the ping-pong signatures and visualizing the piRNA abundance along the Tirant sequence, we used a previously developed pipeline (Kofler et al. 2018). To calculate the abundance of piRNAs complementary to the degraded Tirant fragments, we first extracted the sequences of degraded Tirant insertions (>10% divergence to consensus sequence) from the reference assembly of Iso-1 (v6.22) with RepeatMasker (open-4.0.7; Smit et al. 2013–2015) and bedtools (Quinlan and Hall 2010) (v2.29.2). All sequences longer than 100 bp were concatenated (the reverse complement was adjusted with bedtools) and small RNAs were mapped to these sequences using novoalign. The abundance of all piRNAs complementary to degraded Tirant sequences was summed. We also analyzed the small RNA content of the five GDL strains B10, I06, N10, T05, and ZW155 (data are publicly available; Luo et al. 2020).

TE Abundance and Diversity

The coverage along a TE and the frequencies of SNPs and indels in a TE were computed using our newly developed tool DeviaTE (v0.3.8) (Weilguny and Kofler 2019). Briefly, short reads from a sample were aligned with bwa sw (v0.7.17) (Li and Durbin 2009) to the TE consensus sequences of *Drosophila* (Quesneville et al. 2005) as well as to three single-copy genes (*traffic jam*, *rpl32*, and *rhino*), which allowed

us to infer TE copy numbers by contrasting the coverage of a TE to the coverage of the single-copy genes. The abundance and diversity of TE insertions were visualized with DeviaTE (Weilguny and Kofler 2019). To obtain the normalized number of reads mapping to each TE (rpm), we used PopoolationTE2 (v1.10.03) (Kofler et al. 2016). Based on the visualization of the TE composition with DeviaTE and the estimates of the TE abundance (rpm and DeviaTE using normalization with single-copy genes), we manually classified the presence/absence of Tirant, hobo, the I-element, and the P-element in different *D. melanogaster* strains (supplementary table 1, Supplementary Material online). We used the following three categories: 1) absence of any TE sequences, 2) solely degraded TE sequences are present, 3) nondegraded sequences, with a high similarity to the consensus sequence, are present. For example, see supplementary figures 4 and 6–8, Supplementary Material online. A PCA based on the allele frequencies of SNPs in a TE supports our classification for Tirant and hobo. Since many strains do not contain any P-element sequences, the allele frequencies of SNPs in the P-element could not be calculated for all strains. Despite discernible differences between strains with and without recent I-element insertions, the PCA did not separate these two groups (supplementary figs. 6 and 14, Supplementary Material online). The PCA was performed in R (prcomp) using arcsine and square root transformed allele frequencies of SNPs in TEs (R Core Team 2012). The DSPR lines were not included into the PCA due to their short-read length (50 bp). The pairwise F_{ST} based on the SNPs of TEs was computed with Popoolation2 (v1.2.01) (Kofler et al. 2011) ("fst-sliding.pl" –window-size 8526 –max-coverage 0.1%).

We used PoMo (Schrempf et al. 2016) based on the allele frequencies of Tirant SNPs to generate a tree of the species in the *D. melanogaster* species subgroup. PoMo uses allele frequency data to account for the intraspecific differences while calculating the interspecific variation. We run PoMo with IQ-TREE (v1.6.12) (Nguyen et al. 2015) using polymorphism-aware models (HKY + P). We obtained bootstrap estimates for each node using the ultra-fast bootstrap (-bb) option for 1000 replicates.

Tirant sequences in the assemblies of Canton-S (Wierzbicki et al. 2020) and Iso-1 (v6.22; https://flybase.org/, last accessed December 15, 2020) were identified with RepeatMasker using the TE consensus sequences of *Drosophila* as custom library (Quesneville et al. 2005). To visualize the divergence of annotated Tirant fragments of the Canton-S genome, we extract all sequences annotated with RepeatMasker and map them to the Tirant consensus sequence using bwa sw (Li and Durbin 2009) with a low mismatch penalty (-b) of 0.5. Visualization of the sequence alignment was done with IGV. Colored lines represent SNPs compared with the consensus sequence.

We searched for canonical Tirant insertions in a long-read based assembly of *D. simulans* (strain w^{XD1}; PRJNA383250; Chakraborty et al. 2020) using RepeatMasker (open-4.0.7; Smit et al. 2013–2015). We filtered for complete insertions with a low divergence (<5%).

To estimate the position and population frequency of canonical and degraded Tirant insertions in a natural *D. melanogaster* population, we used PoPoolationTE2 (v1.10.03) (Kofler et al. 2016) and a population collected in 2014 at Viltain (France) by the DrosEU consortium (SRR5647729; Kapun et al. 2020). We generated the artificial reference genome required by PoPoolationTE2, by merging the repeat masked reference genome, the consensus sequence of Tirant and the degraded Tirant sequences with a minimum length of 100 bp (see above) into a single fasta file. The short reads were mapped to this artificial reference genome using bwa mem (v0.7.17) (Li and Durbin 2009) with paired-end mode and the parameter -M. The mapped reads were sorted with samtools (Li, Handsaker, et al. 2009). Finally, we followed the PoPoolationTE2 pipeline using the parameters: –map-qual 15, –min-count 2, –min-coverage 2. We indicated heterochromatic regions following previous work (Riddle et al. 2011; Hoskins et al. 2015).

Hybrid Dysgenesis Assay

To test whether Tirant induces HD symptoms, we performed four reciprocal crosses among *D. melanogaster* strains having canonical Tirant insertions (Urbana-S, Hikone-R, and Iso-1) and strains not having canonical Tirant insertions (Lausanne-S, Canton-S, Crimea). Each cross was performed in three replicates by mating 20 female virgin flies with 15 males. To estimate the number of dysgenic ovaries, 2–5 days old F1 flies (kept at either 20, 25, or 29 °C) were allowed to lay eggs on black agar plates (containing charcoal) for 24 h. The F1 female ovaries were dissected on PBS and scored for the presence of dysgenic (underdeveloped) ovaries. The deposited F2 embryos were counted, incubated for 24 h, and the number of larvae (=hatched eggs) was quantified. Crosses involving Iso-1 were only performed at 25 °C.

Supplementary Material

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

Acknowledgments

We thank Wolfgang Miller for access to his extensive library, Elisabeth Salbaba for technical support, and Divya Selvaraju for providing the Harwich data. We thank all members of the Institute of Population Genetics for feedback and support. This work was supported by the Austrian Science Foundation (FWF) grants P30036-B25 to R.K. and W1225.

Author Contributions

F.S. and R.K. conceived the work. F.S. and F.W. analyzed the data. K.-A.S. provided feedback on the manuscript. F.S. and R.K. wrote the manuscript.

Data Availability

All scripts are available at https://sourceforge.net/projects/te-tools/ (last accesssed December 15, 2020) (folder tirant) and important files (including all DeviaTE outputs) at https://sourceforge.net/projects/te-tools/files/tirant_data/ (last

accesssed December 15, 2020). The sequence data of the old laboratory strains and the piRNA sequences are available at NCBI (PRJNA634847).

References

Agis M, Schlötterer C. 2001. Microsatellite variation in natural *Drosophila melanogaster* populations from New South Wales (Australia) and Tasmania. *Mol Ecol*. 10(5):1197–1205.

Akkouche A, Grentzinger T, Fablet M, Armenise C, Burlet N, Braman V, Chambeyron S, Vieira C. 2013. Maternally deposited germline piRNAs silence the tirant retrotransposon in somatic cells. *EMBO Rep*. 14(5):458–464.

Akkouche A, Rebollo R, Burlet N, Esnault C, Martinez S, Viginier B, Terzian C, Vieira C, Fablet M. 2012. tirant, a newly discovered active endogenous retrovirus in *Drosophila simulans*. *J Virol*. 86(7):3675–3681.

Anxolabéhère D, Kidwell MG, Periquet G. 1988. Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of *Drosophila melanogaster* by mobile P elements. *Mol Biol Evol*. 5(3):252–269.

Arguello JR, Laurent S, Clark AG, Gaut B. 2019. Demographic history of the human commensal *Drosophila melanogaster*. *Genome Biol Evol*. 11(3):844–854.

Barckmann B, El-Barouk M, Pélisson A, Mugat B, Li B, Franckhauser C, Fiston Lavier A-S, Mirouze M, Fablet M, Chambeyron S. 2018. The somatic piRNA pathway controls germline transposition over generations. *Nucleic Acids Res*. 46(18):9524–9536.

Bargues N, Lerat E. 2017. Evolutionary history of LTR-retrotransposons among 20 *Drosophila* species. *Mob DNA*. 8(1):1–15.

Bartolomé C, Bello X, Maside X. 2009. Widespread evidence for horizontal transfer of transposable elements across *Drosophila* genomes. *Genome Biol*. 10(2):R22.

Bartolomé C, Maside X, Charlesworth B. 2002. On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. *Mol Biol Evol*. 19(6):926–937.

Begun DJ, Aquadro CF. 1995. Molecular variation at the vermilion locus in geographically diverse populations of *Drosophila melanogaster* and *D. simulans*. *Genetics* 140(3):1019–1032.

Bergland AO, Behrman EL, O'Brien KR, Schmidt PS, Petrov DA. 2014. Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in *Drosophila*. *PLoS Genet*. 10(11):e1004775.

Bergland AO, Tobler R, González J, Schmidt P, Petrov D. 2016. Secondary contact and local adaptation contribute to genome-wide patterns of clinal variation in *Drosophila melanogaster*. *Mol Ecol*. 25(5):1157–1174.

Bergman CM, Bensasson D. 2007. Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A*. 104(27):11340–11345.

Biémont C. 2010. A brief history of the status of transposable elements: from junk DNA to major players in evolution. *Genetics* 186(4):1085–1093.

Bingham PM, Kidwell MG, Rubin GM. 1982. The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family. *Cell* 29(3):995–1004.

Black DM, Jackson MS, Kidwell MG, Dover GA. 1987. KP elements repress P-induced hybrid dysgenesis in *Drosophila melanogaster*. *EMBO J*. 6(13):4125–4135.

Blackman RK, Grimaila R, Macy M, Koehler D, Gelbart WM. 1987. Mobilization of hobo elements residing within the decapentaplegic gene complex: suggestion of a new hybrid dysgenesis system in *Drosophila melanogaster*. *Cell* 49(4):497–505.

Blumenstiel JP. 2011. Evolutionary dynamics of transposable elements in a small RNA world. *Trends Genet*. 27(1):23–31.

Blumenstiel JP. 2019. Birth, school, work, death and resurrection: the life stages and dynamics of transposable element proliferation. *Genes* 10(5):336.

Bock I, Parsons P. 1981. Species of Australia and New Zealand. In Ashburner M, Carson L, Thompson JJ, editors. Vol. 3a. The genetics and biology of Drosophila. Oxford: Academic Press. p. 349–393.

Bonnivard E, Bazin C, Denis B, Higuet D. 2000. A scenario for the hobo transposable element invasion, deduced from the structure of natural populations of *Drosophila melanogaster* using tandem TPE repeats. *Genet Res*. 75(1):13–23.

Bowen NJ, McDonald JF. 2001. *Drosophila* euchromatic LTR retrotransposons are much younger than the host species in which they reside. *Genome Res*. 11(9):1527–1540.

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* 128(6):1089–1103.

Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ. 2008. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* 322(5906):1387–1392.

Brizuela BJ, Elfring L, Ballard J, Tamkun JW, Kennison JA. 1994. Genetic analysis of the *brahma* gene of *Drosophila melanogaster* and polytene chromosome subdivisions 72AB. *Genetics* 137(3):803–813.

Bucheton A. 1979. Non-Mendelian female sterility in *Drosophila melanogaster*: influence of aging and thermic treatments. III. Cumulative effects induced by these factors. *Genetics* 93(1):131–142.

Bucheton A, Lavige JM, Picard G, L'Heritier P. 1976. Non-Mendelian female sterility in *Drosophila melanogaster*: quantitative variations in the efficiency of inducer and reactive strains. *Heredity* 36(3):305–314.

Bucheton A, Paro R, Sang HM, Pelisson A, Finnegan DJ. 1984. The molecular basis of I-R hybrid Dysgenesis in *Drosophila melanogaster*: identification, cloning, and properties of the I factor. *Cell* 38(1):153–163.

Bucheton A, Simonelig M, Vaury C, Crozatier M. 1986. Sequences similar to the I transposable element involved in I-R hybrid dysgenesis in *D. melanogaster* occur in other *Drosophila* species. *Nature* 322(6080):650–652.

Bucheton A, Vaury C, Chaboissier M-C, Abad P, Pélisson A, Simonelig M. 1992. I elements and the *Drosophila* genome. *Genetica* 86(1-3):175–190.

Calvi BR, Gelbart WM. 1994. The basis for germline specificity of the hobo transposable element in *Drosophila melanogaster*. *EMBO J*. 13(7):1636–1644.

Cañizares J, Grau M, Paricio N, Moltó MD. 2000. Tirant is a new member of the Gypsy family of retrotransposons in *Drosophila melanogaster*. *Genome* 43(1):9–14.

Chakraborty M, Chang C-H, Khost D, Vedanayagam J, Adrion JR, Liao Y, Montooth KL, Meiklejohn CD, Larracuente AM, Emerson JJ. 2020. Evolution of genome structure in the Drosophila simulans species complex. *bioRxiv*.

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):419275.

Cooper JC, Guo P, Bladen J, Phadnis N. 2019. A triple-hybrid cross reveals a new hybrid incompatibility locus between D. melanogaster and D. sechellia. *bioRxiv* 590588.

Czech B, Preall JB, McGinn J, Hannon GJ. 2013. A transcriptome-wide RNAi screen in the *Drosophila* ovary reveals factors of the germline piRNA pathway. *Mol Cell*. 50(5):749–761.

Daniels SB, Chovnick A, Boussy IA. 1990. Distribution of hobo transposable elements in the genus *Drosophila*. *Mol Biol Evol*. 7(6):589–606.

Daniels SB, Peterson KR, Strausbaugh LD, Kidwell MG, Chovnick A. 1990. Evidence for horizontal transmission of the P transposable element between *Drosophila* species. *Genetics* 124(2):339–355.

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284(5757):601–603.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450(7167):203–218.

Eanes WF, Wesley C, Charlesworth B. 1992. Accumulation of P elements in minority inversions in natural populations of *Drosophila melanogaster*. *Genet Res*. 59(1):1–9.

Engels WR. 1992. The origin of P elements in *Drosophila melanogaster*. *Bioessays* 14(10):681–686.
Engels WR. 2007. Hybrid dysgenesis in *Drosophila melanogaster*: rules of inheritance of female sterility. *Genet Res*. 89(5-6):407–424.
Fablet M, Lerat E, Rebollo R, Horard B, Burlet N, Martinez S, Brasset È, Gilson E, Vaury C, Vieira C. 2009. Genomic environment influences the dynamics of the tirant LTR retrotransposon in *Drosophila*. *FASEB J*. 23(5):1482–1489.
Fablet M, McDonald JF, Biémont C, Vieira C. 2006. Ongoing loss of the tirant transposable element in natural populations of *Drosophila simulans*. *Gene* 375(1–2):54–62.
Fablet M, Souames S, Biémont C, Vieira C. 2007. Evolutionary pathways of the tirant LTR retrotransposon in the *Drosophila melanogaster* subgroup of species. *J Mol Evol*. 64(4):438–447.
Galindo MI, Ladevèze V, Lemeunier F, Kalmes R, Periquet G, Pascual L. 1995. Spread of the autonomous transposable element hobo in the genome of *Drosophila melanogaster*. *Mol Biol Evol*. 12(5):723–734.
Garrigan D, Kingan SB, Geneva AJ, Andolfatto P, Clark AG, Thornton KR, Presgraves DC. 2012. Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. *Genome Res*. 22(8):1499–1511.
Garrigan D, Kingan SB, Geneva AJ, Vedanayagam JP, Presgraves DC. 2014. Genome diversity and divergence in *Drosophila mauritiana*: multiple signatures of faster X evolution. *Genome Biol Evol*. 6(9):2444–2458.
Ginzburg LR, Bingham PM, Yoo S. 1984. On the theory of speciation induced by transposable elements. *Genetics* 107(2):331–341.
Goodier JL. 2016. Restricting retrotransposons: a review. *Mob DNA*. 7(1):16.
Grenier JK, Roman Arguello J, 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(4):593–603.
Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, Siomi H, Siomi MC. 2007. A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*. *Science* 315(5818):1587–1590.
Hickey DA. 1982. Selfish DNA: a sexually transmitted nuclear parasite. *Genetics* 101(3-4):519–531.
Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, Booth BW, Pfeiffer BD, George RA, Svirskas R, et al. 2015. The Release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome Res*. 25(3):445–458.
Jakšić AM, Kofler R, Schlötterer C. 2017. Regulation of transposable elements: interplay between TE-encoded regulatory sequences and host-specific trans-acting factors in *Drosophila melanogaster*. *Mol Ecol*. 26(19):5149–5159.
Kang L, Rashkovetsky E, Michalak K, Garner HR, Mahaney JE, Rzigalinski BA, Korol A, Nevo E, Michalak P. 2019. Genomic divergence and adaptive convergence in *Drosophila simulans* from evolution Canyon, Israel. *Proc Natl Acad Sci U S A*. 116(24):11839–11844.
Kapun M, Barrón MG, Staubach F, Obbard DJ, Wiberg RAW, Vieira J, Goubert C, Rota-Stabelli O, Kankare M, Bogaerts-Márquez M, et al. 2020. Genomic analysis of European *Drosophila melanogaster* populations reveals longitudinal structure, continent-wide selection, and previously unknown DNA viruses. *Mol Biol Evol*. 37(9):2661–2678.
Kelleher E, Jaweria J, Akoma U, Ortega L, Tang W. 2018. QTL mapping of natural variation reveals that the developmental regulator Bruno reduces tolerance to P-element transposition in the Drosophila female germline. *PLoS Biol*. 16(10):e2006040.
Keller A. 2007. *Drosophila melanogaster*'s history as a human commensal. *Curr Biol*. 17(3):R77–R81.
Kidwell MG. 1983. Evolution of hybrid dysgenesis determinants in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A*. 80(6):1655–1659.
Kidwell MG, Kidwell JF, Sved JA. 1977. Hybrid dysgenesis in *Drosophila melanogaster*: a syndrome of aberrant traits including mutations, sterility and male recombination. *Genetics* 86(4):813–833.
Kidwell MG, Kimura K, Black DM. 1988. Evolution of hybrid dysgenesis potential following P element contamination in *Drosophila melanogaster*. *Genetics* 119(4):815–828.
Kidwell MG, Novy JB. 1979. Hybrid dysgenesis in Drosophila melanogaster: sterility resulting from gonadal dysgenesis in the P-M system. *Genetics* 92(4):1127–1140.
King EG, Merkes CM, McNeil CL, Hoofer SR, Sen S, Broman KW, Long AD, Macdonald SJ. 2012. Genetic dissection of a model complex trait using the *Drosophila* synthetic population resource. *Genome Res*. 22(8):1558–1566.
Kofler R, Betancourt AJ, Schlötterer C. 2012. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. *PLoS Genet*. 8(1):e1002487.
Kofler R, Gómez-Sánchez D, Schlötterer C. 2016. PoPoolationTE2: comparative population genomics of transposable elements using Pool-Seq. *Mol Biol Evol*. 33(10):2759–2764.
Kofler R, Hill T, Nolte V, Betancourt AJ, Schlötterer C. 2015. The recent invasion of natural *Drosophila simulans* populations by the P-element. *Proc Natl Acad Sci U S A*. 112(21):6659–6663.
Kofler R, Nolte V, Schlötterer C. 2015. Tempo and mode of transposable element activity in *Drosophila*. *PLoS Genet*. 11(7):e1005406–e1005421.
Kofler R, Pandey RV, Schlötterer C. 2011. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* 27(24):3435–3436.
Kofler R, Senti K-A, Nolte V, Tobler R, Schlotterer C. 2018. Molecular dissection of a natural transposable element invasion. *Genome Res*. 28(6):824–835.
Kotov AA, Adashev VE, Godneeva BK, Ninova M, Shatskikh AS, Bazylev SS, Aravin AA, Olenina LV. 2019. PiRNA silencing contributes to interspecies hybrid sterility and reproductive isolation in *Drosophila melanogaster*. *Nucleic Acids Res*. 47(8):4255–4271.
Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, Langley CH, Pool JE. 2015. The *Drosophila* genome nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. *Genetics* 199(4):1229–1241.
Lanno SM, Shimshak SJ, Peyser RD, Linde SC, Coolon JD. 2019. Investigating the role of Osiris genes in *Drosophila sechellia* larval resistance to a host plant toxin. *Ecol Evol*. 9(4):1922–1933.
Le Rouzic A, Capy P. 2005. The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. *Genetics* 169(2):1033–1043.
Le Thomas A, Rogers AK, Webster A, Marinov GK, Liao SE, Perkins EM, Hur JK, Aravin AA, Tóth KF. 2013. Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes Dev*. 27(4):390–399.
Le Thomas A, Stuwe E, Li S, Du J, Marinov G, Rozhkov N, Chen YCA, Luo Y, Sachidanandam R, Toth KF, et al. 2014. Transgenerationally inherited piRNAs trigger piRNA biogenesis by changing the chromatin of piRNA clusters and inducing precursor processing. *Genes Dev*. 28(15):1667–1680.
Lemeunier F, Ashburner M, Thoday JM. 1976. Relationships within the *melanogaster* species subgroup of the genus *Drosophila* (*Sophophora*) – II. Phylogenetic relationships between six species based upon polytene chromosome banding sequences. *Proc R Soc Lond B Biol Sci*. 193(1112):275–294.
Lerat E, Burlet N, Biémont C, Vieira C. 2011. Comparative analysis of transposable elements in the *melanogaster* subgroup sequenced genomes. *Gene* 473(2):100–109.
Li C, Vagin VV, Lee S, Xu J, Ma S, Xi H, Seitz H, Horwich MD, Syrzycka M, Honda BM, et al. 2009. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. *Cell* 137(3):509–521.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14):1754–1760.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
Lindsley DH, Grell EH. 1968. Genetic variations of Drosophila melanogaster. Washington: Carnegie Institute of Washington Publication.

Lohe AR, Moriyama EN, Lidholm DA, Hartl DL. 1995. Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements. *Mol Biol Evol*. 12(1):62–72.
Loreto EL, Carareto CMA, Capy P. 2008. Revisiting horizontal transfer of transposable elements in *Drosophila*. *Heredity* 100(6):545–554.
Luo S, Zhang H, Duan Y, Yao X, Clark AG, Lu J. 2020. The evolutionary arms race between transposable elements and piRNAs in *Drosophila melanogaster*. *BMC Evol Biol*. 20(1):1–18.
Machado HE, Bergland AO, Taylor R, Tilk S, Behrman E, Dyer K, Fabian DK, Flatt T, González J, Karasov TL. 2019. Broad geographic sampling reveals predictable, pervasive, and strong seasonal adaptation in Drosophila. *bioRxiv*, 337543.
Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The *Drosophila melanogaster* genetic reference panel. *Nature* 482(7384):173–178.
Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, Hannon GJ. 2009. Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell* 137(3):522–535.
Maniatis T, Fritsch EF, Sambrook J, et al. 1982. Molecular cloning: a laboratory manual. Vol. 545. NY: Cold Spring Harbor Laboratory.
Marí-Ordóñez A, Marchais A, Etcheverry M, Martin A, Colot V, Voinnet O. 2013. Reconstructing *de novo* silencing of an active plant retrotransposon. *Nat Genet*. 45(9):1029–1039.
Marsano RM, Moschetti R, Caggese C, Lanave C, Barsanti P, Caizzi R. 2000. The complete Tirant transposable element in *Drososphila melanogaster* shows a structural relationship with retrovirus-like retrotransposons. *Gene* 247(1–2):87–95.
Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. 17(1):10–12.
Maruyama K, Hartl DL. 1991. Evidence for interspecific transfer of the transposable element mariner between *Drosophila* and *Zaprionus*. *J Mol Evol*. 33(6):514–524.
Meany MK, Conner WR, Richter SV, Bailey JA, Turelli M, Cooper BS. 2019. Loss of cytoplasmic incompatibility and minimal fecundity effects explain relatively low Wolbachia frequencies in *Drosophila mauritiana*. *Evolution* 73(6):1278–1295.
Melvin RG, Lamichane N, Havula E, Kokki K, Soeder C, Jones CD, Hietakangas V. 2018. Natural variation in sugar tolerance associates with changes in signaling and mitochondrial ribosome biogenesis. *eLife* 7:e40841.
Miller DE, Staber C, Zeitlinger J, Hawley RS. 2018. Highly contiguous genome assemblies of 15 *Drosophila* species generated using nanopore sequencing. *G3 (Bethesda)* 8(10):3131–3141.
Mizrokhi LJ, Mazo AM. 1990. Evidence for horizontal transmission of the mobile element jockey between distant *Drosophila* species. *Proc Natl Acad Sci U S A*. 87(23):9216–9220.
Moltó MD, Paricio N, López-Preciado M. A, Semeshin VF, Martínez-Sebastián MJ. 1996. Tirant: a new retrotransposon-like element in *Drosophila melanogaster*. *J Mol Evol*. 42(3):369–375.
Moon S, Cassani M, Lin YA, Wang L, Dou K, Zhang ZZ. 2018. A robust transposon-endogenizing response from germline stem cells. *Dev Cell*. 47(5):660–671.e3.
Mugnier N, Gueguen L, Vieira C, Biémont C. 2008. The heterochromatic copies of the LTR retrotransposons as a record of the genomic events that have shaped the *Drosophila melanogaster* genome. *Gene* 411(1–2):87–93.
Nefedova LN, Urusov FA, Romanova NI, Shmel'Kova AO, Kim AI. 2012. Study of the transcriptional and transpositional activities of the Tirant retrotransposon in *Drosophila melanogaster* strains mutant for the flamenco locus. *Genetika* 48(11):1271–1279.
Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32(1):268–274.
Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. *Nature* 284(5757):604–607.
Pardue ML, DeBaryshe PG. 2011. Retrotransposons that maintain chromosome ends. *Proc Natl Acad Sci U S A*. 108(51):20317–20324.
Parhad SS, Tu S, Weng Z, Theurkauf WE. 2017. Adaptive evolution leads to cross-species incompatibility in the piRNA transposon silencing machinery. *Dev Cell*. 43(1):60–70.e5.
Parsons P, Stanley S. 1981. Special ecological studies-domesticated and widespread species. In: Ashburner M, Carson L, Thompson JJ, editors. Vol. 3c. The genetics and biology of Drosophila. Oxford: Academic Press. p. 349–393.
Pascual L, Periquet G. 1991. Distribution of hobo transposable elements in natural populations of *Drosophila melanogaster*. *Mol Biol Evol*. 8(3):282–296.
Peccoud J, Loiseau V, Cordaux R, Gilbert C. 2017. Massive horizontal transfer of transposable elements in insects. *Proc Natl Acad Sci U S A*. 114(18):4721–4726.
Periquet G, Hamelin MH, Bigot Y, Lepissier A. 1989. Geographical and historical patterns of distribution of hobo elements in *Drosophila melanogaster* populations. *J Evol Biol*. 2(3):223–229.
Periquet G, Lemeunier F, Bigot Y, Hamelin MH, Bazin C, Ladevéze V, Eeken J, Galindo MI, Pascual L, Boussy I. 1994. The evolutionary genetics of the hobo transposable element in the *Drosophila melanogaster* complex. *Genetica* 93(1–3):79–90.
Petrov DA, Fiston-Lavier A-S, Lipatov M, Lenkov K, González J. 2011. Population genomics of transposable elements in *Drosophila melanogaster*. *Mol Biol Evol*. 28(5):1633–1644.
Post C, Clark JP, Sytnikova YA, Chirn G-W, Lau NC. 2014. The capacity of target silencing by *Drosophila* PIWI and piRNAs The capacity of target silencing by *Drosophila* PIWI and piRNAs. *RNA* 20(12):1977–1986.
Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, Anxolabéhère D. 2005. Combined evidence annotation of transposable elements in genome sequences. *PLoS Comp Biol*. 1(2):e22–e175.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.
R Core Team. 2012. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Rahman R, Chirn GW, Kanodia A, Sytnikova YA, Brembs B, Bergman CM, Lau NC. 2015. Unique transposon landscapes are pervasive across *Drosophila melanogaster* genomes. *Nucleic Acids Res*. 43(22):10655–10672.
Riddle NC, Minoda A, Kharchenko PV, Alekseyenko AA, Schwartz YB, Tolstorukov MY, Gorchakov AA, Jaffe JD, Kennedy C, Linder-Basso D, et al. 2011. Plasticity in patterns of histone modifications and chromosomal proteins in *Drosophila* heterochromatin. *Genome Res*. 21(2):147–163.
Rogers RL, Cridland JM, Shao L, Hu TT, Andolfatto P, Thornton KR. 2014. Landscape of standing variation for tandem duplications in *Drosophila yakuba* and *Drosophila simulans*. *Mol Biol Evol*. 31(7):1750–1766.
Rozhkov NV, Hammell M, Hannon GJ. 2013. Multiple roles for Piwi in silencing *Drosophila* transposons. *Genes Dev*. 27(4):400–412.
Ruebenbauer A, Schlyter F, Hansson BS, Löfstedt C, Larsson MC. 2008. Genetic variability and robustness of host odor preference in *Drosophila melanogaster*. *Curr Biol*. 18(18):1438–1443.
Saint-Leandre B, Levine MT. 2020. The telomere paradox: stable genome preservation with rapidly evolving proteins. *Trends Genet*. 36(4):232–242.
Sánchez-Gracia A, Maside X, Charlesworth B. 2005. High rate of horizontal transfer of transposable elements in *Drosophila*. *Trends Genet*. 21(4):200–203.
Schrempf D, Minh BQ, De Maio N, von Haeseler A, Kosiol C. 2016. Reversible polymorphism-aware phylogenetic models and their application to tree inference. *J Theor Biol*. 407:362–370.
Schrider D, Ayroles J, Matute D, Kern A. 2018. Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. *PLoS Genet*. 14(4):e1007341.
Serrato-Capuchina A, Wang J, Earley E, Peede D, Isbell K, Matute D. 2020. Paternally inherited P-element copy number affects the magnitude of hybrid dysgenesis in *Drosophila simulans* and *D. melanogaster*. *Genome Biol Evol*. 12(6):808–826.

Sienski G, Dönertas D, Brennecke J. 2012. Transcriptional silencing of transposons by Piwi and Maelstrom and its impact on chromatin state and gene expression. *Cell* 151(5):964–980.
Silva JC, Loreto EL, Clark JB. 2004. Factors that affect the horizontal transfer of transposable elements. *Curr Issues Mol Biol*. 6(1):57–72.
Simmons GM. 1992. Horizontal transfer of hobo transposable elements within the *Drosophila melanogaster* species complex: evidence from DNA sequencing. *Mol Biol Evol*. 9(6):1050–1060.
Smit AFA, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0.
Sniegowski PD, Charlesworth B. 1994. Transposable element numbers in cosmopolitan inversions from a natural population of *Drosophila melanogaster*. *Genetics* 137(3):815–827.
Song SU, Kurkulos M, Boeke JD, Corces VG. 1997. Infection of the germ line by retroviral particles produced in the follicle cells: a possible mechanism for the mobilization of the Gypsy retroelement of *Drosophila*. *Development* 124(14):2789–2798.
Srivastav SP, Kelleher ES. 2017. Paternal induction of hybrid dysgenesis in *Drosophila melanogaster* is weakly correlated with both P-element and Hob element dosage. *G3 (Bethesda)* 7(5):1487–1497.
Srivastav SP, Rahman R, Ma Q, Pierre J, Bandyopadhyay S, Lau NC. 2019. Har-P, a short P-element variant, weaponizes P-transposase to severely impair *Drosophila* development. *eLife* 26(6):e49948.
Stewart NB, Rogers RL. 2019. Chromosomal rearrangements as a source of new gene formation in *Drosophila yakuba*. *PLoS Genet*. 15(9):e1008314.
Terzian C, Ferraz C, Demaille J, Bucheton A. 2000. Evolution of the Gypsy endogenous retrovirus in the *Drosophila melanogaster* subgroup. *Mol Biol Evol*. 17(6):908–914.
Terzian C, Pélisson A, Bucheton A. 2001. Evolution and phylogeny of insect endogenous retroviruses. *BMC Evol Biol*. 1(1):3.
Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, Marygold SJ, Matthews BB, Millburn G, Antonazzo G, Trovisco V, et al. 2019. FlyBase 2.0: the next generation. *Nucleic Acids Res*. 47(D1):D759–D765.
Turissini DA, Liu G, David JR, Matute DR. 2015. The evolution of reproductive isolation in the *Drosophila yakuba* complex of species. *J Evol Biol*. 28(3):557–575.
Vieira C, Lepetit D, Dumont S, Biémont C. 1999. Wake up of transposable elements following *Drosophila simulans* worldwide colonization. *Mol Biol Evol*. 16(9):1251–1255.
Viggiano L, Caggese C, Barsanti P, Caizzi R. 1997. Cloning and characterization of a copy of Tirant transposable element in *Drosophila melanogaster*. *Gene* 197(1-2):29–35.
Wang L, Barbash DA, Kelleher ES. 2020. Adaptive evolution among cytoplasmic piRNA proteins leads to decreased genomic autoimmunity. *PLoS Genet*. 16(6):e1008861–e1008922.
Wang L, Dou K, Moon S, Tan FJ, Zhang ZZZ, Wang L, Dou K, Moon S, Tan FJ, Zhang ZZZ. 2018. Hijacking oogenesis enables massive propagation article hijacking oogenesis enables massive propagation of LINE and retroviral transposons. *Cell* 174(5):1082–1094.
Weilguny L, Kofler R. 2019. DeviaTE: assembly-free analysis and visualization of mobile genetic element composition. *Mol Ecol Resour*. 19(5):1346–1354.
Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 8(12):973–982.
Wierzbicki F, Schwarz F, Cannalonga O, Kofler R. 2020. Generating high quality assemblies for genomic analysis of transposable elements. *bioRxiv*.
Yang P, Wang Y, Macfarlan TS. 2017. The role of KRAB-ZFPs in transposable element repression and mammalian evolution. *Trends Genet*. 33(11):871–881.
Yannopoulos G, Stamatis N, Monastirioti M, Hatzopoulos P, Louis C. 1987. hobo is responsible for the induction of hybrid dysgenesis by strains of *Drosophila melanogaster* bearing the male recombination factor 23.5MRF. *Cell* 49(4):487–495.

Discussion

Towards comprehensive analyses of piRNA cluster dynamics

In chapter 1, I present novel, specialized metrics to evaluate the quality of genomic assemblies regarding the representation of repetitive elements and regions. I apply these novel metrics to create genomic assemblies of two *D. melanogaster* strains with an optimized representation of TEs and annotated piRNA clusters. Additionally, I employ these metrics to evaluate the structure of the piRNA cluster 42AB in these assemblies as well as the reference genome of *D. melanogaster*. I thus demonstrate that my quality metrics can not only be used to globally infer the quality of an assembly, but can also be employed to directly infer the assembly quality of a specific region. As previously available quality metrics to evaluate assembly quality were largely uninformative about TEs and repetitive regions, an empirical comparison of piRNA clusters was unable to distinguish natural polymorphisms from technical artefacts or missing sequence. This lack of ability to reliably infer the degree of polymorphisms in piRNA clusters from available genome assemblies was a major motivation for this study. My novel quality metrics will now allow researchers to reliably determine polymorphisms in piRNA clusters by ensuring that all compared regions are reliably assembled. The degree of polymorphism of TE insertions in piRNA clusters within and between populations of *D. melanogaster* is still major open question in TE research (Liu et al., 2014; Zhang et al., 2020). A prominent theoretical model which predicts TE-host interactions with piRNA silencing is called the 'trap model'. The trap model is build on the assumption that the successful establishment of piRNA-mediated prevention of TE activity in the presence of functional TE copies necessitates an insertion of the respective TE in a piRNA cluster (Bergman et al., 2006; Malone and Hannon, 2009). The trap model assumes that piRNA clusters are the exclusive source regions for piRNA production and predicts that a single insertion of a TE into a piRNA cluster will result in the production of a sufficient level of piRNAs to prevent the activity of the respective TE (Ronsseray et al., 1991; Josse et al., 2007; Zanni et al., 2013). Recent simulation studies have more precisely predicted the dynamics expected to be observed under the classic trap model. They predict that within natural populations it is likely to observe various segregating, compensating piRNA cluster insertions (Kelleher et al., 2018; Kofler, 2019). For example, Kofler (2019) predicts that

a population will need several independent piRNA cluster insertions to effectively silence a respective TE. These simulations further predict that these insertions will be segregating within a population for quite some time until one of them will be randomly driven to fixation within the population by drift.

Testing these predictions in a future experiment by comparing reliably assembled piRNA clusters in different assemblies (i.e. different populations) of *D. melanogaster* could unravel the true population dynamics of piRNA clusters. Complications for testing these predictions could arise due to paramutations (i.e. piRNA-producing TE insertions outside of piRNA clusters) as well as the existence of additional piRNA clusters not included in the CUSCO set. This could be accounted for, e.g. by the inclusion of small RNA sequencing data and the annotation of novel piRNA clusters. Within this experimental framework it would then also be possible to test for a potential correlation of the degree of transposition activity and the abundance of corresponding piRNAs. Such a correlation is predicted to be absent under the classical trap model, as a single insertion should produce sufficient piRNAs (Josse et al., 2007). An analysis following this approach is currently being undertaken within our laboratory. First inferences based on the reliably assembled piRNA clusters seem to dissent from theoretical and simulation predictions (Filip Wierzbicki, personal communication). However, these inferences are thus far incomplete and the evaluation of the trap model will require additional scrutiny. Regardless, I think that the empirical data (i.e. the genome assemblies) as well as the methodological advances (i.e. the global and local assembly quality metrics) I established in this thesis are valuable tools that will finally allow researchers to unravel the complex dynamics of piRNA clusters.

The future of genome assemblies

In chapter 1, I create two genome assemblies with a high quality, specifically in terms of their representation of TEs and piRNA cluster. So far, the creation of a high-quality genome assembly was challenging. Thus, even to date high-quality reference genomes are rather rare and mainly exist for model organisms like humans (Lander et al., 2001), *D. melanogaster* (Hoskins et al., 2015) or *A. thaliana* (The Arabidopsis Genome Initiative, 2000). However, with the advent of technologies like long-read sequencing, the generation of high-quality genome assemblies

is steadily becoming more accessible (Wong et al., 2018; Levy-Sakin et al., 2019; Almarri et al., 2020; Kim et al., 2021). Notably, the recent human telomere-to-telomere assembly (Miga et al., 2020) seems even more contiguous than the human reference genome (see also Chapter 1).As I demonstrated in chapter 1, assessing the quality of a genome assembly is challenging, because any quality metric only describes 'quality' in a very narrow scope. For example, BUSCO focuses solely on the representation of conserved genes, usually within euchromatic stretches of the genome. Similarly, NG50 solely describes the length of the largest contiguous stretches within an assembly, without considering potential misassemblies underlying these values. Additionally, as illustrated in the previous section, measures of global assembly quality do not necessarily reflect the assembly quality of a specific region of interest. Thus, I believe that a major challenge for comparative genomic research in the near future will not be the creation of genome assemblies, but rather to find criteria on which to decide which genome assemblies are of high-enough quality to be included within an analysis. I argue that my newly established quality metrics could be extended to become a cornerstone of a routinely applied pipeline to test the quality of any newly created genome assemblies. Towards the end of Chapter 1, I already demonstrate that my assembly quality metrics developed for the application in *D. melanogaster* can be extended to different species. I use existing TE consensus sequences and existing annotations of piRNA cluster regions to calculate TE landscape metrics and CUSCO for various human assemblies. Additionally, I demonstrate that the approach underlying CUSCO can be used to compare the assembly quality of other regions of interest than piRNA clusters by using annotations of the Knot-Engaged Element regions forming the KNOT structure in *A. thaliana* (Grob et al., 2014). Since my quality metrics require either an existing library of TE consensus sequences and/or conserved flanking sequences of regions of interest, their most straightforward application will likely be in studies inferring population-variation or individual variation in organisms with pre-existing genome assemblies (from which the flanking sequences can be inferred) and available TE consensus sequences. However, I think that my metrics could also be used routinely in the creation of novel genome assemblies for species without a reference genome . For example, genome assemblies of closely related species could be used to infer flanking sequences. Similarly, TE consensus sequences could be inferred *de-novo*, or consensus sequences of related species could be used for inferences (e.g. like *D. melanogaster*

consensus sequences are often used when comparing TE abundance within the *Drosophila* species group (e.g. Kofler et al. (2015)). Ultimately, the community could build a database of CUSCO annotations for a variety of species, similar to the large BUSCO databases. Such a database could then allow researchers to routinely employ CUSCO or TE-landscape approaches within their assembly pipelines.

Further unravelling of the complex dynamics of Tirant

In chapter 2, I describe the recent invasion of the LTR retrotransposon Tirant into the genome of *D. melanogaster*. I called the invasion 'stealthy' to contrast the invasion of Tirant with the other known recent invasions of TEs in *D. melanogaster*, the P-element, the I-element and hobo. The invasions of these three TEs were discovered and described in detail several decades ago (Kidwell, 1983; Daniels et al., 1990a,b; Bucheton et al., 1992). Contrarily, the Tirant invasion remained undetected (i.e. stealthy), despite the high amount of studies describing TE dynamics in *D. melanogaster*. It is possible that the late discovery of Tirant is indicative of an important difference between Tirant and the other recently invaded TEs. Within chapter 2, I showed that Tirant did not induce HD phenotypes in the investigated crosses. The discovery of HD usually led to the discovery of the other TE invasions. Thus, the lack of HD phenotypes induced by Tirant mobilization could explain the late discovery of the Tirant invasion. However, it is possible that Tirant mobilization induces HD symptoms, but the strains and environmental conditions explored within chapter 2 were not sufficient to induce such phenotypic effects. It is also possible, that piRNAs derived from degraded Tirant sequences are sufficient to prevent strong phenotypic consequences of Tirant activation or can even prevent Tirant activation altogether. Generally, the molecular mechanism of Tirant activity are still partially elusive (Malone et al., 2009). For example, it is possible that Tirant activity is restricted to a certain developmental timeframe and/or certain tissues. If Tirant activity strongly differs compared to the other HD-inducing TEs, resulting phenotypic consequences (i.e. HD phenotypes) might also be inherently different. As only phenotypes observed in previously described HD systems were examined in Chapter 2, different dynamics of Tirant activity might thus also explain the observed absence of HD phenotypes. To determine if Tirant indeed shows inherently different

dynamics compared to the other recently invaded TEs, it would thus first be necessary to experimentally perform a detailed molecular examination of the dynamics of Tirant activity in a natural population. A potentially interesting future experiment to describe these dynamics could be the artificial introduction of functional Tirant into the genome of strains sampled before the Tirant invasion. Using an experimental evolution setup, researchers could recreate the natural invasion of Tirant in a controlled laboratory environment. Analyses of piRNA levels during the experiment could quantify after how many generations piRNA defense against the novel Tirant is established as well as test if residual piRNAs are relevant for preventing Tirant activity. Performing the experiment in a replicated manner at different temperatures could reveal environmental influences on Tirant activity. Also, the examination of Tirant activity in flies sampled at different developmental stages or in specific tissues, e.g. using single-cell RNA-seq, could allow for a detailed characterization of the molecular mechanism of Tirant activity. Finally, using several strains with different genetic backgrounds could help to classify the suitability of different strains to induce Tirant activity, similar to the P, Q and M strain classification used to indicate susceptibility to P-element-induced HD (Kidwell, 1983). Overall, such an experiment would help to continue the detailed examination of the dynamics and mechanisms of Tirant activity conducted in this work and further improve our understanding of the dynamics of Tirant activity in natural populations.

A generalized workflow to unravel complex TE dynamics

I believe that the description of the Tirant invasion in chapter 2 exemplifies how to empirically unravel seemingly cryptic TE dynamics. The characterization of the recent, stealthy invasion of Tirant in *D. melanogaster* is unprecedentedly detailed due to the combination of numerous resources. I create a highly informative dataset combining newly produced and publicly available DNA, RNA and small RNA sequencing data of long-established as well as recently collected fly strains, representing worldwide extant natural populations as well as historic snapshots of past populations during the last century. I additionally utilize the novel high-quality genome assemblies produced in chapter 1, allowing to determine the genomic position and relative age of Tirant insertions in different strains. This approach not only provides a narrow estimate

of the timeframe of the Tirant invasion, but also to reconstruct previously published findings regarding the timeframes for other recent TE invasions in *D. melanogaster*. This indicates that the established approach could be used to unravel the detailed dynamics of any recently active TE in natural populations. I would like to highlight that if even in *D. melanogaster*, arguably the best studied organism regarding TE invasions, an invasion as recent as 100 years ago was overlooked so far, the potential for discoveries of recent TE invasions in other, less well studied species utilizing a similar approach could be immense. Thus, I believe that establishing a catalogue combining time-series data with modern genomic and transcriptomic analyses and high-quality genome assemblies for a variety of species could greatly improve our understanding of how and why TE invasions occur. Certain species or populations might be identified to show significantly higher or lower frequency of recent TE invasions, allowing the characterization of shared underlying ecological or genetic properties influencing the frequency and intensity of TE invasions. Establishing such datasets for a variety of organisms could thus finally allow detailed characterizations of the natural dynamics of many different TEs. Thus, these analyses could be able to finally unravel the full degree of variability of TE dynamics in different populations and species.

Concluding remarks

In this thesis, I developed novel methodologies and established new resources to unravel complex TE dynamics. Utilizing these resources, I discover a novel, recent TE invasion in the highly studied model organism *D. melanogaster* and provide a detailed characterization of its invasion dynamics. I additionally present first biological inferences about piRNA cluster dynamics. While most inferences in this work are restricted to *D. melanogaster*, I demonstrate how the underlying methodologies and developed tools can be applied to a broad range of questions in various organisms. In summary, this thesis not only provides novel biological insight into peculiarities of TE dynamics, but also lays the groundwork for future research, particularly for describing the dynamics of piRNA clusters and TE invasions.

References

Alkan, C., Sajjadian, S., and Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nature Methods*, 8:61–65.

Almarri, M. A., Bergström, A., Prado-Martinez, J., Yang, F., Fu, B., Dunham, A. S., Chen, Y., Hurles, M. E., Tyler-Smith, C., and Xue, Y. (2020). Population Structure, Stratification, and Introgression of Human Structural Variation. *Cell*, pages 189–199.

Aravin, A. A., Hannon, G. J., and Brennecke, J. (2007). The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science*, 318(5851):761–764.

Arkhipova, I. R. (2018). Neutral Theory, Transposable Elements, an Eukaryotic Genome Evolution. *Mol Biol Evol*, 35(6):1332–37.

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6(1):11.

Barrón, M. G., Fiston-Lavier, A.-S., Petrov, D. A., and González, J. (2014). Population Genomics of Transposable Elements in *Drosophila*. *Annual Review of Genetics*, 48(1):561–581.

Bartolomé, C., Maside, X., and Charlesworth, B. (2002). On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. *Molecular Biology and Evolution*, 19(6):926–937.

Bergman, C. M. and Bensasson, D. (2007). Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, 104(27):11340–11345.

Bergman, C. M., Quesneville, H., Anxolabéhère, D., and Ashburner, M. (2006). Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome biology*, 7(11):R112.

Biémont, C. and Vieira, C. (2006). Junk DNA as an evolutionary force. *Nature*, 443(7111):521–524.

Black, D. M., Jackson, M. S., Kidwell, M. G., and Dover, G. A. (1987). KP elements repress P-induced hybrid dysgenesis in *Drosophila melanogaster*. *The EMBO Journal*, 613(13):4125–4135.

Blackman, R. K., Grimaila, R., Macy, M., Koehler, D., and Gelbart, W. M. (1987). Mobilization of hobo elements residing within the decapentaplegic gene complex: Suggestion of a new hybrid dysgenesis system in *Drosophila melanogaster*. *Cell*, 49(4):497–505.

Blommaert, J., Riss, S., Hecox-Lea, B., Mark Welch, D. B., and Stelzer, C. P. (2019). Small, but surprisingly repetitive genomes: transposon expansion and not polyploidy has driven a doubling in genome size in a metazoan species complex. *BMC Genomics*, 20(1):466.

Blumenstiel, J. P. (2011). Evolutionary dynamics of transposable elements in a small RNA world. *Trends in Genetics*, 27(1):23–31.

Brandt, J., Schrauth, S., Veith, A.-M., Froschauer, A., Haneke, T., Schultheis, C., Gessler, M., Leimeister, C., and Volff, J.-N. (2005). Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. *Gene*, 345(1):101–111.

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G. J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell*, 128(6):1089–1103.

Brennecke, J., Malone, C. D., Aravin, A. A., Sachidanandam, R., Stark, A., and Hannon, G. J. (2008). An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science*, 322(5906):1387–1392.

Bucheton, A. (1979). Non-Mendelian female sterility in *Drosophila melanogaster*: Influence of aging and thermic treatments. III. Cumulative effects induced by these factors. *Genetics*, 93(1):131–142.

Bucheton, A., Lavige, J., Picard, G., and L'heritier, P. (1976). Non-mendelian female sterility in *Drosophila melanogaster*: quantitative variations in the efficiency of inducer and reactive strains. *Heredity*, 36(3):305–314.

Bucheton, A., Vaury, C., Chaboissier, M. C., Abad, P., Pélisson, A., and Simonelig, M. (1992). I elements and the *Drosophila* genome. *Genetica*, 86(1-3):175–190.

Canapa, A., Barucca, M., Biscotti, M. A., Forconi, M., and Olmo, E. (2015). Transposons, Genome Size, and Evolutionary Insights in Animals. *Cytogenetic and Genome Research*, 147(4):217–239.

Charlesworth, B. (1991). Transposable elements in natural populations with a mixture of selected and neutral insertion sites. *Genetical research*, 57(2):127–34.

Charlesworth, B. and Charlesworth, D. (1983). The population dynamics of transposable elements. *Genetical Research*, 42(01):1–27.

Charlesworth, B. and Langley, C. H. (1989). The population genetics of *Drosophila* transposable elements. *Annual review of genetics*, 23:251–87.

Daniels, S. B., Chovnick, A., and Boussyy, I. A. (1990a). Distribution of hobo transposable elements in the genus *Drosophila. Molecular Biology and Evolution*, 7(6):589–606.

Daniels, S. B., Peterson, K. R., Strausbaugh, L. D., Kidwell, M. G., and Chovnick, A. (1990b). Evidence for horizontal transmission of the P transposable element between *Drosophila* species. *Genetics*, 124(2):339–355.

Doolittle, W. F. and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284(5757):601–3.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Engels, W. R. (1983). The P family of transposable elements in *Drosophila*. *Annual review of genetics*, 17:315–44.

Feschotte, C. and Pritham, E. J. (2007). DNA Transposons and the Evolution of Eukaryotic Genomes. *Annual Review of Genetics*, 41(1):331–368.

Finnegan, D. J. (1992). Transposable elements. *Current Opinion in Genetics and Development*, 2(6):861–867.

Grob, S., Schmid, M., and Grossniklaus, U. (2014). Hi-C Analysis in *Arabidopsis* Identifies the *KNOT*, a Structure with Similarities to the *flamenco* Locus of *Drosophila*. *Molecular Cell*, 55(5):678–693.

Hof, A. E. t., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., Hall, N., Darby, A. C., and Saccheri, I. J. (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nature*, 534(7605):102–105.

Hoskins, R. A., Carlson, J. W., Wan, K. H., Park, S., Mendez, I., Galle, S. E., Booth, B. W., Pfeiffer, B. D., George, R. A., Svirskas, R., et al. (2015). The Release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome research*, 25(3):445–458.

Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C. A., Carretero-Paulet, L., Chang, T.-H., Lan, T., Welch, A. J., Juárez, M. J. A., Simpson, J., et al. (2013). Architecture and evolution of a minute plant genome. *Nature*, 498(7452):94–98.

Jakšić, A. M., Kofler, R., and Schlötterer, C. (2017). Regulation of transposable elements: Interplay between TE-encoded regulatory sequences and host-specific trans-acting factors in *Drosophila melanogaster*. *Molecular Ecology*, 26(19):5149–5159.

Jo, B.-S. and Choi, S. S. (2015). Introns: The Functional Benefits of Introns in Genomes. *Genomics & informatics*, 13(4):112–118.

Josse, T., Teysset, L., Todeschini, A.-L., Sidor, C. M., Anxolabéhère, D., and Ronsseray, S. (2007). Telomeric trans-silencing: an epigenetic repression combining RNA silencing and heterochromatin formation. *PLoS Genetics*, 3(9):1633–43.

Kaminker, J. S., Bergman, C. M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D. A., Lewis, S. E., Rubin, G. M., Ashburner, M., and Celniker, S. E. (2002). The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome biology*, 3(12).

Kazazian, H. H. (1998). Mobile elements and disease. *Current opinion in genetics & development*, 8(3):343–350.

Kelleher, E. S., Azevedo, R. B. R., and Zheng, Y. (2018). The Evolution of Small-RNA-Mediated Silencing of an Invading Transposable Element. *Genome Biology and Evolution*, 10(11):3038–57.

Kidwell, M. G. (1983). Evolution of hybrid dysgenesis determinants in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, 80(6):1655–1659.

Kidwell, M. G., Kidwell, J. F., and Sved, J. A. (1977). Hybrid dysgenesis in *Drosophila melanogaster*: A syndrome of aberrant traits including mutations, sterility and male recombination. *Genetics*, 86(4):813–833.

Kim, B. Y., Wang, J. R., Miller, D. E., Barmina, O., Delaney, E., Thompson, A., Comeault, A. A., Peede, D., D'Agostino, E. R. R., Pelaez, J., et al. (2021). Highly contiguous assemblies of 101 drosophilid genomes. *eLife*, 10:e66405.

Kofler, R. (2019). Dynamics of Transposable Element Invasions with piRNA Clusters. *Molecular Biology and Evolution*, 36(7):1457–1472.

Kofler, R., Betancourt, A. J., and Schlötterer, C. (2012). Sequencing of Pooled DNA Samples (Pool-Seq) Uncovers Complex Dynamics of Transposable Element Insertions in *Drosophila melanogaster*. *PLoS genetics*, 8(1):e1002487.

Kofler, R., Nolte, V., and Schlötterer, C. (2015). Tempo and mode of transposable element activity in *Drosophila*. *PLoS Genetics*, 11(7):e1005406.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Lee, C. C., Mul, Y. M., and Rio, D. C. (1996). The Drosophila P-element KP repressor protein dimerizes and interacts with multiple sites on P-element DNA. *Molecular and cellular biology*, 16(10):5616–5622.

Lee, S.-I. and Kim, N.-S. (2014). Transposable elements and genome size variations in plants. *Genomics & informatics*, 12(3):87–97.

Lee, Y. C. G. and Langley, C. H. (2010). Transposable elements in natural populations of *Drosophila melanogaster*. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1544):1219–28.

Levy-Sakin, M., Pastor, S., Mostovoy, Y., Li, L., Leung, A. K. Y., McCaffrey, J., Young, E., Lam, E. T., Hastie, A. R., Wong, K. H. Y., Chung, C. Y. L., Ma, W., Sibert, J., Rajagopalan, R., Jin, N., Chow, E. Y. C., Chu, C., Poon, A., Lin, C., Naguib, A., Wang, W.-P., Cao, H., Chan, T.-F., Yip, K. Y., Xiao, M., and Kwok, P.-Y. (2019). Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nature Communications*, 10(1):1025.

Liu, J., Schnakenberg, S. L., Xing, J., Chen, K. C., Song, J., and Ha, H. (2014). Variation in piRNA and Transposable Element Content in Strains of *Drosophila melanogaster*. *Genome Biology and Evolution*, 6(10):2786–2798.

López-Flores, I. and Garrido-Ramos, M. A. (2012). The repetitive DNA content of eukaryotic genomes. *Genome Dynamics*, 7:1–28.

Malone, C. D., Brennecke, J., Dus, M., Stark, A., McCombie, W. R., Sachidanandam, R., and Hannon, G. J. (2009). Specialized piRNA Pathways Act in Germline and Somatic Tissues of the *Drosophila* Ovary. *Cell*, 137(3):522–535.

Malone, C. D. and Hannon, G. J. (2009). Small RNAs as Guardians of the Genome. *Cell*, 136(4):656–668.

Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annual review of genomics and human genetics*, 7:29–59.

Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–564.

Mayer, K. F. X., Waugh, R., Langridge, P., Close, T. J., Wise, R. P., Graner, A., Matsumoto, T., Sato, K., Schulman, A., Muehlbauer, G. J., et al. (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature*, 491(7426):711–716.

McClintock, B. (1956). Controlling elements and the gene. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 21, pages 197–216. Cold Spring Harbor Laboratory Press.

McCullers, T. J. and Steiniger, M. (2017). Transposable elements in *Drosophila*. *Mobile Genetic Elements*, 7(3):1–18.

Medstrand, P., van de Lagemaat, L. N., Dunn, C. A., Landry, J.-R., Svenback, D., and Mager, D. L. (2005). Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenetic and genome research*, 110(1-4):342–352.

Mendel, G. (1866). Versuche Über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines zu Brünn*, 4:3–47.

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G. A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, 585(7823):79–84.

Mills, R. E., Bennett, E. A., Iskow, R. C., and Devine, S. E. (2007). Which transposable elements are active in the human genome? *Trends in Genetics*, 23(4):183–191.

Moore, G. P. (1984). The C-Value Paradox. *BioScience*, 34(7):425–429.

Nuthikattu, S., McCue, A. D., Panda, K., Fultz, D., DeFraia, C., Thomas, E. N., and Slotkin, R. K. (2013). The Initiation of Epigenetic Silencing of Active Transposable Elements Is Triggered by RDR6 and 21-22 Nucleotide Small Interfering RNAs. *Plant Physiology*, 162(1):116–131.

Ohno, S. (1972). So much "junk" DNA in our genome. In *Brookhaven Symp Biol*, volume 23, pages 366–370.

Pardue, M.-L. and DeBaryshe, P. G. (2003). Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annual review of genetics*, 37:485–511.

Platt, R. N., Vandewege, M. W., and Ray, D. A. (2018). Mammalian transposable elements and their impacts on genome evolution. *Chromosome Research*, 26(1-2):25–43.

Ronsseray, S., Lehman, M., and Anxolabéhère, D. (1991). The Maternally Inherited Regulation of P Elements in *Drosophila melanogaster* Can Be Elicited by Two P Copies at Cytological Site 1A on the X Chromosome. *Genetics*, 129:501–512.

Rowe, H. M., Jakobsson, J., Mesnard, D., Rougemont, J., Reynard, S., Aktas, T., Maillard, P. V., Layard-Liesching, H., Verp, S., Marquis, J., Spitz, F., Constam, D. B., and Trono, D. (2010). KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature*, 463(7278):237–240.

Ruiz, M. T. and Carareto, C. M. A. (2003). Copy number of P elements, KP/full-sized P element ratio and their relationships with environmental factors in Brazilian Drosophila melanogaster populations. *Heredity*, 91(6):570–576.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science (New York, N.Y.)*, 326(5956):1112–1115.

Schrader, L. and Schmitz, J. (2019). The impact of transposable elements in adaptive evolution. *Molecular Ecology*, 28(6):1537–1549.

Sedlazeck, F. J., Lee, H., Darby, C. A., and Schatz, M. C. (2018). Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, 19(6):329–346.

Seppey, M., Manni, M., and Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods in molecular biology*, 1962:227–245.

Shanmugam, A., Nagarajan, A., and Pramanayagam, S. (2017). Non-coding DNA – a brief review. *Journal of Applied Biology & Biotechnology*, 5(05):42–47.

Sigman, M. J. and Slotkin, R. K. (2016). The First Rule of Plant Transposable Element Silencing: Location, Location, Location. *The Plant cell*, 28(2):304–313.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814):796–815.

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A., and Kakutani, T. (2009). Bursts of retrotransposition reproduced in arabidopsis. *Nature*, 461(7262):423–426.

Urrutia, R. (2003). KRAB-containing zinc-finger repressor proteins. *Genome Biology*, 4(10):231.

Wegrzyn, J. L., Lin, B. Y., Zieve, J. J., Dougherty, W. M., Martínez-García, P. J., Koriabine, M., Holtz-Morris, A., DeJong, P., Crepeau, M., Langley, C. H., Puiu, D., Salzberg, S. L., Neale, D. B., and Stevens, K. A. (2013). Insights into the Loblolly Pine Genome: Characterization of BAC and Fosmid Sequences. *PLOS ONE*, 8(9):e72439.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12):973–982.

Wong, K. H. Y., Levy-Sakin, M., and Kwok, P.-Y. (2018). De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nature Communications*, 9(1):3040.

Wong, W. Y., Simakov, O., Bridge, D. M., Cartwright, P., Bellantuono, A. J., Kuhn, A., Holstein, T. W., David, C. N., Steele, R. E., and Martínez, D. E. (2019). Expansion of a single transposable element family is associated with genome-size increase and radiation in the genus *Hydra*. *Proceedings of the National Academy of Sciences*, 116(46):22915 LP – 22917.

Yannopoulos, G., Stamatis, N., Monastirioti, M., Hatzopoulos, P., and Louis, C. (1987). hobo is responsible for the induction of hybrid dysgenesis by strains of *Drosophila melanogaster* bearing the male recombination factor 23.5MRF. *Cell*, 49(4):487–495.

Zanni, V., Eymery, A., Coiffet, M., Zytnicki, M., Luyten, I., Quesneville, H., Vaury, C., and Jensen, S. (2013). Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. *Proceedings of the National Academy of Sciences*, 110(49):19842–19847.

Zhang, S., Pointer, B., and Kelleher, E. S. (2020). Rapid evolution of piRNA-mediated silencing of an invading transposable element was driven by abundant de novo mutations. *Genome research*, 30(4):566–575.

Appendix

Chapter 1

Supplementary figures and tables

August 2, 2021

1 Supplementary figures

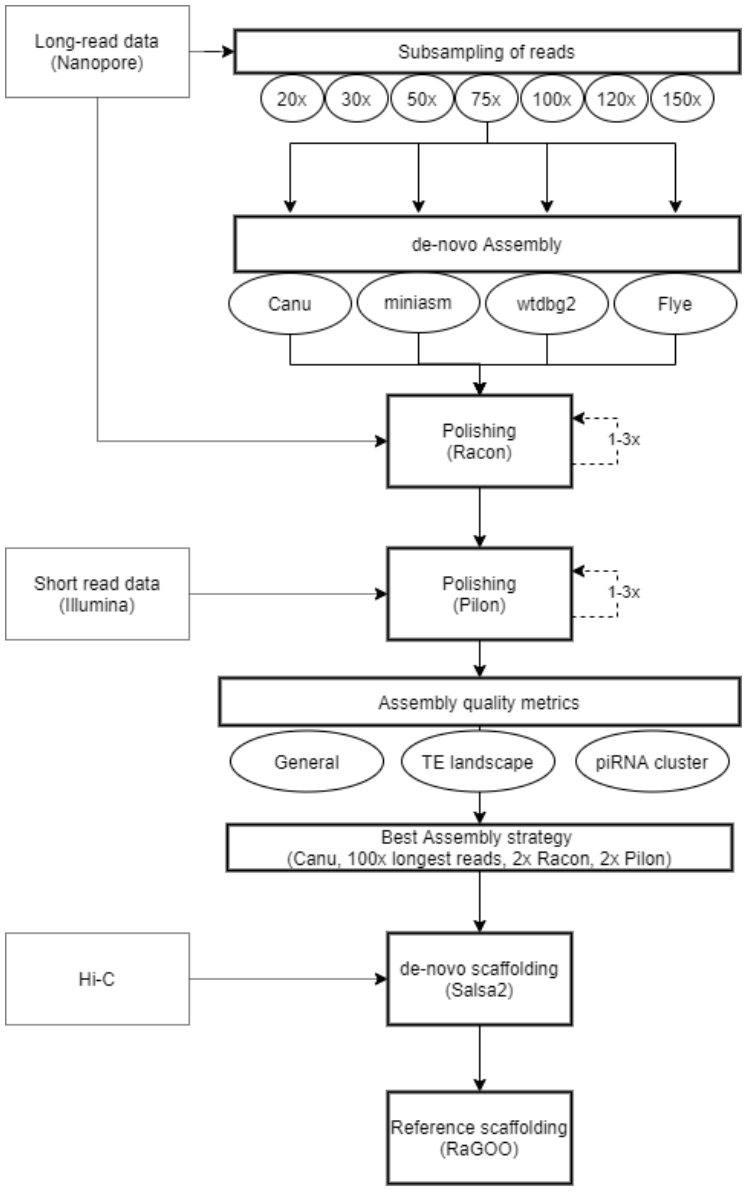

Figure S1: Overview of the assembly pipeline used in the manuscript.

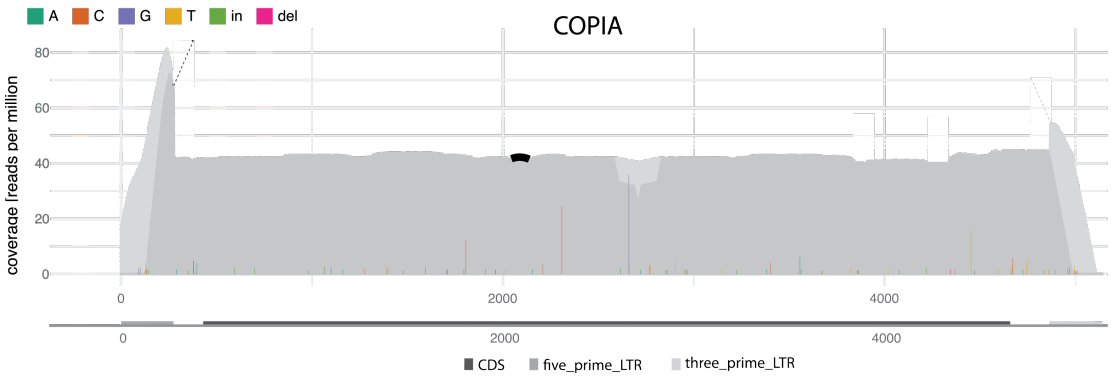

Figure S2: Abundance and diversity for *copia* elements in the *D. melanogaster* strain Canton-S. The coverage (TE abundance in rpm), the position of SNPs (colored lines) and the position of indels (bold arc at the top) are shown. The coverage based on unambiugously (dark grey) and ambiguously (light grey) mapped reads is shown. The plot was generated by DeviaTE (Weilguny and Kofler, 2019) based on Illumina reads mapped to the consensus sequence of *copia* (30 coverage; 2x125bp)

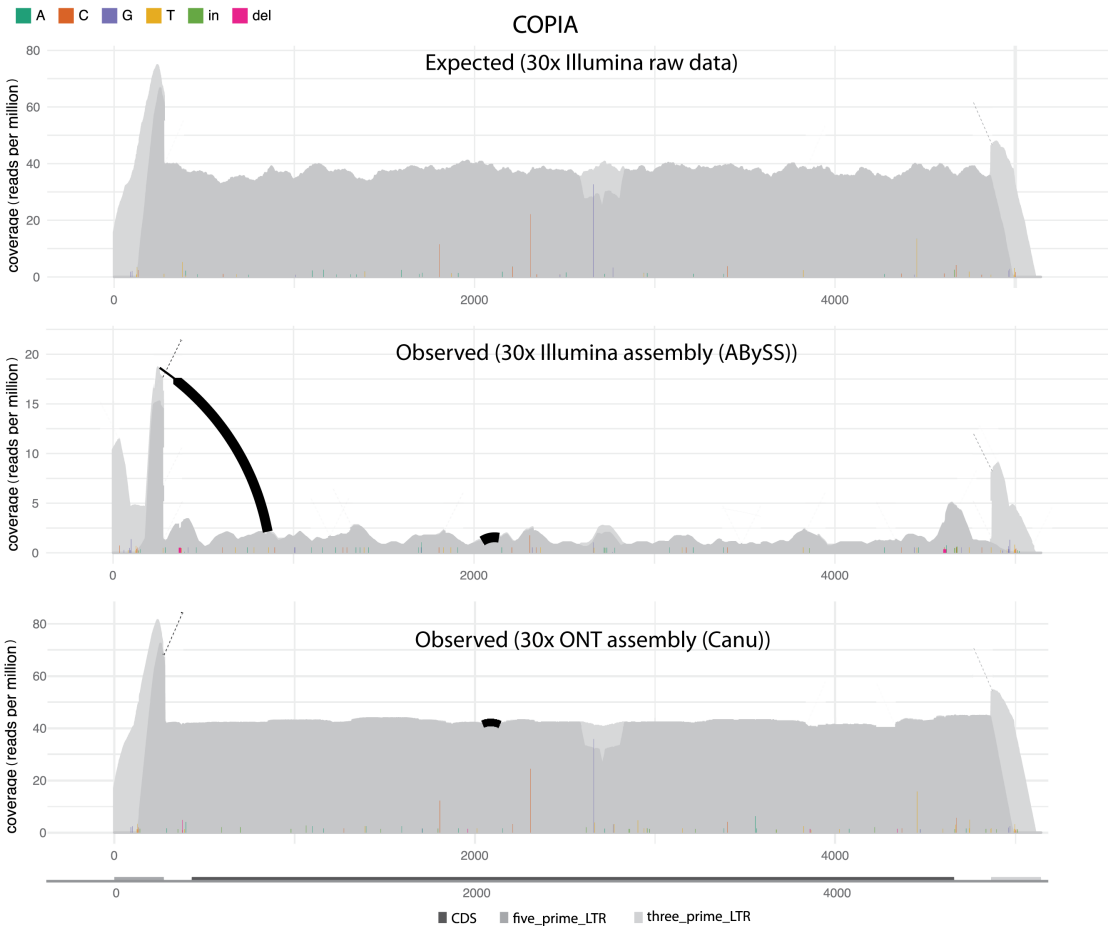

Figure S3: Expected and observed abundance and diversity of *copia* elements in Canton-S. Expected values are based on Illumina raw reads aligned to the consensus sequence of *copia*. Observed values are shown for assemblies based on short (ABySS) and long (Canu) reads. The normalised coverage is shown for ambiguously (light grey) and unambiguously (dark grey) mapped reads. The positions of SNPs (colored lines) and the position of indels (bold arcs) are shown. Note that both, the expected coverage (TE abundance) and diversity (SNPs and indels) of copia, are best reproduced by the long-read assembly (Canu).

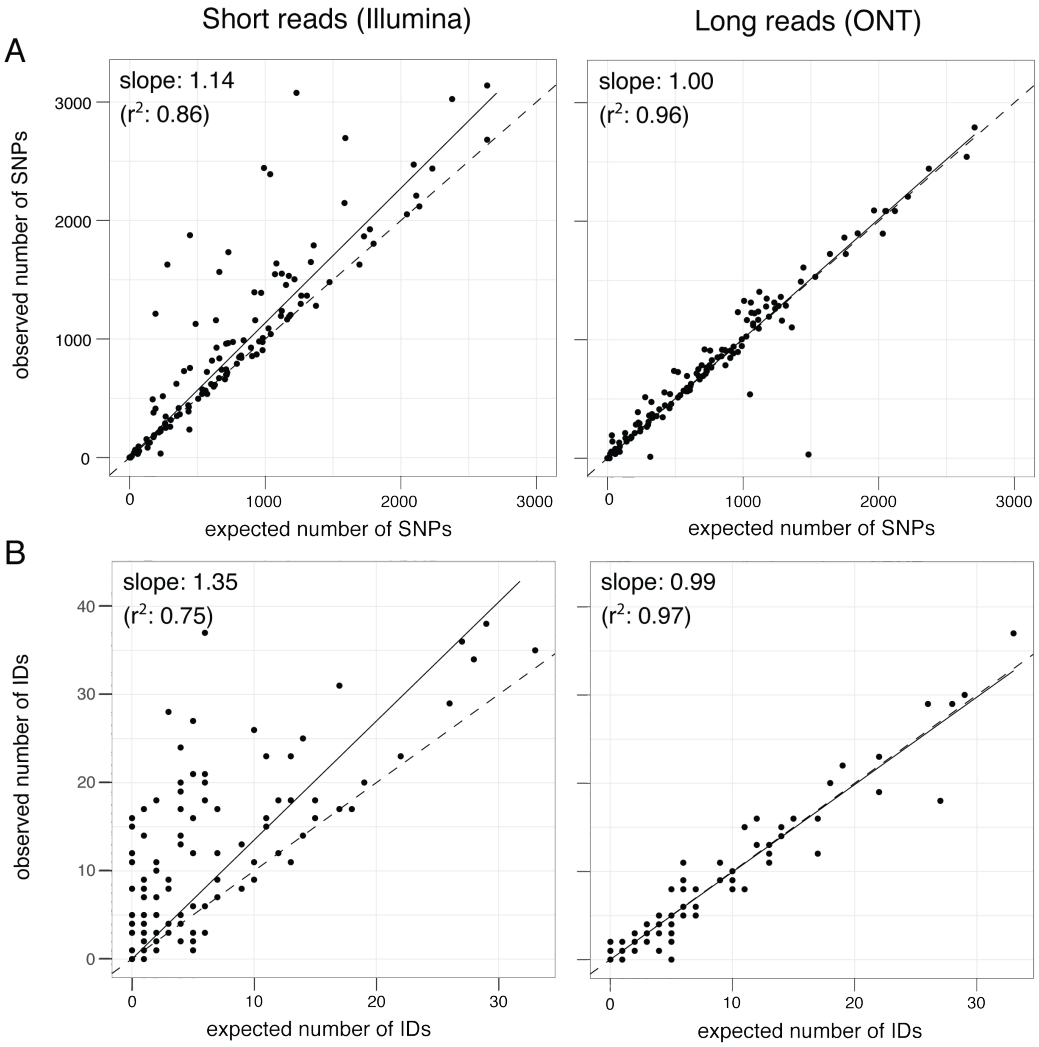

Figure S4: Expected and observed abundance of SNPs and IDs for a short-read and a long-read assembly of Canton-S. The slope of the regressions represent our novel quality metric for the abundance of SNPs and IDs. Each dot represents a distinct TE family and the dashed line shows the optimal representation of the TE. Note that the long-read assembly captures the abundance of SNPs and IDs more accurately than the short-read assembly, which overestimate the abundance of SNPs and IDs (slope > 1)

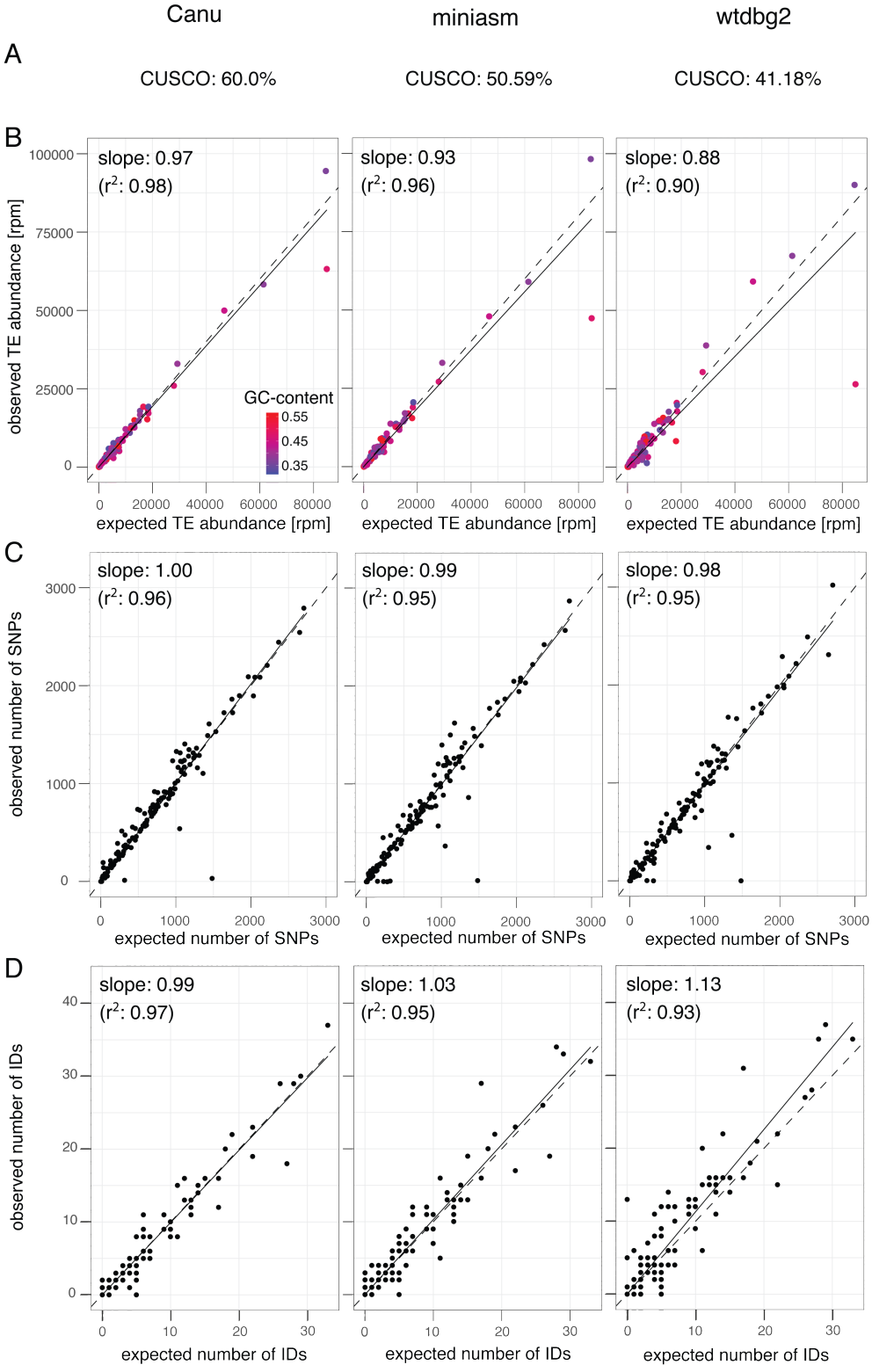

Figure S5: Influence of the assembly algorithm on the quality of assemblies of the *D. melanogaster* strain Canton-S. Assemblies are based on 30x coverage with ONT reads.

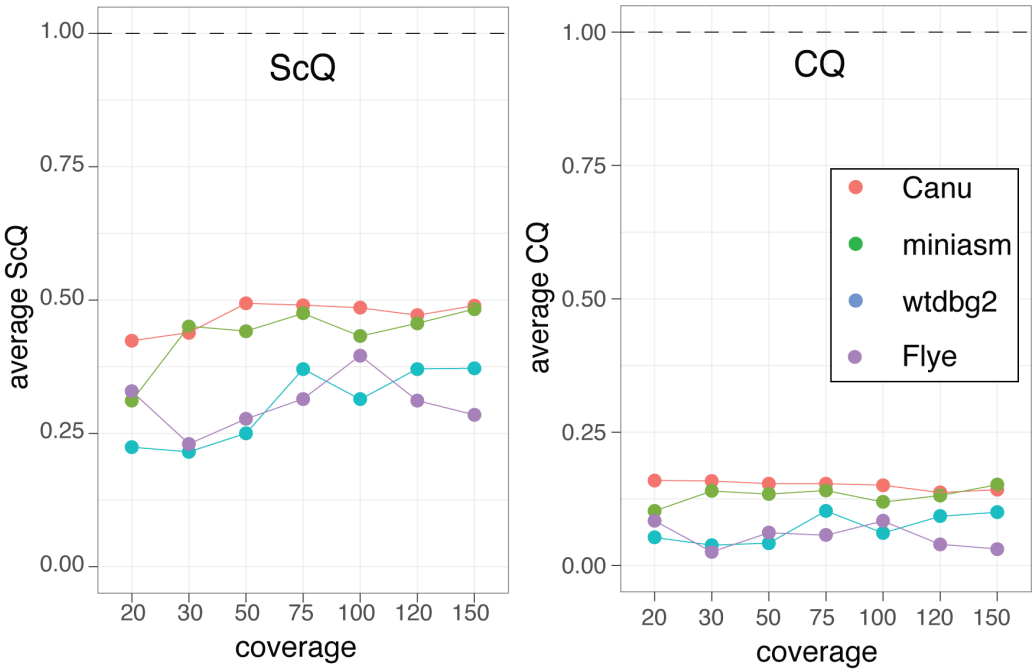

Figure S6: Influence of the assembly algorithm (Canu, miniasm, wtdbg2, Flye) and the coverage on the quality of assemblies. Results are shown for the average CQ (coverage quality) and the average ScQ (soft-clip quality). Note that Canu consistently generates the most reliable assemblies of piRNA clusters.

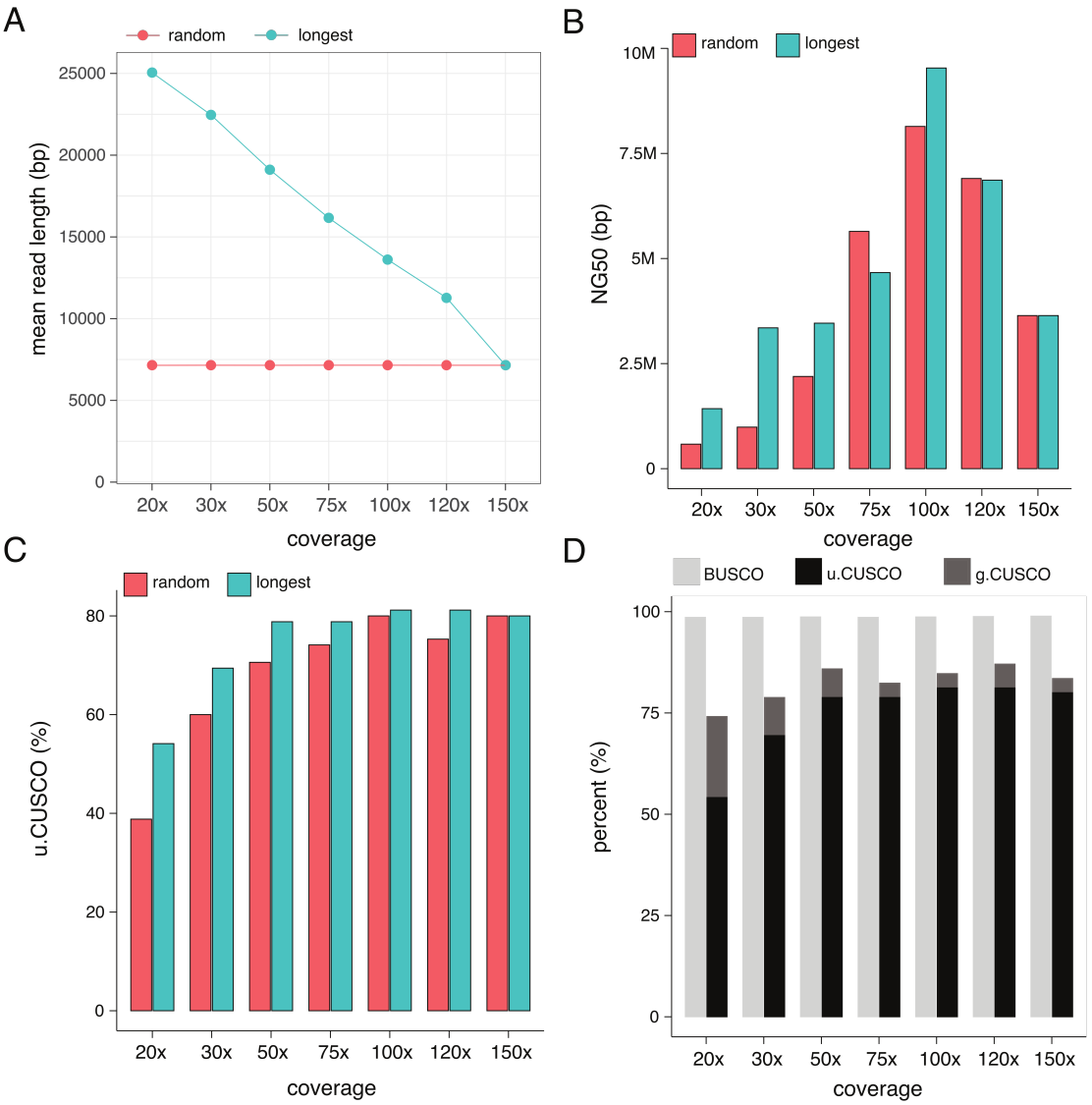

Figure S7: Assemblies with Canu have the highest quality when solely a subset of the longest reads is used. Assemblies were either based on a random subset of the reads (random) or on the longest reads (longest). A) Mean read length of the investigated subsets of reads. B) NG50 values of assemblies generated with different subsets of reads. C) ungapped-CUSCO values of assemblies generated with different subsets of reads. D) CUSCO and BUSCO values of assemblies generated with different subsets of the longest reads. Scaffolding was based on Hi-C data. Note that 150x coverage corresponds to the full dataset. Hence, no differences between assemblies based on random reads and the longest reads are expected at this coverage.

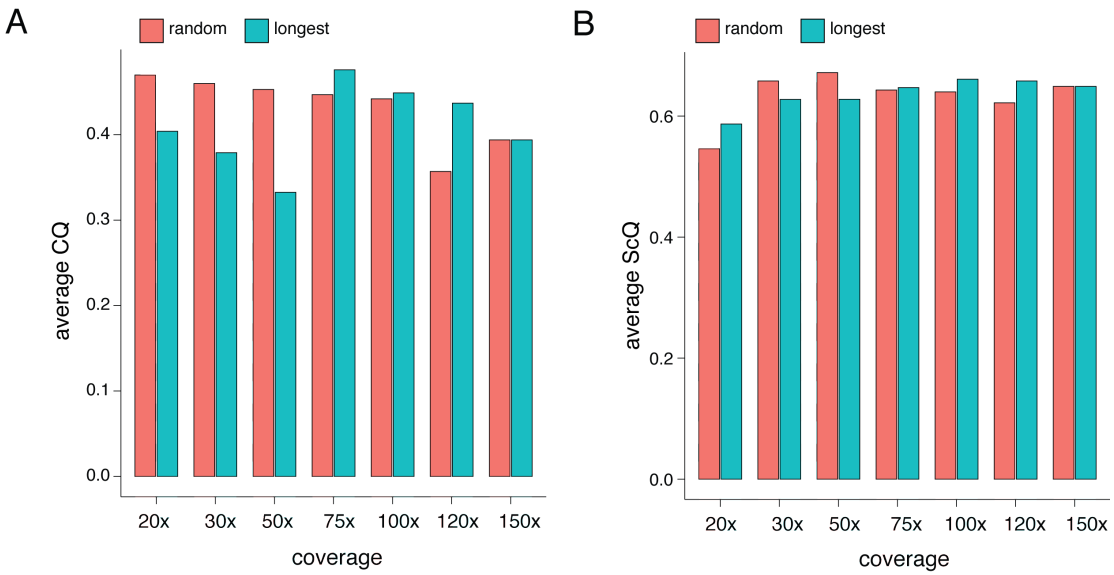

Figure S8: Quality of the assembled piRNA clusters with different subsets of reads. Either random reads (random) or the longest reads (longest) were used to generate assemblies with Canu. The assembly quality was assessed using the average CQ and the average ScQ.

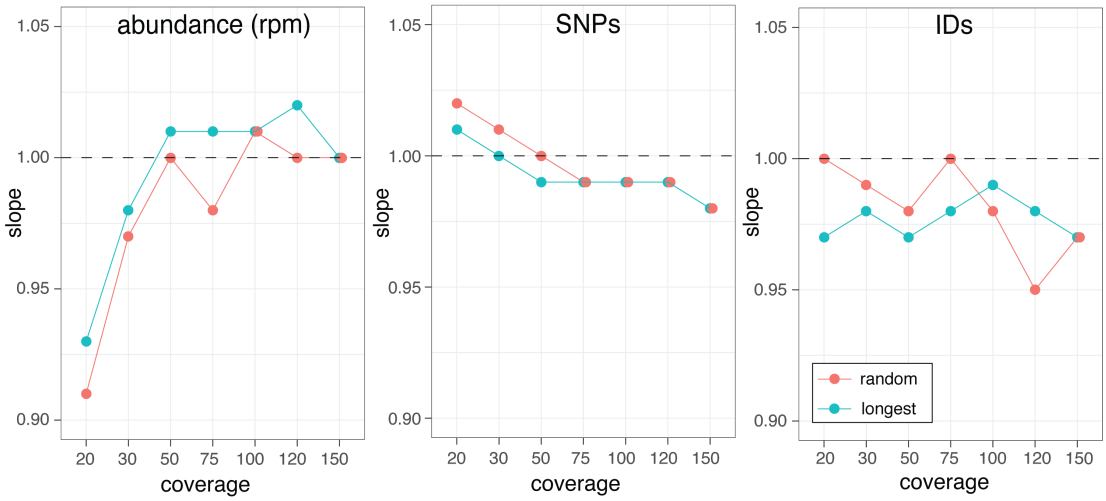

Figure S9: Assembly quality with different subsets of reads. Either random reads (random) or the longest reads (longest) were used to generate assemblies with Canu. The assembly quality is assessed using three of our TE-centered quality metrics (abundance, SNPs, IDs). The dashed lines indicate the optimal representation of TEs.

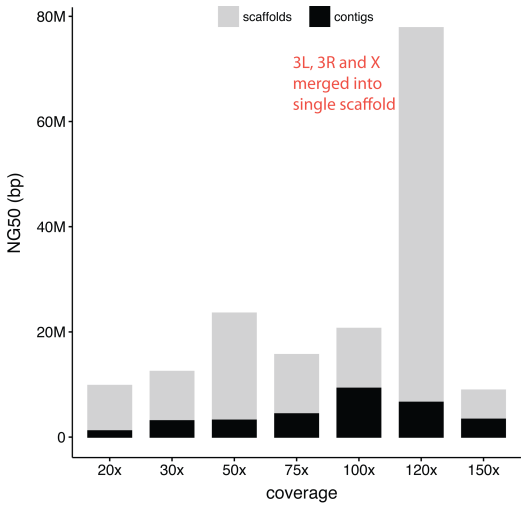

Figure S10: NG50 values of Canton-S assemblies generated with Canu and different subsamples of the longest reads. Values are shown before (contigs) and after Hi-C based scaffolding.

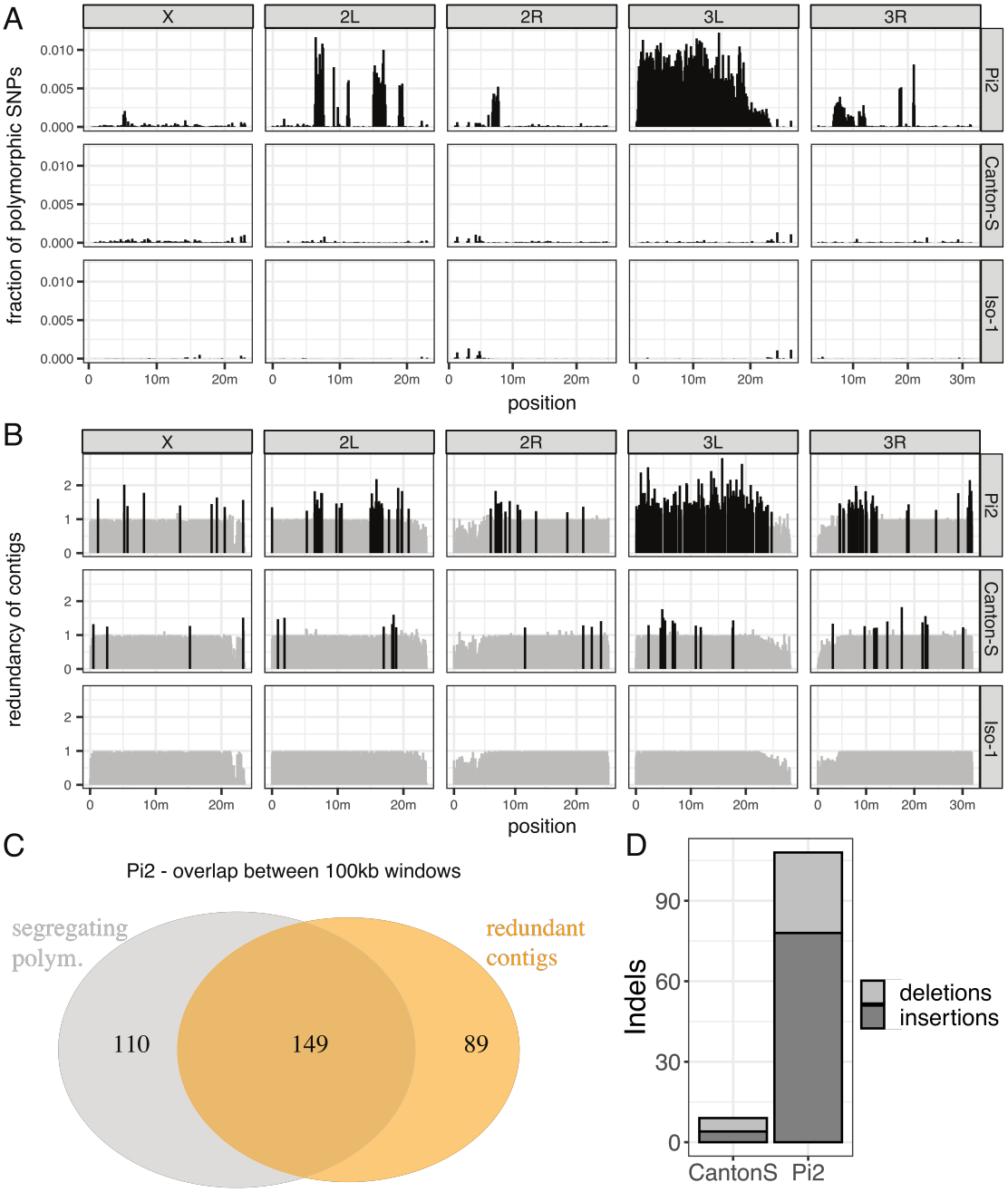

Figure S11: Segregating polymorphisms may lead to redundant contigs. A) The fraction of segregating SNPs for the *D. melanogaster* strains Canton-S and Pi2. The results are shown for 100kb windows. The highly isogenic strain Iso-1 is included as a reference. B) Origin of redundant contigs. Non-overlapping 1kb subsequences of an assembly were aligned to the reference. The average coverage per 100kb window is shown. Coverages > 1.2 indicate redundant contigs (shown in black). C) Redundant contigs are mostly found for windows with segregating polymorphism. D) Number of large indels ($\geq$ 1kb) in the assemblies of Canton-S and Pi2.

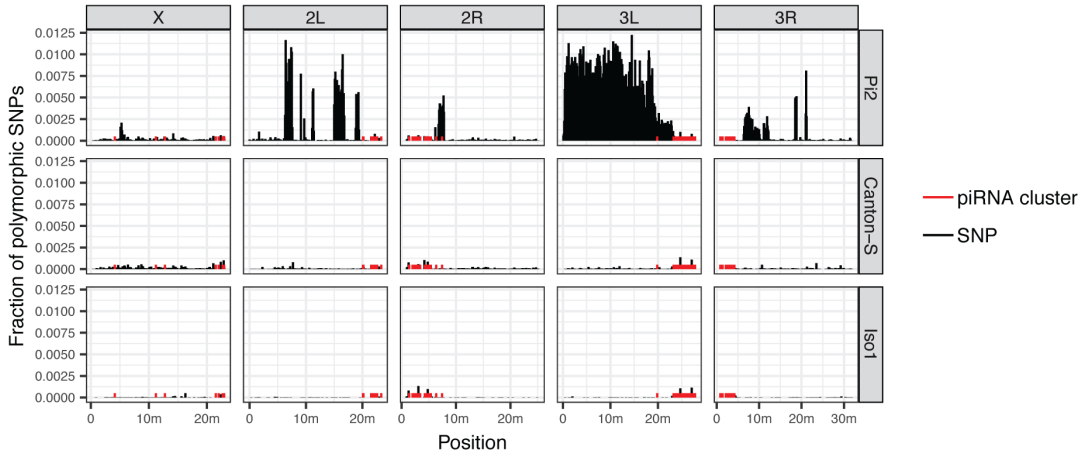

Figure S12: Location of piRNA clusters (red) and of regions with segregating polymorphisms (black) for several *D. melanogaster* strains. Segregating polymorphisms are shown for 100kb windows.

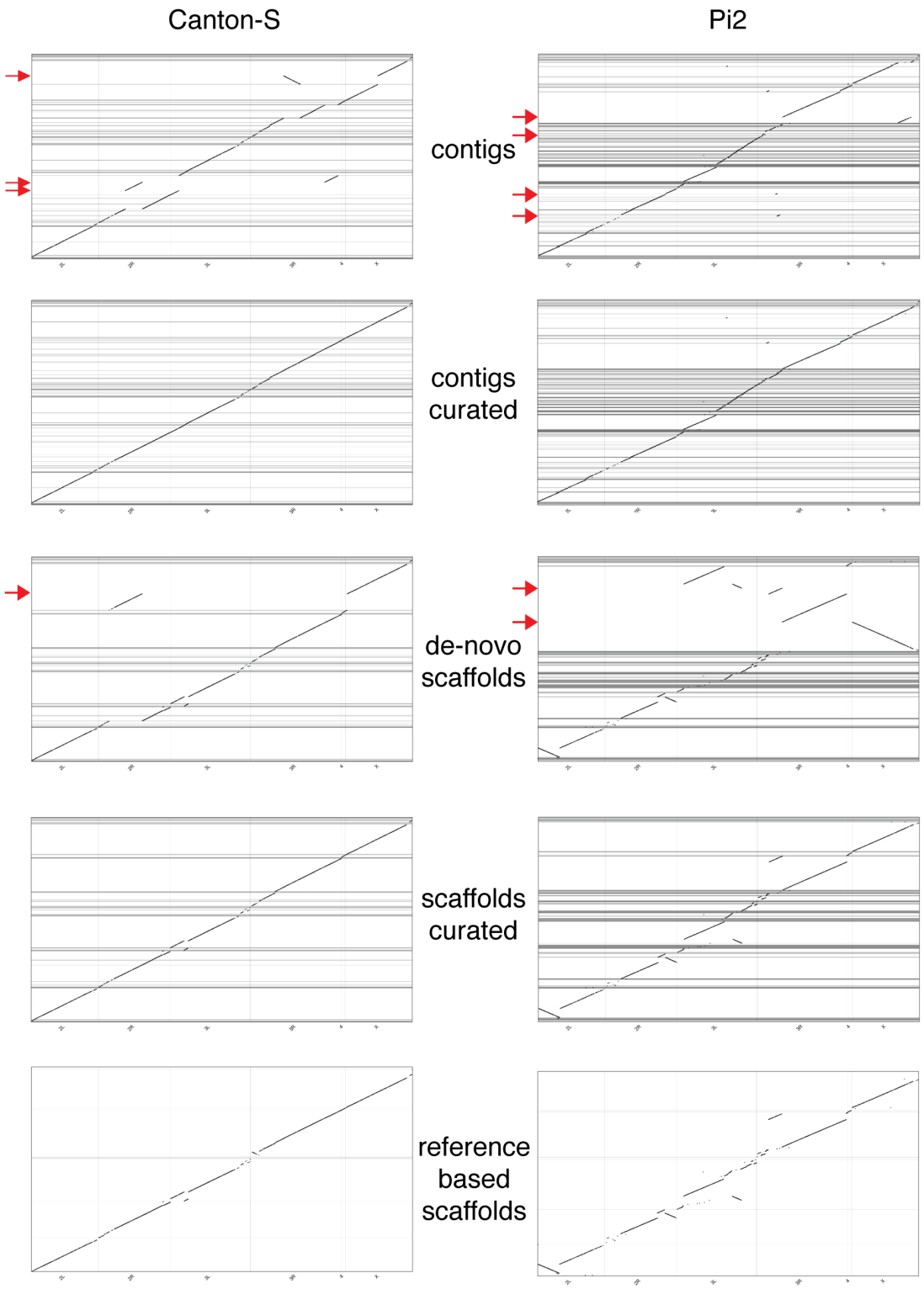

Figure S13: Manual curation steps of the final assemblies of Pi2 and Canton-S. Misassemblies (red arrows) were manually broken up at each step.

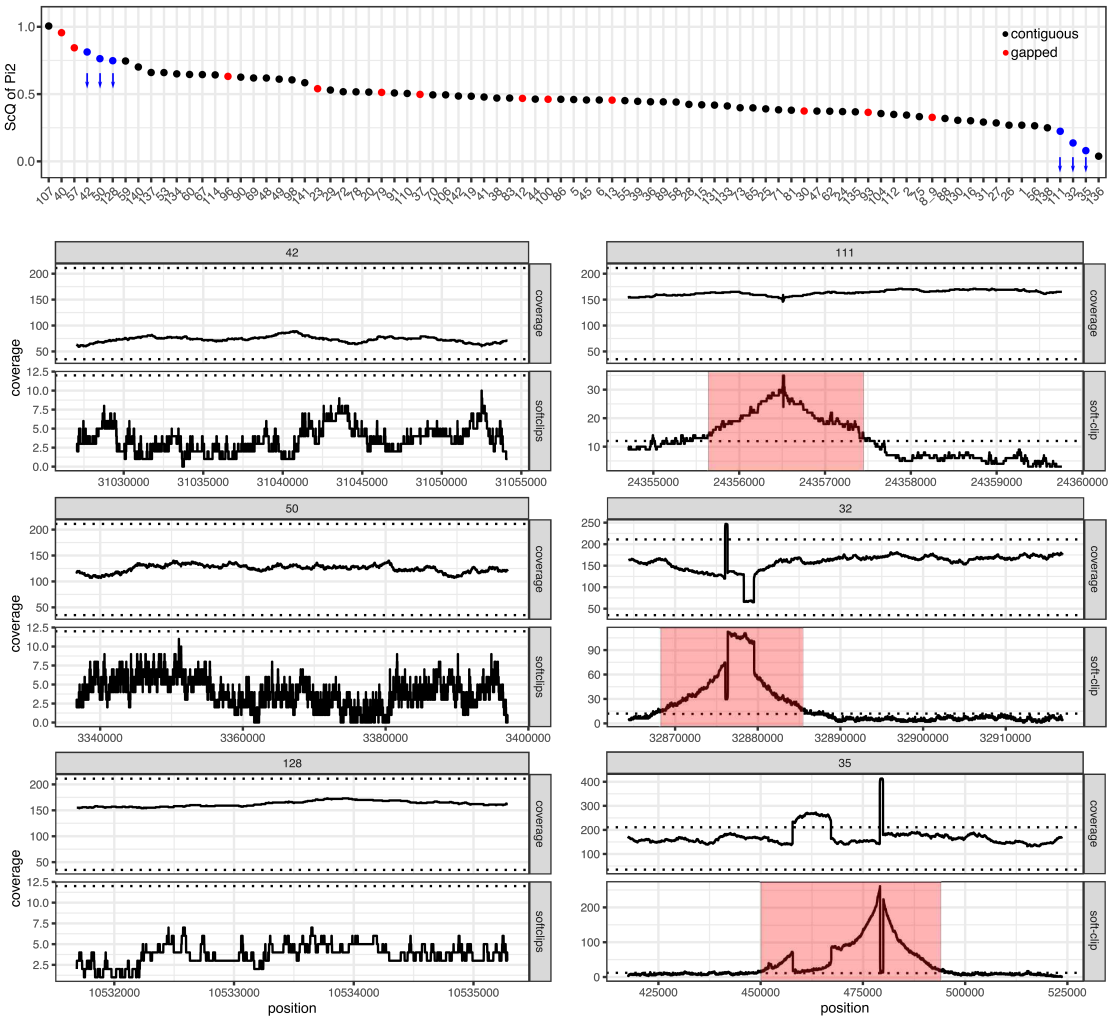

Figure S14: The base coverage and the soft-clip coverage are shown for three piRNA clusters having either high or low ScQ values in Pi2 (blue with arrows). Regions with potential assembly issues are marked red.

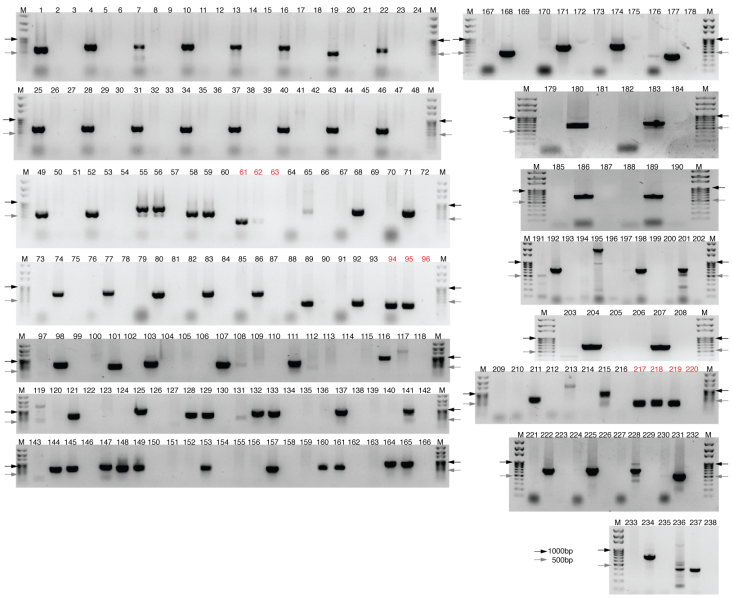

Figure S15: PCR validation of polymorphic TE insertions in piRNA clusters. Numbers above lanes refer to entries in supplementary tables S3; Arrows indicate the position of the 1000bp and 500bp size markers. Positive controls (*RpL32*) are labeled in red.

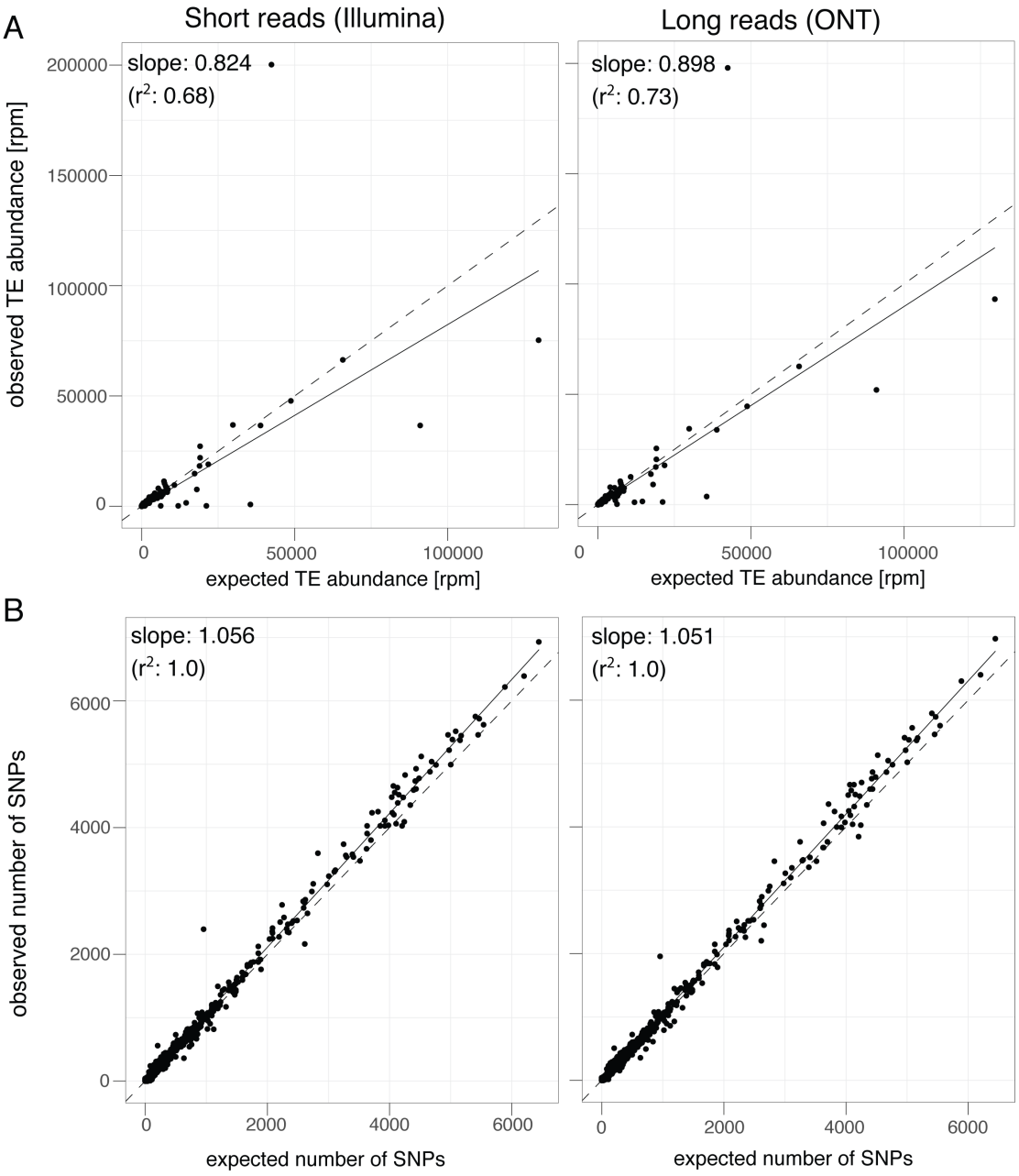

Figure S16: Expected and observed abundance of TEs (in reads per million [rpm]) and SNPs for a short-read and a long-read assembly of the Korean Reference Genome (KOREF1.0 and PT64x (Cho et al., 2016; Kim et al., 2019)). The slope of the regressions represent our novel quality metric for rpm and SNPs. Each dot represents a distinct TE family and the dashed line shows the optimal representation.

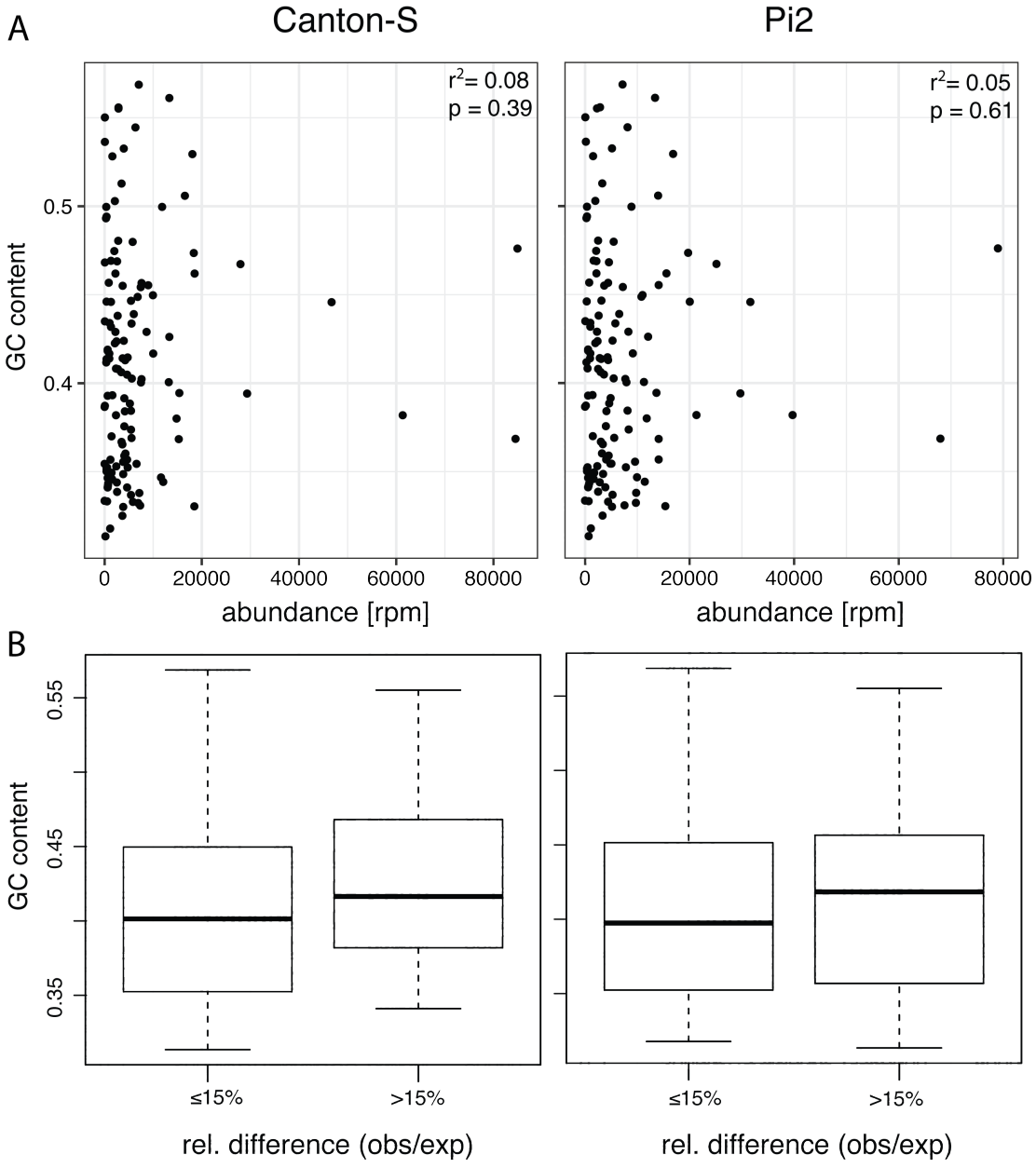

Figure S17: Influence of the GC-content on TE abundance for our assemblies of Canton-S and Pi2. A) Relationship between the the GC-content of a TE and the estimated abundance of a TE. Correlations were calculated with the Spearman method. B) GC-content of TEs that strongly deviate from the expected abundance (difference between observed and expected TE abundance $> 15\%$) compared to TEs that do not deviate from expectations (difference $< 15\%$). For both Canton-S and Pi2 the differences between deviating and not-deviating TEs was not significant (Wilcoxon rank-sum test, two-sided: $p_{CS} = 0.109$, $p_{Pi2} = 0.3813$)

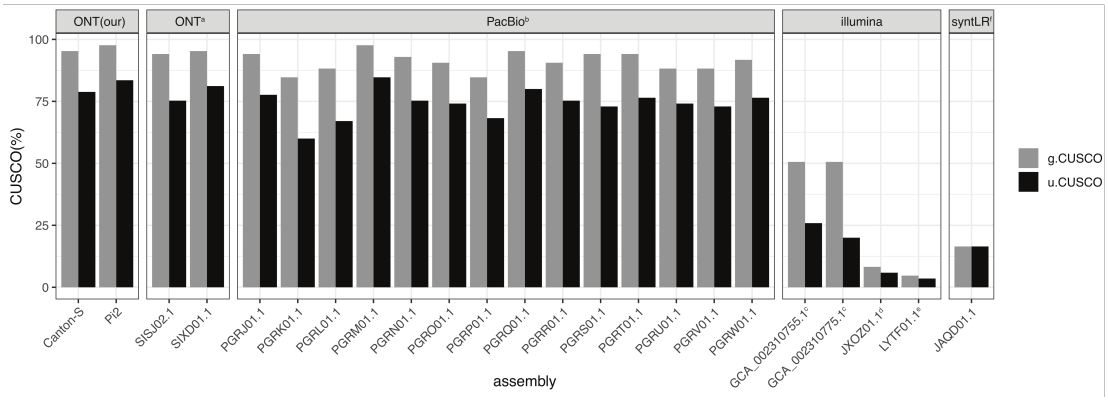

Figure S18: CUSCO values for our assemblies and publicly available assemblies of different *D. melanogaster* strains. We used assemblies from NCBI databases with following accession numbers: [a]WGS: SIXD01000000 and SISJ02000000 (Ellison and Cao, 2020) for ONT; [b]Bioproject: PRJNA418342 (Chakraborty et al., 2019) for PacBio; [c]Genbank: GCA_002310755.1 and GCA_002310775.1 (Anreiter et al., 2017), [d]WGS: JXOZ01000000 (Vicoso and Bachtrog, 2015),[e]WGS: LYTF01000000 (Singhal et al., 2017) for illumina; [f]WGS: JAQD01000000 (McCoy et al., 2014) for illumina synthetic long reads.

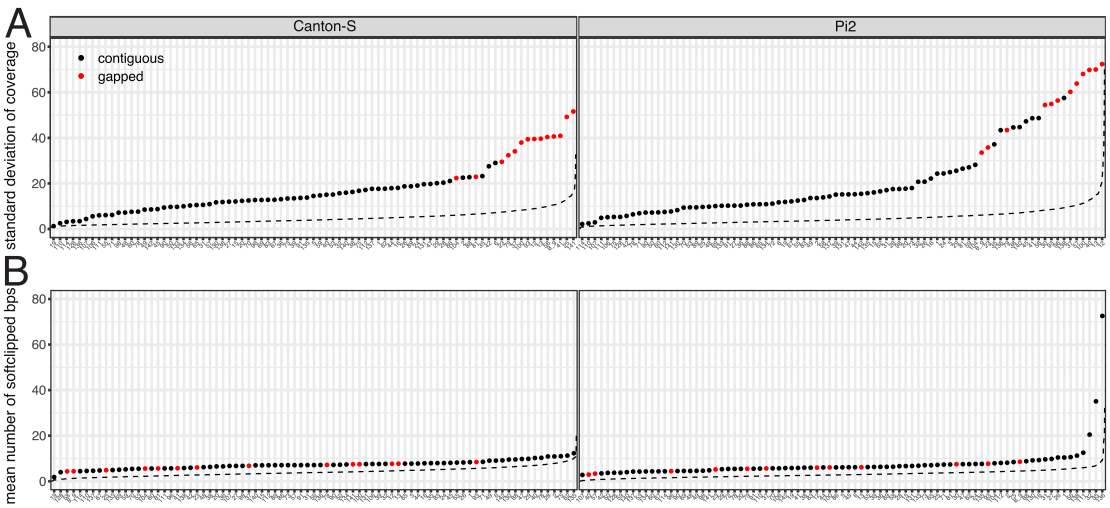

Figure S19: Raw data of piRNA cluster metrics based on the final assemblies of Canton-S and Pi2. Different piRNA clusters are shown on the x-axis. A) The standard deviation of the base coverage B) Average soft-clip coverage. Dashed lines show the corresponding distributions for completely assembly BUSCO genes (Diptera).

2 Supplementary tables

Table S1: Overview of the raw data used for the assemblies; PE paired ends

	CantonS	Pi2
ONT, coverage	149x	199x
ONT, flow cells	2	3
ONT, mean read length	7146bp	8045bp
Illumina PE, coverage	30x	40x
Illumina PE, read length	125	125
Hi-C, coverage	591x	260x

Table S2: Effect of polishing on BUSCO values. We applied three rounds of polishing with Racon, picked the assembly with the highest BUSCO value (bold) and polished this assembly three times with Pilon, where we again kept the assembly with the highest BUSCO values (bold). In case BUSCO values did not improve between two successive iterations we kept the assembly requiring the fewest polishing steps. The finally used polishing strategy (polishing s.) for each assembly is shown at the bottom (R .. Racon, P.. Pilon). All ONT reads (150x coverage) were used for these assemblies.

	CantonS			
	Canu	miniasm	wtdbg2	Flye
unpolished	83.0	1.1	74.9	90.3
1x Racon	91.0	80.2	90.5	**91.5**
2x Racon	91.0	90.2	91.4	90.9
3x Racon	**91.6**	**91.5**	**91.7**	91.5
1x Pilon	98.6	98.2	98.5	98.5
2x Pilon	98.8	98.6	**98.9**	**98.7**
3x Pilon	**98.9**	**98.8**	98.8	98.7
polishing s.	3R,3P	3R,3P	3R,2P	1R,2P

Table S3: Overview of validated presence/absence polymorphisms in piRNA clusters. For each polymorphism, we show the primers, the cluster, the TE family, the length of the expected fragment, the position and indicate whether or not the polymorphism is present in Canton-S, Pi2 or Iso-1 (y..yes, n..no). Numbers in brackets refer to lane numbers in the PCR gels (supplementary fig. S15).

PCR-ID	strain of prescence	Canton-S	Pi2	Iso-1	piRNA cluster	TE family	start	end	length	primer #1	primer #2
cs19L	Canton-S	y(13)	n(14)	-	cl112	doc	14527	15230	704	GCAGAGAGGGAGAGCAAGAA	TTTGACTGGCCTTCTTACGG
cs19R	Canton-S	y(16)	n(17)	-	cl112	doc	9735	10435	701	TATTCAACTGGCCCTTCCTG	AATCCCCTCGCAAGAAAACT
cs16L	Canton-S	y(7)	n(8)	-	cl133	mdg1	15190	15885	696	TCACGGTCGCCATGTAGTTA	TCCTCGGTTGGTCCTAATTC
cs16R	Canton-S	y(10)	n(11)	-	cl133	mdg1	7521	8235	715	GCTGAAAGAACCCACCGATA	GCACTTGGCTGTCACAAGAG
cs14R	Canton-S	y(4)	n(5)	-	cl140	doc	58054	58752	699	TGCGCGTAATATATGTCGATG	GTGCTCGATCACCGATTTG
cs14L	Canton-S	y(1)	n(2)	-	cl140	doc	62711	63311	601	ACTTATGGCTCCGAGCTGTG	CCCGAAAACCGTTAAATCAG
cs8R	Canton-S	y(52)	n(53)	-	cl15	412	52318	52917	600	AGCGTGATTTGGTGATAGCC	GGTCGCCATGTAATGATGAA
cs8L	Canton-S	y(49)	n(50)	-	cl15	412	59624	60219	596	CCCACTCGAAGGCAAAAGTA	GAGCCCATTTATGCCAAGTT
cs7R	Canton-S	y(46)	n(47)	-	cl15	F-element	74840	75448	609	GCTACTGCCTGATCCGATGT	AGCGTCTCTTTCGCTTTCAG
cs7L	Canton-S	y(43)	n(44)	-	cl15	F-element	79627	80231	605	TTCAGCGAGCACAATCAAAG	TCAAAACGCAATCGGTCATA
cs9R	Canton-S	y(58)	y(59)	-	cl15	springer	27385	27982	598	AAGCTGCTTGTGCAAGTTGA	ACTTACGCCCCTGTATGTCG
cs9L	Canton-S	y(55)	y(56)	-	cl15	springer	30676	31482	807	GGTCAATTTTCGCACCTACC	CCTAGCAATGGCTCAGGAAA
cs6R	Canton-S	y(40)	n(41)	-	cl16	297	40121	40727	607	TACTCACGCACTTTCCATCG	TCAAAACACACCACAACAAACA
cs6L	Canton-S	y(37)	n(38)	-	cl16	297	47115	47710	596	GCAGCTGGGATACGTTATGG	CCAAGGCGGTCTGTATTGTT
cs4R	Canton-S	y(34)	n(35)	-	cl20	roo	119943	120547	605	GATGGCGTCCAAGTTTGTTT	CCTTTGGTAGGGGGAAACTG
cs4L	Canton-S	y(31)	n(32)	-	cl20	roo	129025	129624	600	CGATAAGGCGGGGACTATTT	CCTTCGGTCAAACACCTTGT
cs3R	Canton-S	y(28)	n(29)	-	cl26	juan	78770	79374	605	CAAATAAAGCGGCGAACTGT	GGTCACCTGTTTGGGGTAGA
cs3L	Canton-S	y(25)	n(26)	-	cl26	juan	82718	83336	619	AGCTGCATTGAAACAGTCCA	GTTTCAGCGGATCCCAAATA
cs1L	Canton-S	y(19)	n(20)	-	cl6	quasimodo	49383	49885	503	GCCCTTCGCTTTAAGCTTTC	AGCGTTCGCTTTGCAAATTA
cs1R	Canton-S	y(22)	n(23)	-	cl6	quasimodo	56976	57581	606	TTGTCACCAGAATTTGGGTTC	GGCCGTTTTTAACGGAAAG
pi22L	Pi2	n(179)	y(180)	-	cl1	invader1	18945	19546	602	TCTTTCCGTCCATCCGTATC	CGGCAGACAACCGTTAGAGT
pi22R	Pi2	n(182)	y(183)	-	cl1	invader1	22817	23509	693	CATCAGAATGCCCTTCTTCC	CTTATTGAGTCGCGAGAAACG
pi25L	Pi2	n(97)	y(98)	-	cl1	P-element	267796	-	609	GGGATTCTGCGATTTGATTC	CTTCGGTAAGCTTCGGC
pi25R	Pi2	n(100)	y(101)	-	cl1	P-element	-	268773	474	GTGGATGTCTCTTGCCG	TTGGTTTAAGCGATGCTTTG
pi23	Pi2	n(191)	y(192)	-	cl1	rover	30040	30743	704	CAAATCCAACTGCATCACCA	GTTCTCAGTTCGGGTTCGAT
pi24	Pi2	n(194)	y(195)	-	cl1	duplication	53683	55687	2005	GCATTAGCTCAGGCAGGAAA	TTGCTGTTGCCATTGTTGTT
pi17R	Pi2	n(224)	y(225)	-	cl130	copia	18413	19115	703	AGGTCTGCTCGGTGACATTC	CCACGACCTACTCACAGCAA
pi17L	Pi2	n(221)	y(222)	-	cl130	copia	23883	24580	698	AATGGCCACACCTTTTATGC	TTCAAACATTATCCCCTGCAC
pi18L	Pi2	n(227)	y(228)	-	cl130	idefix	13300	14003	704	ATCCAAGAGAAACGCAAAGC	ATATGTCGGCGAAACAGGAG
pi18R	Pi2	n(230)	y(231)	-	cl130	idefix	5858	6371	514	CCGGATGTCAAGAGGAAGAG	GGTGAGGCTTAAGCAAGGAA
pi15R	Pi2	n(91)	y(92)	-	cl137	297	11479	11880	402	GCAAACACACGAAATCGAAG	TCCAGCGCTCAAGTAATACG
pi15L	Pi2	n(88)	y(89)	-	cl137	297	2696	3085	390	CTGTCACGGTCGCCATATAA	TAGCGTTGAAAGAGGGCAGT
pi12R	Pi2	n(79)	y(80)	-	cl140	F-element	28851	29541	691	CACACAAACGGTCCATATCAA	CCGTTTATTGCTAGCGGTTT
pi12L	Pi2	n(76)	y(77)	-	cl140	F-element	33426	34123	698	CAACGCAATCACGGAACTTA	TTGACCCTTTGGTGGGAATA
pi13R	Pi2	n(85)	y(86)	-	cl140	F-element	3919	4620	702	TTCGGATTGACATTGGTTGA	TATTGACCGCTCGATGTCTG
pi13L	Pi2	n(82)	y(83)	-	cl140	F-element	8614	9318	705	TTCAGCGAGCACAATCAAAG	AACAAACGTGTGTGGCGTTA
pi11R	Pi2	n(73)	y(74)	-	cl140	gypsy	69269	69959	691	CGAAACAATCGCCATCCTAT	GGCTCATTGCCGTTAAACAT
pi11L	Pi2	n(70)	y(71)	-	cl140	gypsy	76745	77449	705	GGCGATAGCGATTTGATTGT	AACGCTCCACGTTTCATAGG
pi10R	Pi2	n(67)	y(68)	-	cl15	F-element	77867	78569	703	CTTCTCCGTGCTATCCTTCG	CTCCGCTCAGCCTAGATCAC
pi10L	Pi2	n(65)	y(65)	-	cl15	F-element	80809	81505	697	TCTCCAGCTGTTGTTGTTGG	TATGGATGCCTGAAGGGCTA
pi28L	Pi2	n(197)	y(198)	-	cl16	mdg1	21932	22610	679	CGACTCCCACAACTTCATCA	AGCACTTGGCTGTCACAAGA
pi28R	Pi2	n(200)	y(201)	-	cl16	mdg1	29643	30335	693	CTCTTGTGACAGCCAAGTGC	GAGAATCGAAACGGAAATCG
pi5R	Pi2	n(188)	y(189)	-	cl19	gypsy5	12115	12808	694	AAGTAGTGGCGTGGACTGCT	TATACCGGGGAAGTGACTGG
pi5L	Pi2	n(185)	y(186)	-	cl19	gypsy5	4604	5309	706	TGTTCTGACCCACTCCACTG	GGCCGTCACTTTATTCAAGG
pi2R	Pi2	n(206)	y(207)	-	cl26	mdg3	75530	76216	687	GAGGCATCCGTTTGGTAAAA	AGGACGGCCGAGTTGTAGTA
pi2L	Pi2	n(203)	y(204)	-	cl26	mdg3	81089	81790	702	GGTTGTCAGCGAAAAGTCGT	CACAATGTATCGCCCATCTG
pi21R	Pi2	y(176)	y(177)	-	cl5	copia	12622	13018	397	ATTTTCCCTTGCACGAAATG	ACCAGCACCACGACCTACTC
pi21L	Pi2	n(173)	y(174)	-	cl5	copia	17489	18199	711	CCAATCGAATGCTGAATGAA	AGGATCATCTGCGACTCAGG
pi20R	Pi2	n(170)	y(171)	-	cl5	gtwin	29656	30354	699	CACCAACGATAGCTGATCCA	GCTCGCAGCCGTAAAATATC
pi20L	Pi2	n(167)	y(168)	-	cl5	gtwin	51179	51671	493	CAATGCATCACGCCATAAAG	AGCAATCATGATGTCGTCCA
cs29R	CS	y(237)	y(236)	-	cl1	copia	267802	268194	393	GTTTATTAGGCATGGACTGG	CAGCAGCAAGAATACTCC
cs29L	CS	y(234)	n(233)	-	cl1	copia	272690	273429	740	CGCTTGTAGTCTATCCCTAAC	CCACCCACATTTGATAGTTAC
iso34L	Iso-1	n(135)	n(136)	y(137)	cl1	1731	91808	92595	788	GAATCTGTTACGGCCCATTC	CATGAAAGAGGGTCACGTTG
iso34R	Iso-1	n(139)	n(140)	y(141)	cl1	1731	96320	97224	905	TCATTCCGGCAATCTTGAGT	TGTCCGACATTGTGGGTAAT
iso32L	Iso-1	n(119)	n(120)	y(121)	cl1	F-element	57380	57965	586	AATGGGAATTTGTGCTTTCG	TCCTGGGCTATGGGTTATTG
iso32R	Iso-1	n(123)	n(124)	y(125)	cl1	F-element	61836	62648	813	CTCCGCTCAGCCTAGATCAC	CTTTGGAGGCAAATTTCCAA
iso35L	Iso-1/Pi2	n(143)	y(144)	y(145)	cl1	F-element	131490	132095	606	CGAACTCCATCCCATTCCTA	CCTACCAACCCAGCGAATAA
iso35R	Iso-1/Pi2	y(147)	y(148)	y(149)	cl1	F-element	136291	136888	598	TCCTGGGCTATGGGTTATTG	ATTTGGTCTGGTGACGCTCT
iso38L	Iso-1	n(209)	n(210)	y(211)	cl1	F-element	239670	240262	593	TTGTCCAGAGTGACCCTTCC	CACTGGCTCATCAACTCTGG
iso38R	Iso-1	n(213)	n(214)	y(215)	cl1	F-element	244111	245067	957	TCCTGGGCTATGGGTTATTG	TGTGCTCACCATAAAATGTGG
iso37L	Iso-1/Pi2	n(159)	y(160)	y(161)	cl1	invader3	215863	216462	600	GCTACGCCTGTGGACAACTT	TGCTGCTGATCCGTTATTTG
iso37R	Iso-1/Pi2	n(163)	y(164)	y(165)	cl1	invader3	221147	221746	600	GCCGAGTTGTGGTAGGATGA	CCCAAGGTGTACGGGTCTAA
iso36L	Iso-1	n(151)	n(152)	y(153)	cl1	juan	156015	156659	645	TTCACAATGCATGCCGTAAT	CCTGTTTGGGGTAGATGTGC
iso36R	Iso-1	n(155)	n(156)	y(157)	cl1	juan	160009	160613	605	AGCTGCATTGAAACAGTCCA	CACGACAGAGAGGGATGACA
iso33L	Iso-1/Pi2	n(127)	y(128)	y(129)	cl1	stalker4	75075	75679	605	CCTCGTAGACATCGGACTGC	AAGTGGTCAATGCCTTCTGC
iso33R	Iso-1/Pi2	n(131)	y(132)	y(133)	cl1	stalker4	82606	83307	702	GCAGAAGGCATTGACCACTT	GACCACCTGACGATCAAACC
cs30R	Canton-S	y(107)	n(108)	n(109)	cl1	doc	148642	149240	599	ATACGAATGAGACCCGCAGT	CGTGGTACCTTCGAAACGA
cs30L	Canton-S	y(103)	n(104)	n(105)	cl1	doc	153030	153629	600	CGGTCAGTGCTGCTGAATTA	ATTACGGCATGCATTGTGAA
cs31R	Canton-S	n(115)	y(116)	n(117)	cl1	roo	137094	137698	605	ATACGAATGAGACCCGCAGT	TTGGGCTCCGTTCATATCTT
cs31L	Canton-S	y(111)	n(112)	n(113)	cl1	roo	145714	146422	709	GGGCACATCTGCCTATCTTG	AACGGGAATGTGTGGCTATT

Table S4: Influence of different polishing steps on the quality of Canton-S assemblies (30x coverage with long reads). Assembly quality is estimated with BUSCO and our four TE centered quality metrics (CUSCO, TE abundance, SNPs and IDs in TEs).

quality	raw	racon	pilon
BUSCO	76.8	85.6	98.4
CUSCO	60.0	60.0	60.0
TE abu.	0.97	0.97	0.97
TE SNPs	0.93	0.99	1.00
TE IDs	0.91	0.97	0.99

Table S5: Overview of the final assemblies of Canton-S and Pi2. The assembly quality is assessed with classic quality metrics (NG50, BUSCO) as well as our TE centered quality metrics. Misassembled contigs and scaffolds were broken manually (based on dot-plots; supplementary fig. S13). u.CUSCO ungapped-CUSCO, g.CUSCO gapped-CUSCO

		CantonS	Pi2
ONT	Assembler	Canu	Canu
	Coverage	100x	100x
	contigs	335	625
	NG50	6.8m	4.1m
	length	148m	169m
	BUSCO	82.3	85.8
	u.CUSCO	80.00	77.65
	TE abu.	1.02	1.04
	TE SNPs	0.95	0.97
	TE IDs	0.93	0.95
pol.	strategy	2R,2P	2R,2P
	contigs	335	625
	NG50	4.6m	4.1m
	length	149m	169m
	BUSCO	98.7	98.3
	u.CUSCO	81.18	83.53
	TE abu.	1.01	1.04
	TE SNPs	0.99	1.01
	TE IDs	0.99	1.01
Hi-C	scaffolds	266	483
	NG50	21.4m	23.6m
	length	149m	169m
	BUSCO	98.7	98.3
	g.CUSCO	84.71	91.76
	TE abu.	1.01	1.04
	TE SNPs	0.99	1.01
	TE IDs	0.99	1.01
ref. scaf.	scaffolds	15	16
	NG50	28.2m	37.2m
	length	149m	168m
	BUSCO	98.6	98.2
	g.CUSCO	95.29	97.65
	TE abu.	1.01	1.04
	TE SNPs	0.99	1.01
	TE IDs	0.98	1.00
	CQ	0.092	0.065
	ScQ	0.567	0.427

Table S6: BUSCO values for assemblies generated with different assemblers and coverages for Canton-S. For each assembly the optimized number of polishing rounds performed with Racon (R) and Pilon (P) are shown (for optimization procedure see supplementary table S2).

	coverage	20x	30x	50x	75x	100x	120x	150x
Canu	polishing	3R,3P	3R,3P	3R,3P	2R,3P	3R,2P	3R,3P	3R,3P
	BUSCO	98.4	98.4	98.6	98.6	98.6	98.8	98.9
miniasm	polishing	3R,3P	3R,2P	3R,3P	3R,2P	3R,3P	3R,3P	3R,3P
	BUSCO	98.4	98.4	98.6	98.6	98.6	98.8	98.8
wtdbg2	polishing	1R,3P	2R,3P	2R,3P	3R,3P	3R,2P	2R,3P	3R,2P
	BUSCO	98.8	98.6	98.7	98.7	98.7	98.6	98.9
Flye	polishing	1R,3P	2R,3P	1R,3P	1R,3P	2R,2P	2R,3P	1R,2P
	BUSCO	98.6	98.9	98.7	98.6	98.7	98.8	98.7

References

Anreiter, I., Kramer, J. M., and Sokolowski, M. B. (2017). Epigenetic mechanisms modulate differences in *Drosophila* foraging behavior. *Proceedings of the National Academy of Sciences*, 114(47):12518–12523.

Chakraborty, M., Emerson, J. J., Macdonald, S. J., and Long, A. D. (2019). Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nature Communications*, 10(1):4872.

Cho, Y. S., Kim, H., Kim, H.-M., Jho, S., Jun, J., Lee, Y. J., Chae, K. S., Kim, C. G., Kim, S., Eriksson, A., et al. (2016). An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. *Nature communications*, 7(1):1–13.

Ellison, C. E. and Cao, W. (2020). Nanopore sequencing and Hi-C scaffolding provide insight into the evolutionary dynamics of transposable elements and piRNA production in wild strains of *Drosophila melanogaster*. *Nucleic Acids Research*, 48(1):1–14.

Kim, H.-S., Jeon, S., Kim, C., Kim, Y. K., Cho, Y. S., Kim, J., Blazyte, A., Manica, A., Lee, S., and Bhak, J. (2019). Chromosome-scale assembly comparison of the Korean Reference Genome KOREF from PromethION and PacBio with Hi-C mapping information. *GigaScience*, 8(12):giz125.

McCoy, R. C., Taylor, R. W., Blauwkamp, T. A., Kelley, J. L., Kertesz, M., Pushkarev, D., Petrov, D. A., and Fiston-Lavier, A.-S. (2014). Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PLoS ONE*, 9(9):e106689.

Singhal, K., Khanna, R., and Mohanty, S. (2017). Is *Drosophila*-microbe association species-specific or region specific? A study undertaken involving six Indian *Drosophila* species. *World Journal of Microbiology and Biotechnology*, 33(6):103.

Vicoso, B. and Bachtrog, D. (2015). Numerous Transitions of Sex Chromosomes in Diptera. *PLOS Biology*, 13(4):e1002078.

Weilguny, L. and Kofler, R. (2019). DeviaTE: Assembly-free analysis and visualization of mobile genetic element composition. *Molecular Ecology Resources*, 19(5):1346–1354.

Chapter 2

Supplement to Tirant stealthily invaded natural *Drosophila melanogaster* populations during the last century

Florian Schwarz, Filip Wierzbicki, Kirsten-André Senti and Robert Kofler

October 29, 2020

Supplementary figures

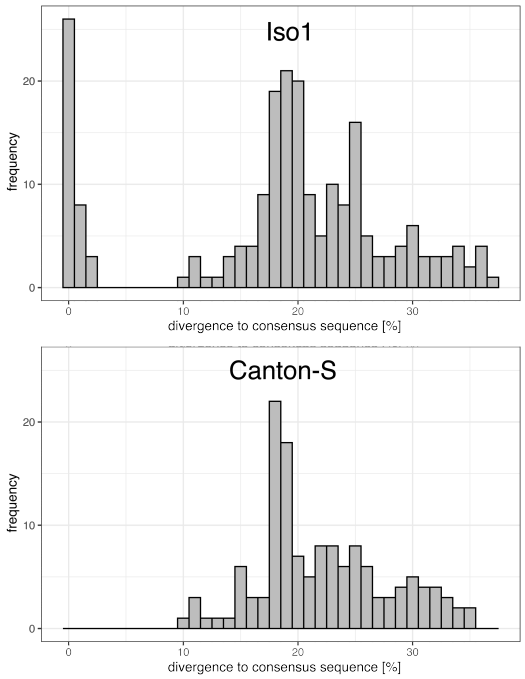

Figure 1: Histogram of the divergence of Tirant sequences relative to the canonical sequence. Tirant sequences were annotated with RepeatMasker in the assemblies of the *D. melanogaster* strains Iso-1 and Canton-S. Note that Iso-1 contains canonical (divergence < 5%) as well as degraded (divergence > 10%) Tirant insertions, whereas Canton-S solely contains degraded Tirant insertions.

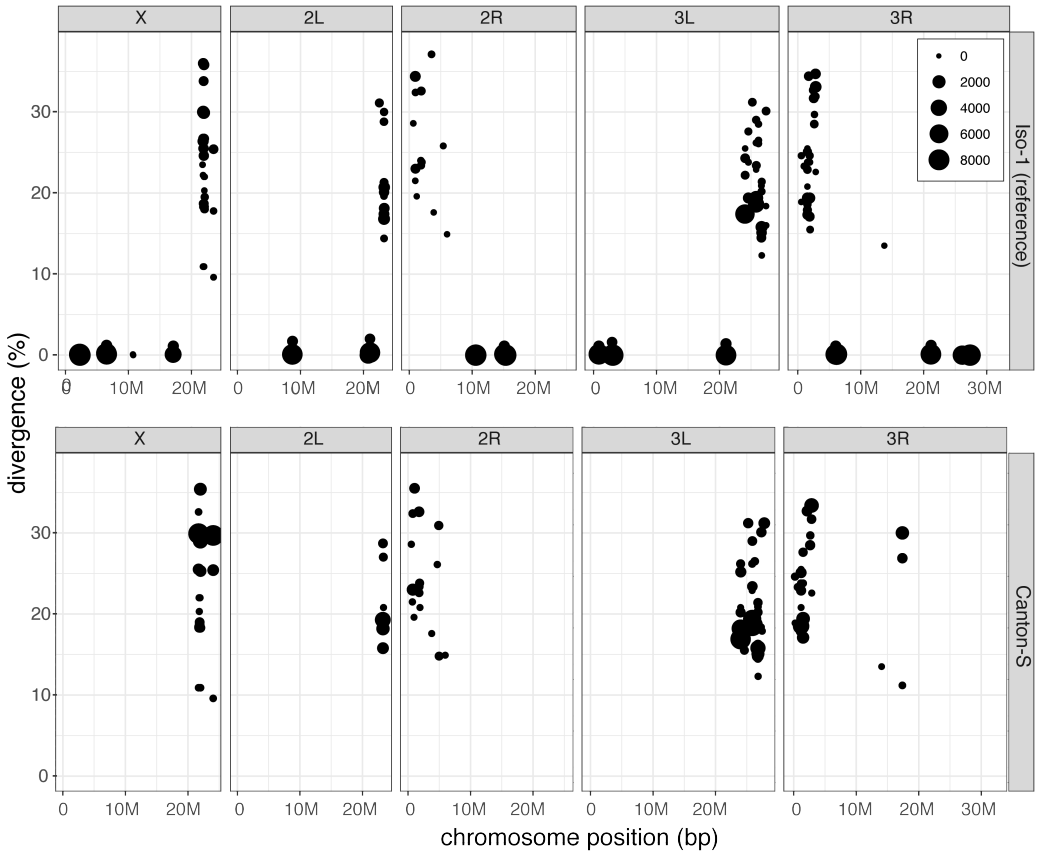

Figure 2: Canonical Tirant insertions are present in Iso-1 but not in Canton-S. The Canton-S assembly was generated by Chakraborty et al. (2019) with PacBio reads (the Canton-S assembly shown in the main manuscript was generated by Wierzbicki et al. (2020) with ONT reads). For each Tirant insertion we show the position in the assembly, the length (size of dot), and the similarity to the consensus sequence (divergence).

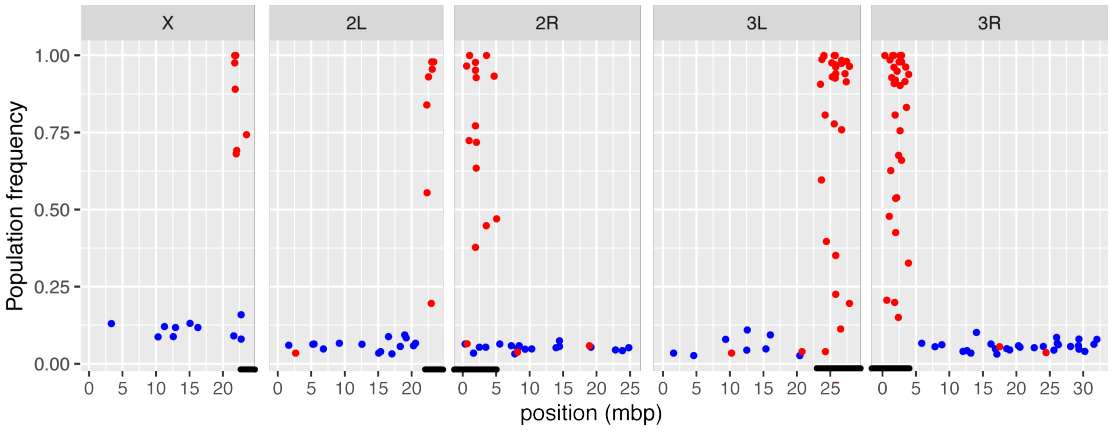

Figure 3: Position and population frequency of canonical (blue) and degraded (red) Tirant insertions in a population from France (Viltain) (Kapun et al., 2020). Canonical Tirant insertions are mostly euchromatic and segregating at a low population frequency whereas degraded insertions are mostly heterochromatic and segregating at high frequency. Black bars indicate (peri)centric heterochromatin (Riddle et al., 2011; Hoskins et al., 2015).

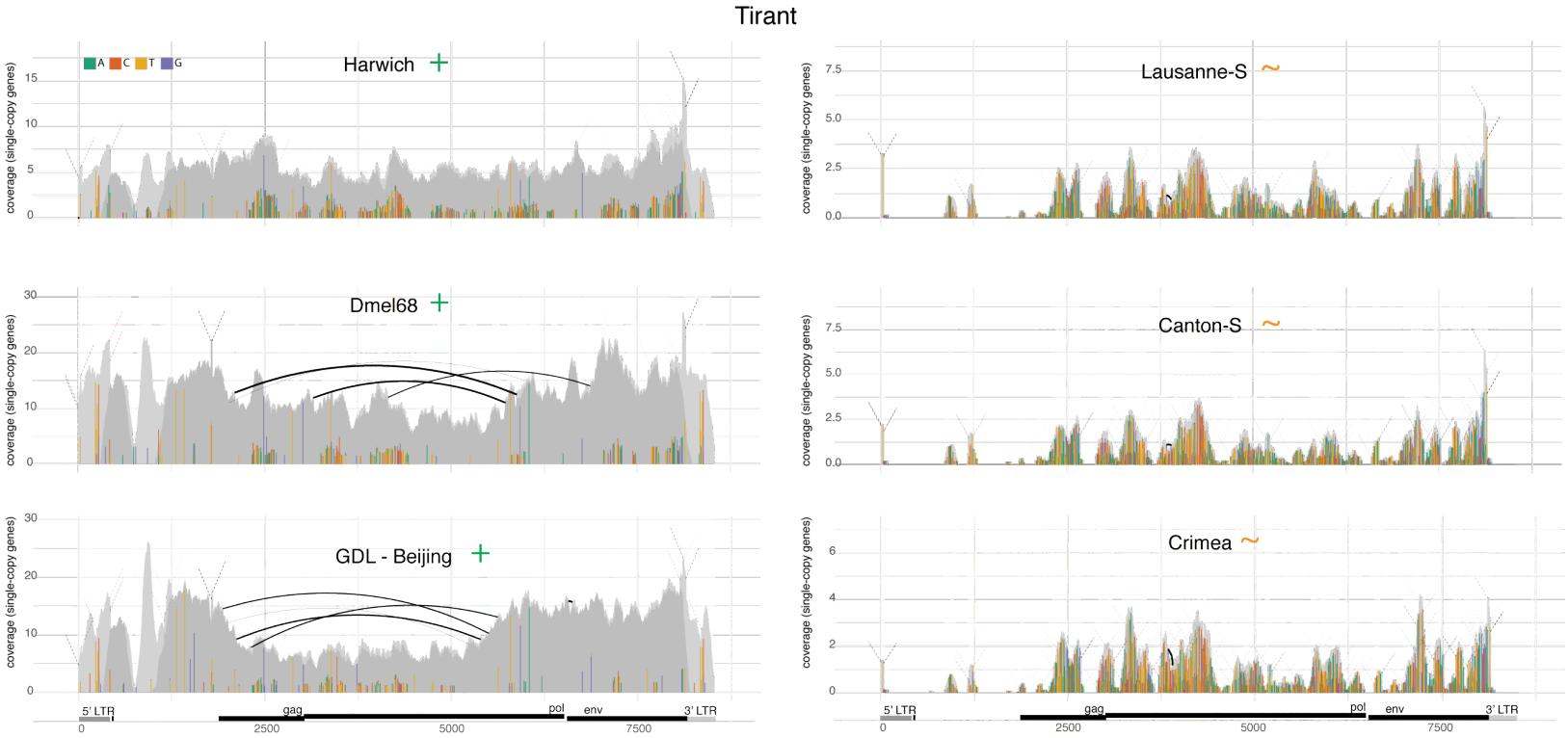

Figure 4: DeviaTE plots for three strains having non-degraded (i.e. canonical) Tirant sequences (+) and three strains solely having degraded Tirant sequences (~). Such plots were used for classifying the Tirant content of the different *D. melanogaster* strains (see supplementary table 1).

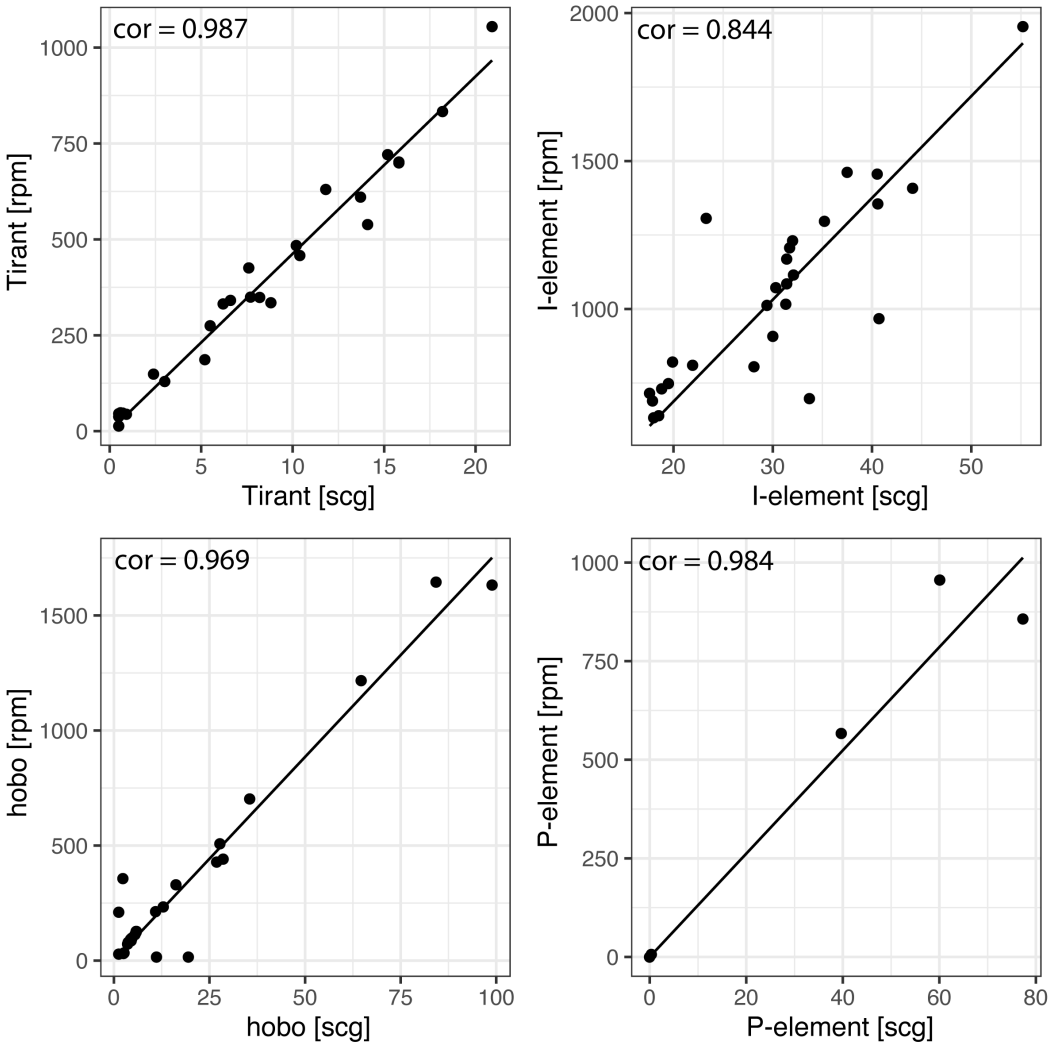

Figure 5: Correlation between the TE abundance estimated by DeviaTE using single copy gene normalization (scg) and the raw abundance of reads mapping to each TE normalized to a million reads (rpm: reads per million). Data are reported for Tirant, the I-element, hobo and the P-element. cor, Pearson's correlation coefficient

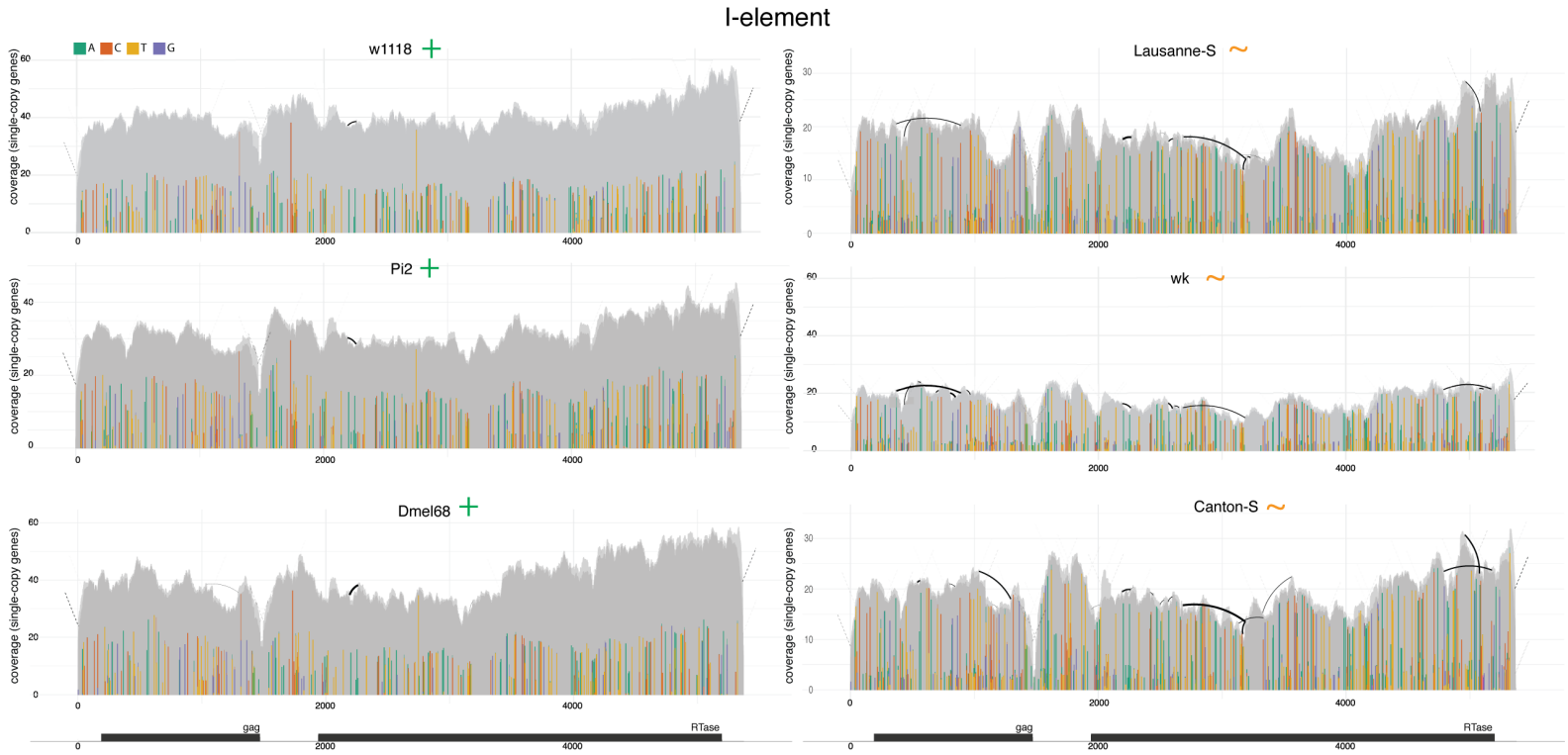

Figure 6: DeviaTE plots for three strains having non-degraded I-element sequences (+) and three strains solely having degraded I-element sequences (~). Such plots were used for classifying the I-element content of the different *D. melanogaster* strains (see supplementary table 1).

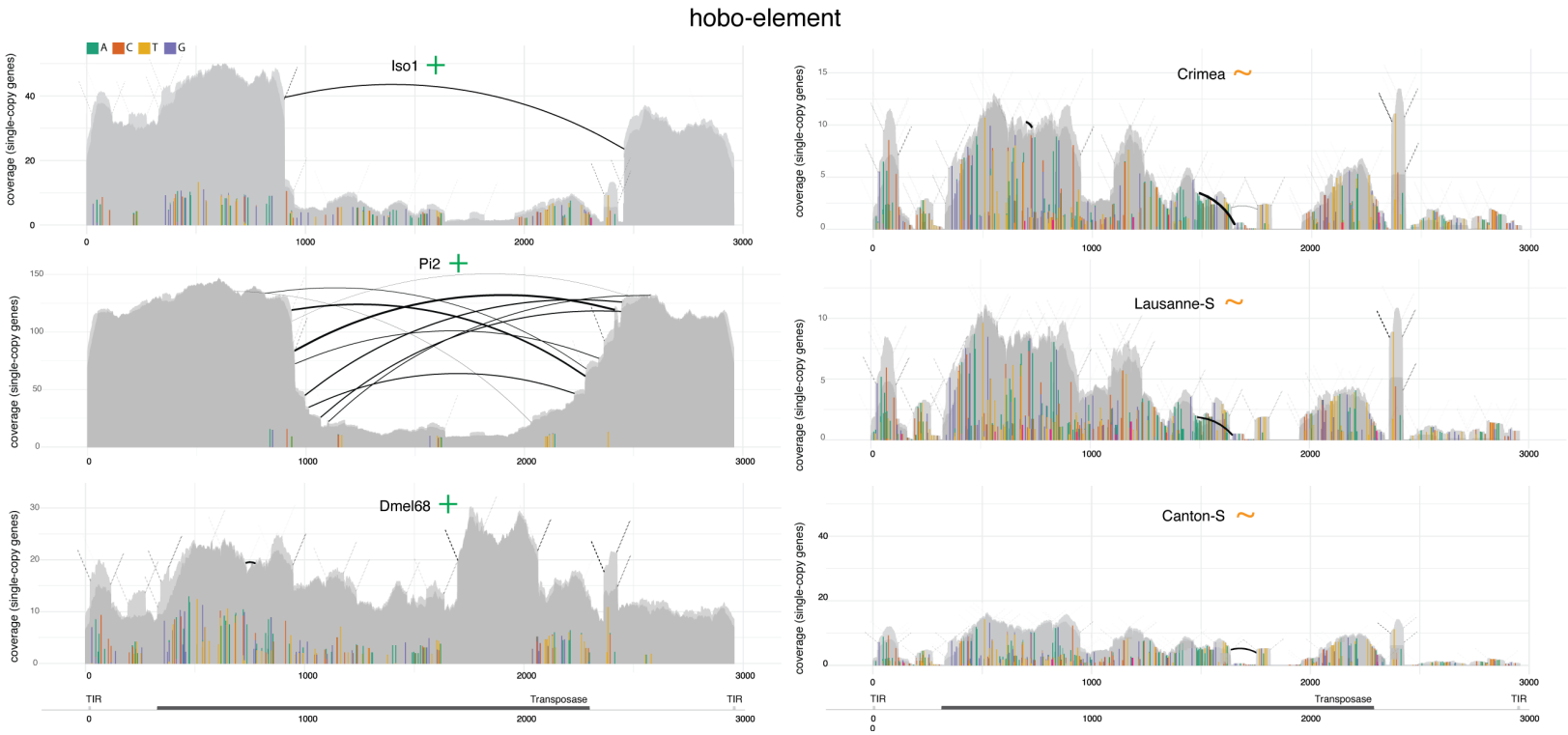

Figure 7: DeviaTE plots for three strains having non-degraded hobo sequences (+) and three strains solely having degraded hobo sequences (~). Such plots were used for classifying the hobo content of the different *D. melanogaster* strains (see supplementary table 1).

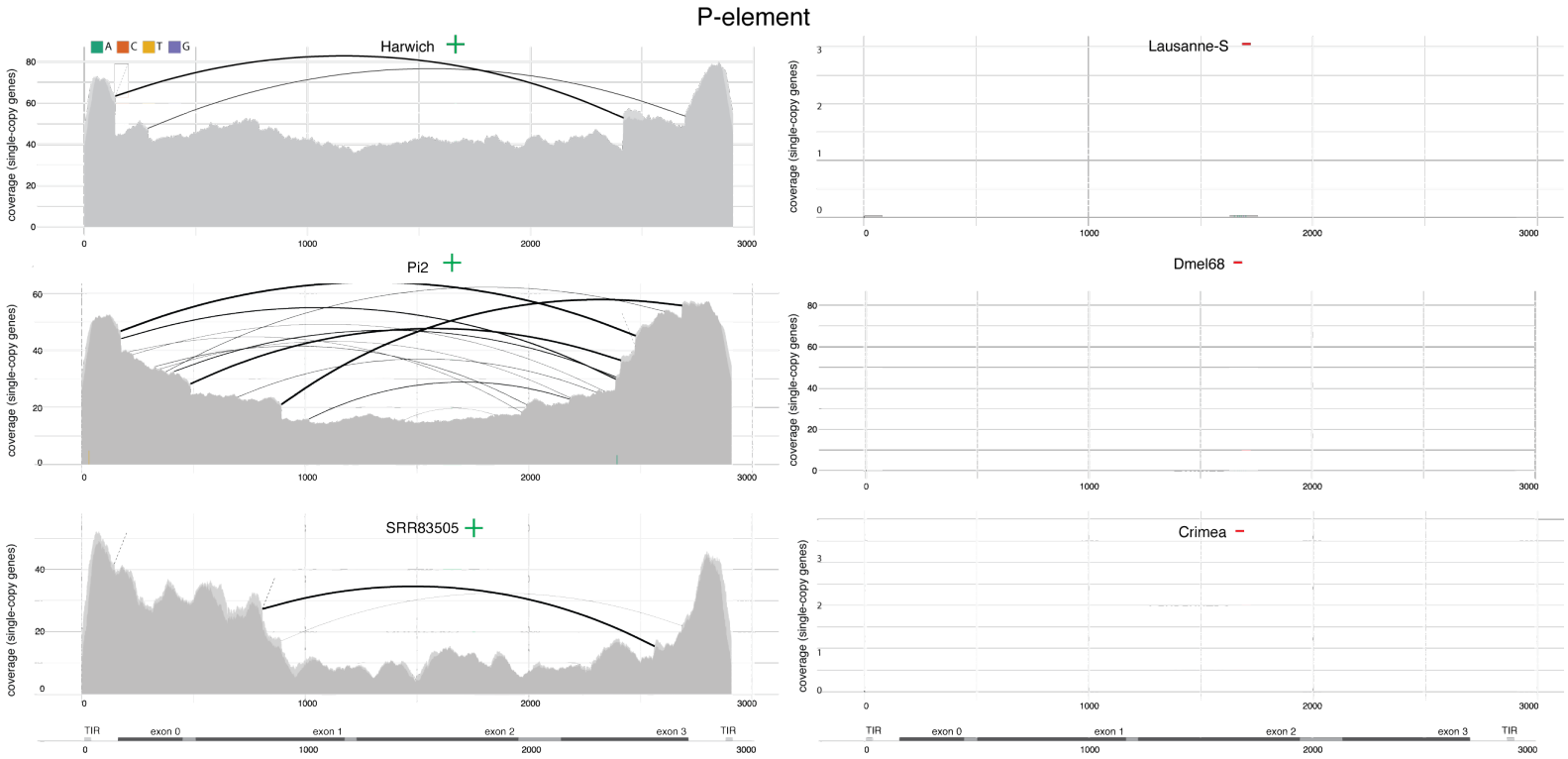

Figure 8: DeviaTE plots for three strains having P-element sequences (+) and three strains not having P-element sequences (−). Such plots were used for classifying the P-element content of the different *D. melanogaster* strains (see supplementary table 1).

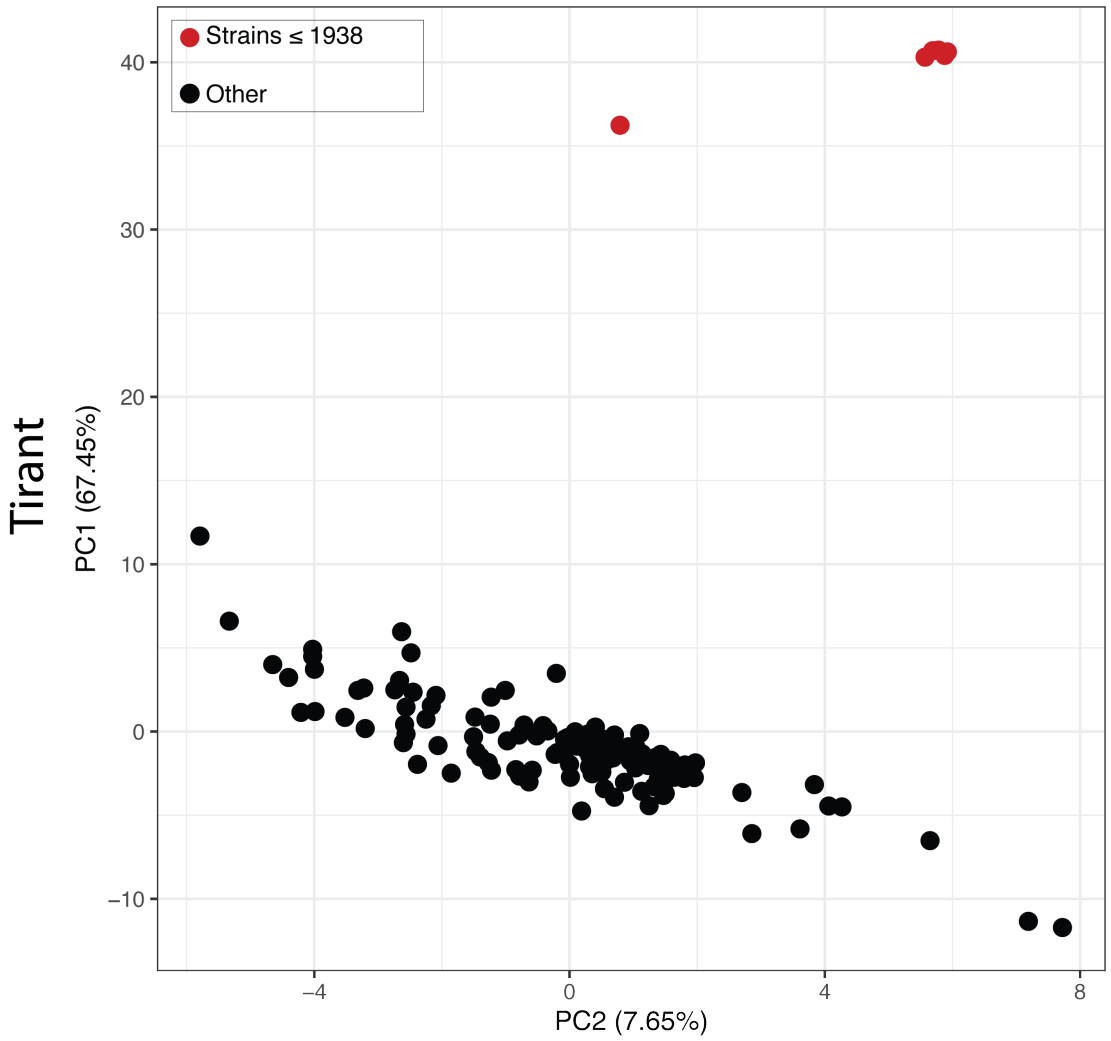

Figure 9: PCA based on the allele frequencies of SNPs in Tirant for different *D. melanogaster* strains and population samples. Strains sampled before or at 1938 form a separate cluster (due to the absence of canonical Tirant insertions). In addition to the strains shown in the manuscript (fig. 3), we used DGRP, DrosEU and Dros-RTEC lines well as lines sampled by Bergland et al. (2014) and Lack et al. (2015) (see supplementary table 1 for details).

Figure 10: Unrooted tree detailing similarity in the Tirant composition among different *D. melanogaster* strains and population samples. The tree is based on a pairwise distance matrix of F_{ST} values computed from the allele frequencies of Tirant SNPs. Names of strains are colored by spatial or temporal origin. Old strains (< 1938) are additionally shown in bold. Note that old strains (< 1938) and most strains from Tasmania form distinct groups.

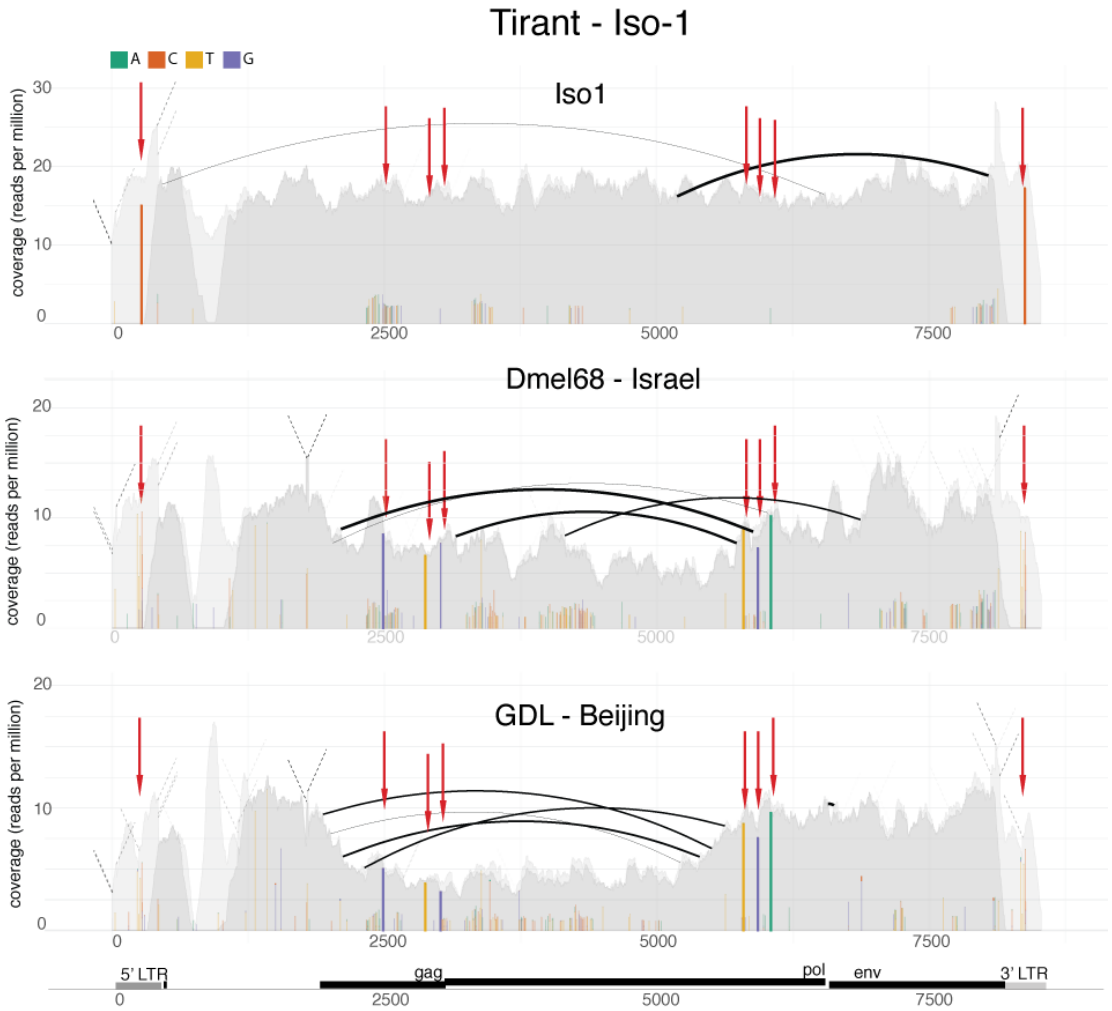

Figure 11: Abundance and diversity of Tirant in the reference strain Iso-1 and two strains collected from natural *D. melanogaster* populations (Dmel68 and a GDL line from Beijing). Eight SNPs found in natural populations but not in Iso-1 are marked by red arrows. To enhance the visibility of these SNPs, opacity of the background was reduced and the size of the SNPs was increased.

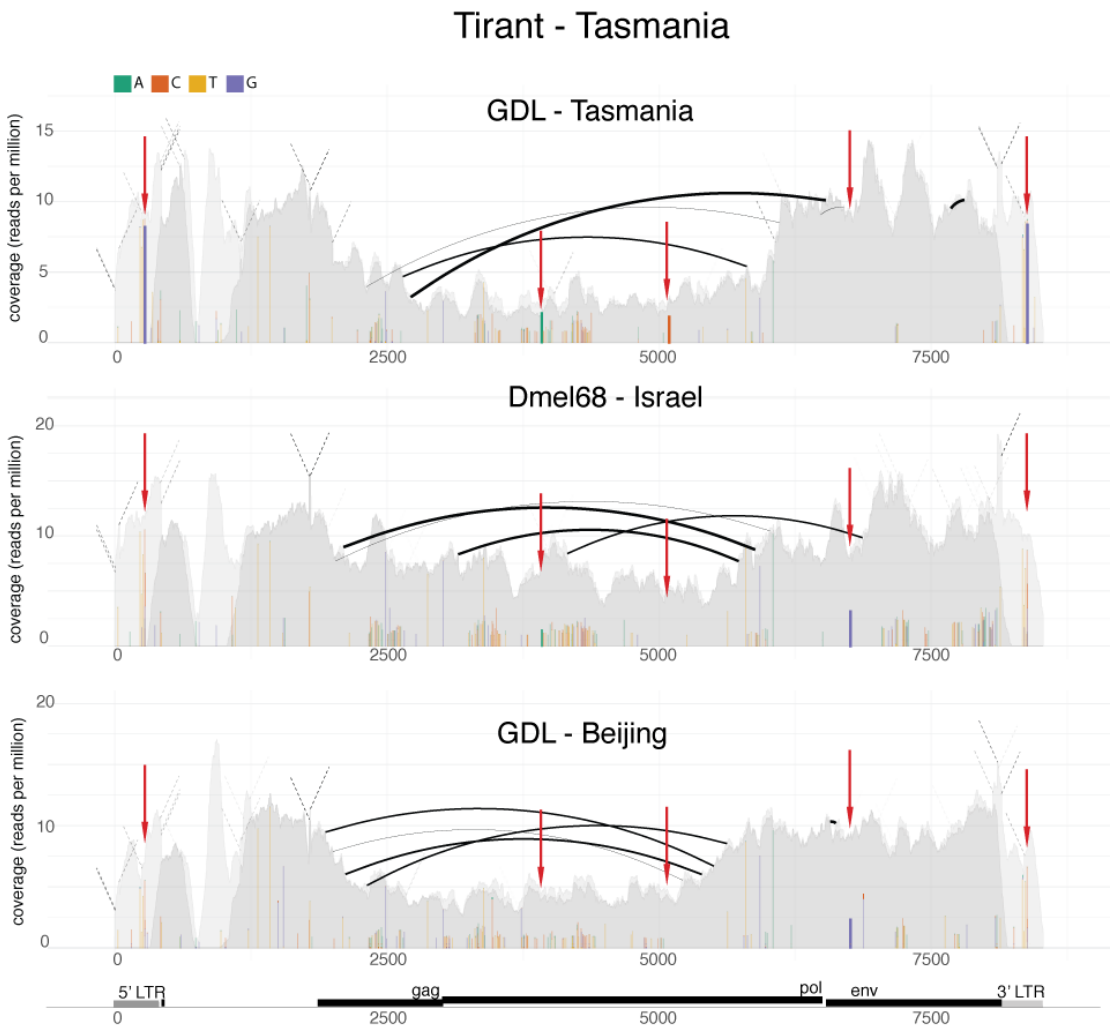

Figure 12: Abundance and diversity of Tirant sequences in a natural population from Tasmania and from other geographic locations. Five SNPs, marked by red arrows, have notably different allele frequencies between populations from Tasmania and the other geographic locations. To enhance the visibility of these SNPs, opacity of the background was reduced and the size of the SNPs was increased.

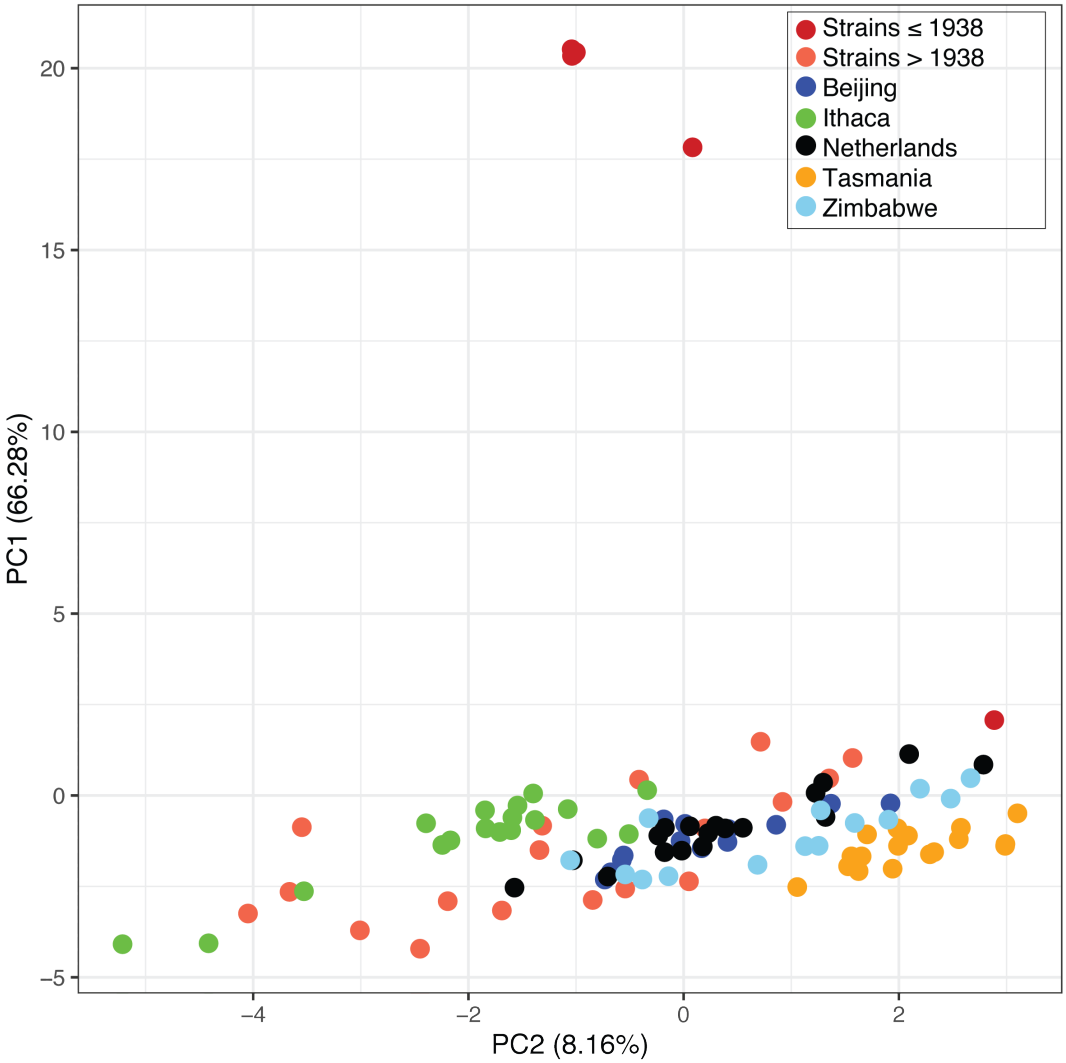

Figure 13: PCA based on the allele frequencies of SNPs in Tirant for different *D. melanogaster* strains and population samples as displayed in Figure 2. The five variants described in supplementary table 4 were omitted from the analysis, which removes the distinct clustering of the Tasmanian strains.

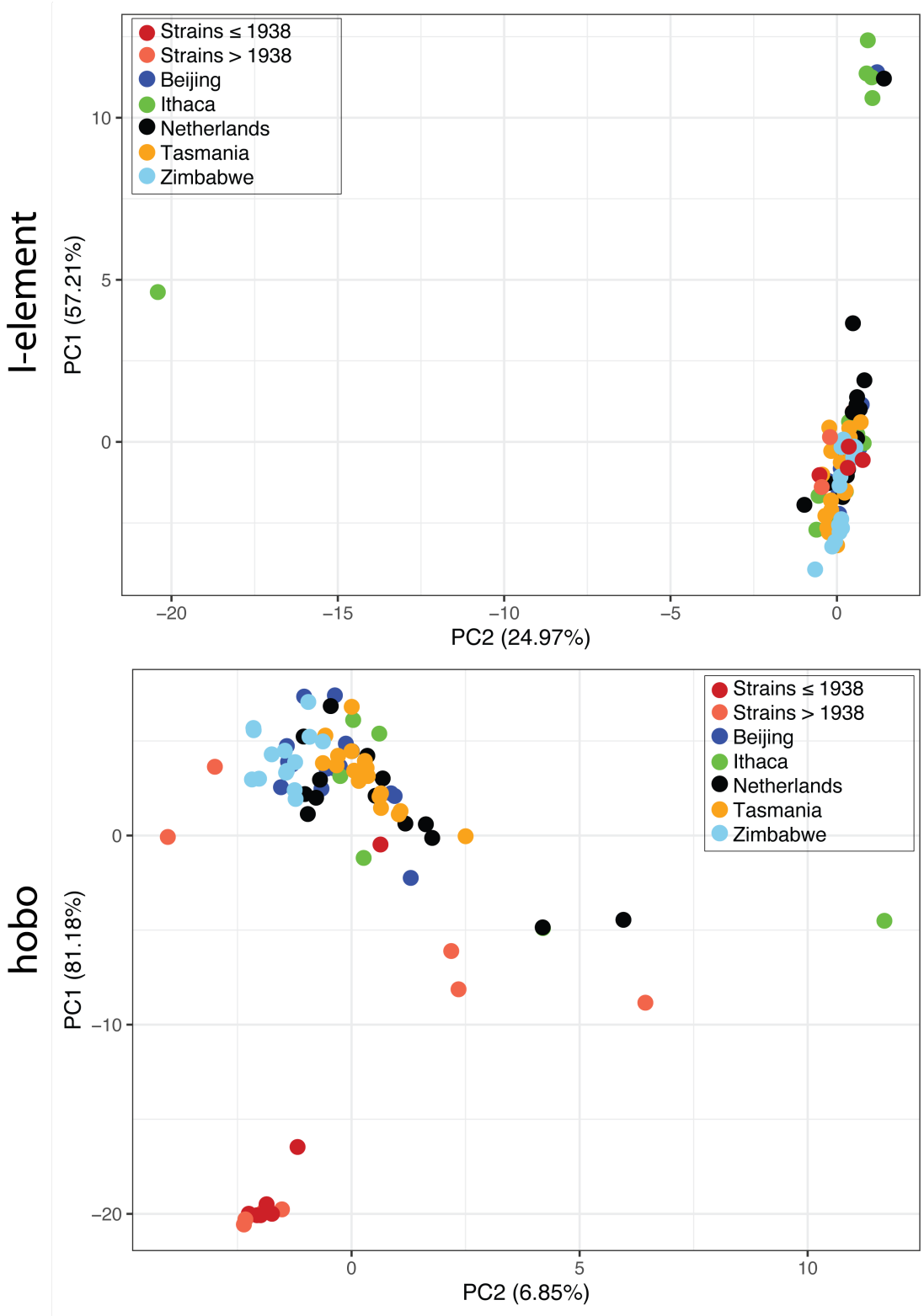

Figure 14: PCA based on the allele frequencies of SNPs in the I-element and hobo. Tasmanian populations cluster with strains from other geographic regions for both TEs.

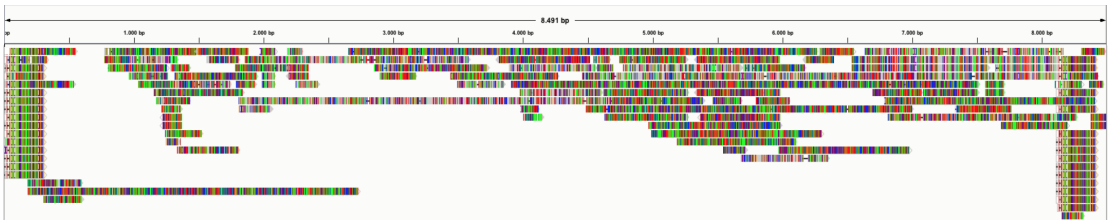

Figure 15: Alignment of all Tirant sequences annotated in Canton-S with the canonical Tirant sequence using permissive parameters. Colored lines represent SNPs with respect to the canonical Tirant. Note that a high sequence divergence can be found for all fragments over the entire sequence.

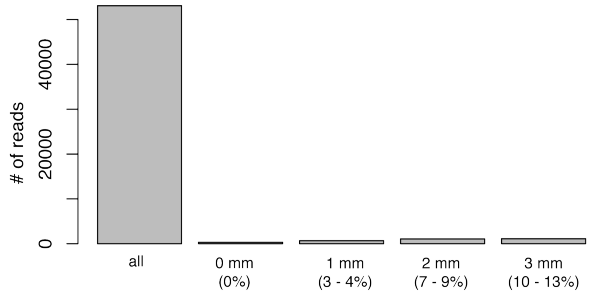

Figure 16: Number of piRNAs from heterochromatic Tirant insertions (all) mapping to the canonical Tirant sequence using different numbers of mismatches tolerated. The maximum divergence of the mapped piRNAs is shown in a brackets (based on the given number of mismatches and a piRNA length of 23-29nt).

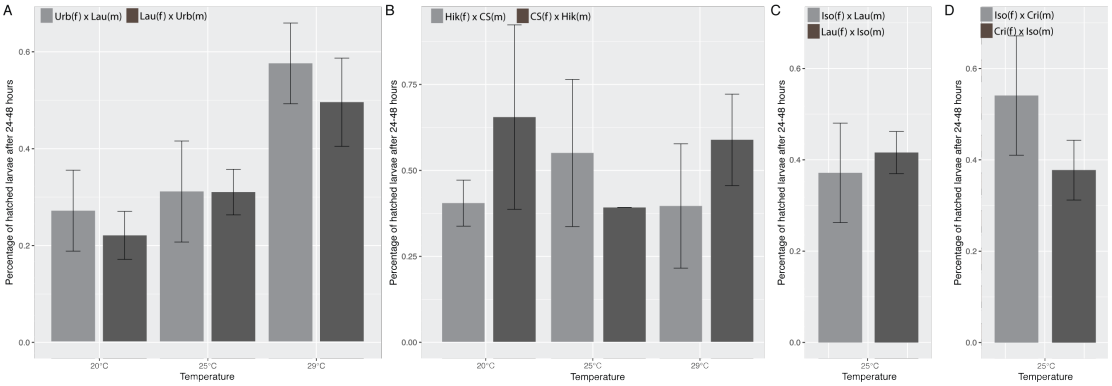

Figure 17: Fraction of hatched F2 eggs for reciprocal crosses between a strains having canonical Tirant insertions (Urb: Urbana-S, Hik: Hikone-R, Iso:Iso-1) and a strains not having canonical Tirant insertions (Lau: Lausanne-S, CS: Canton-S, Cri: Crimea). Crosses were performed at up to three different temperatures and three replicates were used for each cross. We did not detect significant differences in the abundance of hatched F2 eggs between the reciprocal crosses; Wilcoxon rank sum test; Urbana-S and Lausanne-S: $p_{20} = 0.4$, $p_{25} = 0.7$, $p_{29} = 0.4$; Hikone-R and Canton-S: $p_{20} = 0.4$, $p_{25} = 0.3$, $p_{29} = 0.4$; Iso-1 x Lausanne-S: $p_{25} = 1$; Iso-1 x Crimea: $p_{25} = 0.2$; m males, f females

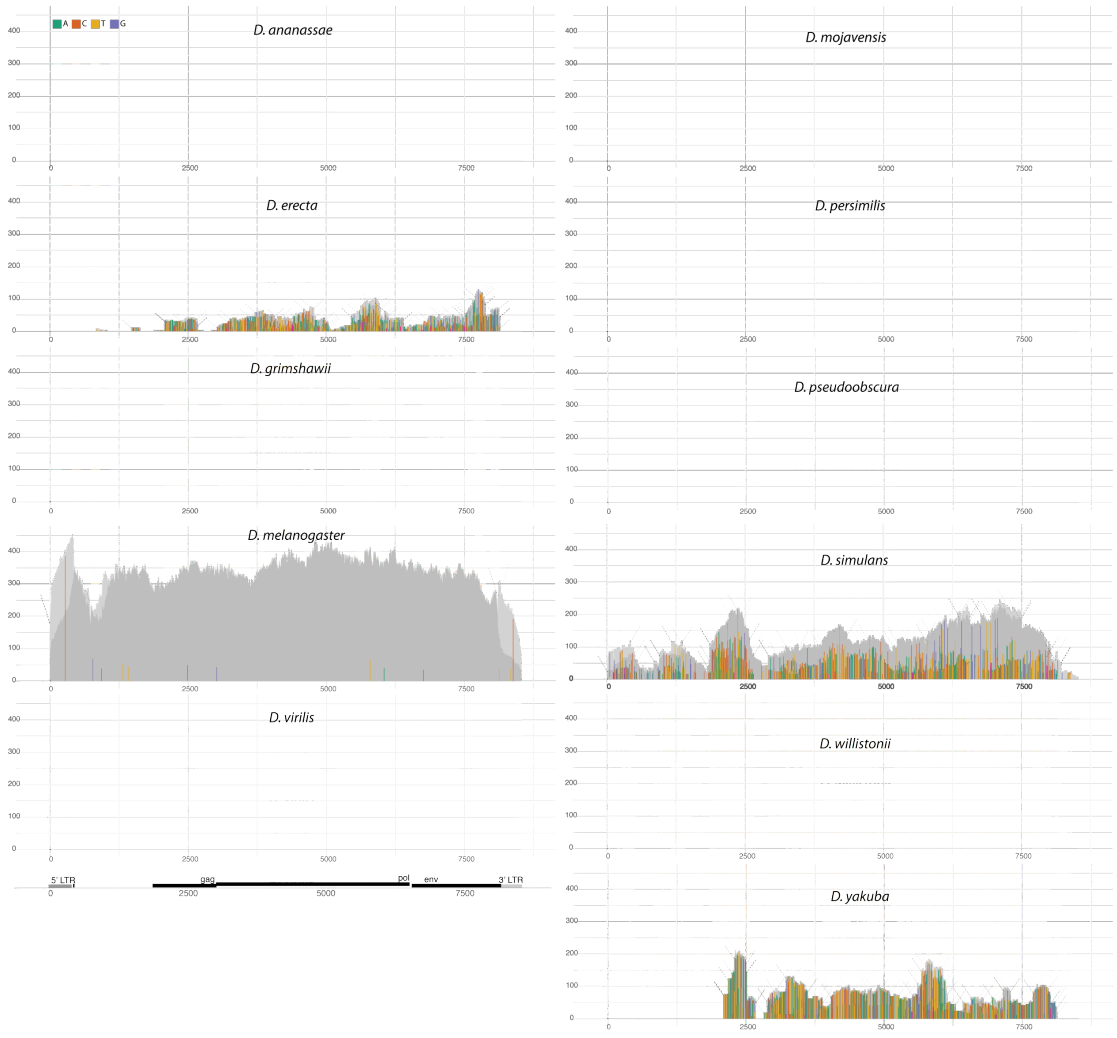

Figure 18: Abundance and diversity of Tirant sequences in 11 Drosophila species (*Drosophila* 12 Genomes Consortium, 2007). Tirant sequences can solely be found in the *Drosophila melanogaster* species subgroup. Note that solely *D. melanogaster* and *D. simulans* may have full-length insertions of Tirant. Furthermore, some insertions in *D. simulans* have a high similarity to the consensus sequence of Tirant (few SNPs in the upper regions of the DeviaTE plot).

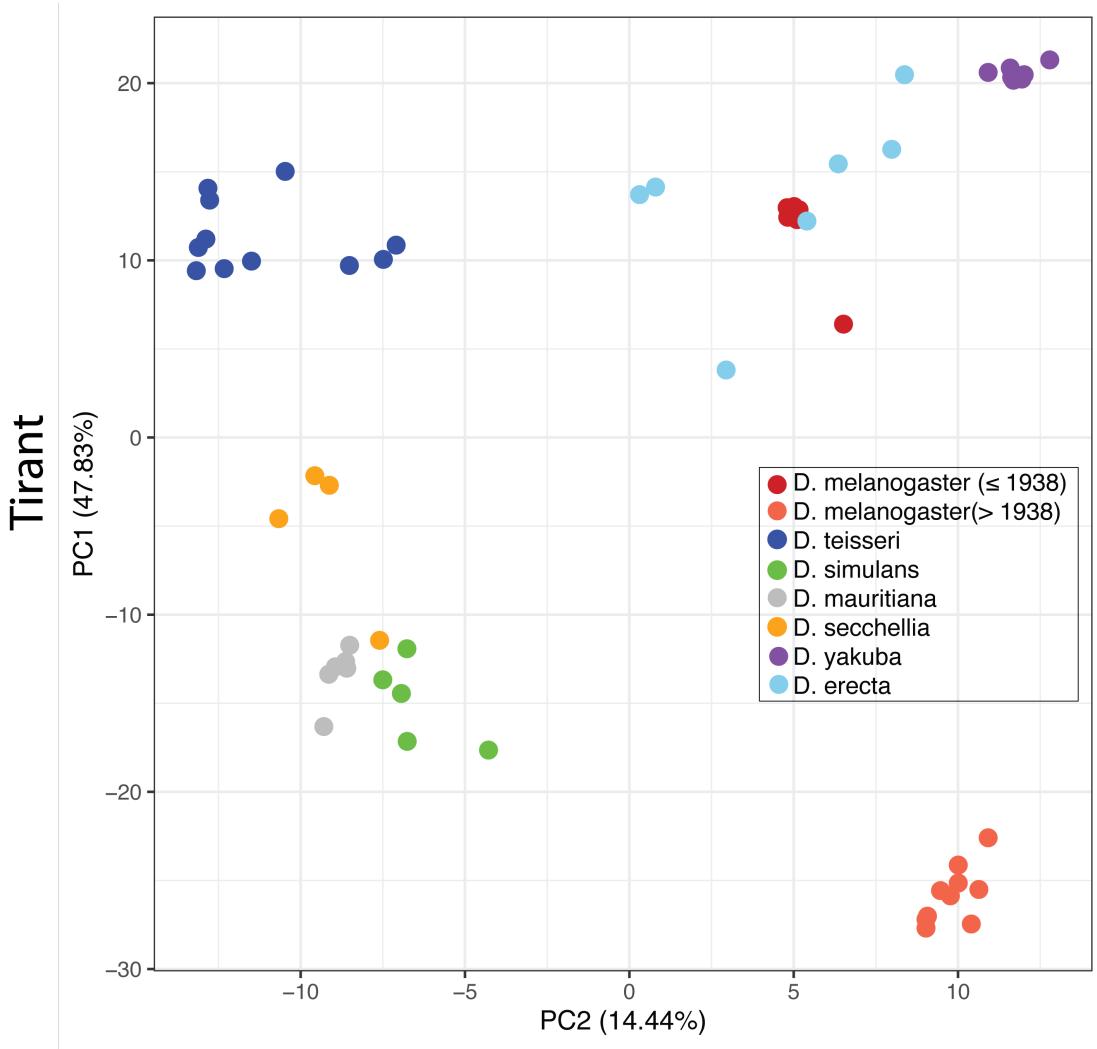

Figure 19: PCA based on the allele frequencies of SNPs in Tirant. Data are shown for several lines of different species from the *Drosophila melanogaster* species subgroup. Note that old lab strains of *D. melanogaster* cluster with *D. erecta* while more recently collected strain are closest to *D. simulans*.

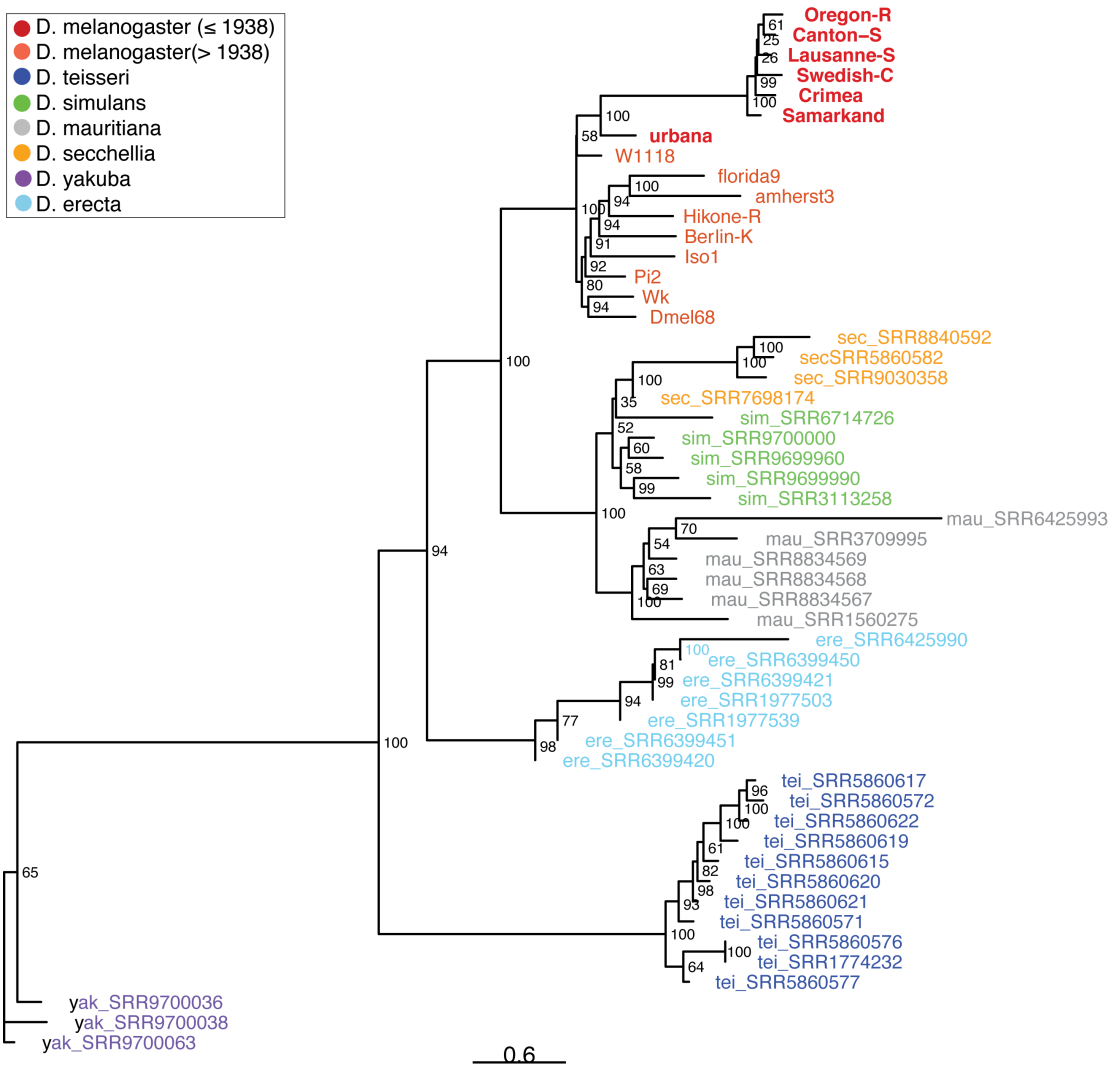

Figure 20: Phylogenetic tree calculated with PoMo based on the allele frequencies of SNPs in Tirant. Data are shown for several lines of different species from the *D. melanogaster* species subgroup. Note that the tree mostly follows the current species phylogeny of the *D. melanogaster* species subgroup (Obbard et al., 2012). *D. melanogaster* and *D. simulans* carry two distinct variants of Tirant sequences that are combined in this analysis. However, diverged Tirant sequences are most abundant in both species (82.55% of annotated Tirant sequences in *D. melanogaster* and 98.3% in *D. simulans*).

Supplementary tables

Table 1: Overview of the abundance of Tirant, I-element, hobo, and P-element sequences in different *D. melanogaster* strains. Strains are ordered by their estimated collection date. For each family and strain, we classified the TE content into three distinct categories: 'red' absence of any TE sequence, 'yellow' solely degraded TE sequences are present, 'green' non-degraded sequences, with a high similarity to the consensus sequence are present. Numbers in brackets represent the average coverage normalized to single-copy genes (≈ TE copy numbers per haploid genome). Strains sequenced in this work are marked by a star (*). † latest possible collection date was inferred from death of C. Bridges (1938), who collected the strain (Lindsley and Grell, 1968). coll. date collection date, FlyBase `https://flybase.org/`, NDSSC `https://www.drosophilaspecies.com/`

strain	coll. date	Tirant	I-ele.	hobo	P-ele.	location	source
Oregon-R	1925	~ (0.6)	~ (19.9)	~ (5.8)	− (0)	Oregon, USA	Lindsley and Grell 1968
Canton-S	1935	~ (0.9)	~ (19.5)	~ (5.9)	− (0)	Ohio, USA	Anxolabéhère *et al.* 1988
Samarkand	1936	~ (0.7)	~ (17.6)	~ (4.3)	− (0)	Samarkand, Uzbekistan	Lindsley and Grell 1968
Crimea*	1936	~ (0.5)	~ (18.5)	~ (4.5)	− (0)	Crimea, Eastern Europe	Anxolabéhère *et al.* 1988
Lausanne-S*	1938	~ (0.5)	~ (18.8)	~ (3.7)	− (0)	Wisconsin, USA	Lindsley and Grell 1968
Swedish-C*	<1938(1923)	~ (0.6)	+ (37.5)	+ (27.7)	~ (0.3)	Stockholm, Sweden	Lindsley and Grell 1968
Urbana-S*	<1938	+ (2.4)	~ (21.9)	~ (5.5)	− (0)	Illinois, USA	Bridges†, (Lindsley and Grell, 1968)
Berlin-K*	<1950	+ (6.6)	+ (32.1)	~ (3.6)	− (0)	Berlin, Germany	Ruebenbauer *et al.* 2008
Hikone-R*	1950-59	+ (6.2)	~ (17.9)	~ (5.8)	− (0)	Japan	Galindo *et al.* 1995
Florida-9*	<1952	+ (7.6)	+ (31.4)	+ (26.9)	+ (77.3)	Florida, USA	Lindsley and Grell 1968
Dmel68*	1954	+ (14.1)	+ (40.6)	+ (16.2)	− (0)	Israel	NDSSC
B1(BER1)	1954	+ (15.8)	+ (30.0)	~ (2.6)	− (0)	Bermuda	FlyBase
A3(BS1)	1954	+ (13.7)	+ (23.3)	~ (2.4)	− (0)	Barcelona, Spain	FlyBase
B2(CA1)	1954	+ (8.2)	+ (33.7)	~ (1.3)	− (0)	Capetown, South Africa	FlyBase
B3(QI2)	1954	+ (10.2)	+ (30.3)	+ (11.1)	− (0)	Israel	FlyBase
A2(BOG1)	1962	+ (18.2)	+ (40.7)	+ (19.5)	− (0)	Bogota, Colombia	FlyBase
A4(KSA2)	1963	+ (3.0)	+ (29.4)	~ (1.3)	− (0)	Koriba Dam, Zimbabwe	FlyBase
B4(RVC3)	1963	+ (15.8)	+ (44.1)	+ (64.7)	− (0)	California, USA	FlyBase
A5(VAG1)	1965	+ (7.7)	+ (31.4)	+ (10.9)	− (0)	Athens, Greece	FlyBase
A6(wild5B)	1966	+ (15.2)	+ (31.3)	+ (84.2)	− (0)	Georgia, USA	FlyBase
Harwich	1967	+ (5.5)	+ (55.2)	+ (12.9)	+ (60.1)	Massachusets, USA	NDSSC
Pi2*	1975	+ (8.8)	+ (31.7)	+ (98.9)	+ (39.7)	N.A.	Engels 1979
w1118*	<1987	+ (5.2)	+ (40.5)	+ (35.5)	− (0)	N.A.	first used by Black *et al.* 1987
AB8 (Sam;ry506)	N.A.	~ (0.5)	~ (28.1)	~ (2.5)	− (0)	N.A.	N.A.
wk*	N.A.	+ (10.4)	~ (18.0)	~ (4.7)	− (0)	N.A.	N.A.
Amherst-3*	N.A.	+ (11.8)	+ (35.2)	~ (4.7)	− (0)	Massachusets, USA	N.A.
Iso1	N.A.	+ (20.9)	+ (32.0)	+ (28.6)	− (0)	N.A.	N.A.

Table 2: The abundance of Tirant, I-element, hobo, and P-element sequences in different *D. melanogaster* strains estimated with two different approaches: i) with DeviaTE based on the coverage of single copy genes (scg) and ii) as the normalized number of reads mapping to each TE (rpm; reads per million).For the sampling data of strains see supplementary table 1

	DeviaTE with scg.				rpm			
strain	Tirant	I-element	hobo	P-element	Tirant	I-element	hobo	P-element
Oregon-R	0.6	19.9	5.8	0.0	43.04	820.95	126.72	0.025
Canton-S	0.9	19.5	5.9	0.0	43.64	747.89	126.67	0.0
Samarkand	0.7	17.6	4.3	0.0	46.63	715.08	92.29	0.025
Crimea*	0.5	18.5	4.5	0.0	37.44	639.57	85.52	0.0
Lausanne-S*	0.5	18.8	3.7	0.0	46.13	729.53	79.53	0.054
Swedish-C*	0.6	37.5	27.7	0.3	47.70	1461.67	506.78	6.33
Urbana-S*	2.4	21.9	5.5	0.0	149.03	809.59	113.07	0.076
Berlin-K*	6.6	32.1	3.6	0.0	341.16	1114.81	71.35	0.0
Hikone-R*	6.2	17.9	5.8	0.0	332.03	689.45	122.49	0.0
Florida-9*	7.6	31.4	26.9	77.3	425.74	1168.60	428.50	857.25
Dmel68*	14.1	40.6	16.2	0.0	538.57	1355.19	329.42	0.049
B1(BER1)	15.8	30.0	2.6	0.0	701.70	907.19	32.63	0.0
A3(BS1)	13.7	23.3	2.4	0.0	609.78	1306.47	356.14	0.0
B2(CA1)	8.2	33.7	1.3	0.0	348.53	697.04	27.95	0.0
B3(QI2)	10.2	30.3	11.1	0.0	483.64	1072.19	15.93	0.0
A2(BOG1)	18.2	40.7	19.5	0.0	832.94	967.37	15.76	0.0
A4(KSA2)	3.0	29.4	1.3	0.0	129.29	1012.27	209.47	0.0
B4(RVC3)	15.8	44.1	64.7	0.0	698.92	1408.05	1216.57	0.0
A5(VAG1)	7.7	31.4	10.9	0.0	349.46	1084.80	211.92	0.0
A6(wild5B)	15.2	31.3	84.2	0.0	720.72	1016.0	1645.18	0.0
Harwich	5.5	55.2	12.9	60.1	275.29	1953.82	232.75	955.57
Pi2*	8.8	31.7	98.9	39.7	334.27	1206.16	1632.15	566.52
w1118*	5.2	40.5	35.5	0.0	186.42	1456.42	703.03	0.0
AB8	0.5	28.1	2.5	0.0	13.74	805.10	31.17	0.0
wk*	10.4	18.0	4.7	0.0	457.35	632.20	94.52	0.062
Amherst-3*	11.8	35.2	4.7	0.0	629.89	1296.54	96.80	0.0
Iso1	20.9	32.0	28.6	0.0	1054.42	1231.07	440.42	0.0

Table 3: Position in Tirant (pos.), reference allele (ref.) and frequency of the reference allele for SNPs with notable allele frequency differences between Iso-1 and natural populations (GDL, DrosEU and Dros-RTEC). For an overview of all SNPs in Iso-1 and some GDL lines see supplementary fig. 11.

pos	refbase	Iso-1 (SRR1663590)	GDL-Other (SRR1663540)	GDL-Other (SRR1663560)	GDL-Other (SRR1663600)	DrosEU (SRR5647729)	DrosEU (SRR5647776)	Dros-RTEC (SRR3590550)	Dros-RTEC (SRR3939104)
231	C	0.948	0.013	0	0	0.011	0	0	0
2486	A	0.792	0	0	0	0.046	0.036	0.038	0.057
2873	G	0.937	0	0	0	0.054	0.027	0.028	0.073
3015	A	0.891	0	0.167	0	0.032	0.023	0.033	0.05
5794	C	0.908	0	0	0	0	0	0	0
5926	A	0.931	0.134	0.152	0.227	0.131	0.14	0.042	0.103
6046	G	0.882	0	0	0	0	0.012	0.053	0.047
8338	C	0.931	0	0	0	0.004	0	0	0.003

Table 4: Position in Tirant (pos), reference allele (ref.) and frequency of the reference allele for SNPs with notable allele frequency differences between populations from Tasmania and other geographic locations (GDL). For an overview of all SNPs in Tasmanian and non-Tasmanian populations see supplementary fig. 12.

pos	ref.	GDL-Tasm. (SRR1663590)	GDL-Tasm. (SRR1663591)	GDL-Tasm. (SRR1663592)	GDL-Other (SRR1663540)	GDL-Other (SRR1663560)	GDL-Other (SRR1663600)
276	T	0.071	0.04	0.097	0.93	0.336	0.873
3922	G	0.075	0.111	0.084	0.68	0.824	0.791
5092	T	0.396	0.682	0.923	1	1	1
6758	A	0.932	1	0.993	0.734	0.366	0.439
8383	T	0.068	0.066	0.137	0.871	0.354	0.788

Table 5: Number of dysgenic and not-dysgenic ovaries in the F1 of reciprocal crosses between strains having canonical Tirant insertions (Urbana-S, Hikone-R or Iso-1) and strains not having canonical Tirant insertions (Lausanne-S, Canton-S or Crimea). Crosses were performed at up to two temperatures and three replicates were used for each cross. The direction of the cross had no significant influence on the fraction of dysgenic ovaries at both temperatures (Cochran–Mantel–Haenszel test; Urbana-S x Lausanne-S: $p_{25} = 0.736$, $p_{29} = 0.742$; Hikone-R x Canton-S: $p_{29} = 0.9611$; Iso-1 x Lausanne-S: $p_{25} = 0.867$; Iso-1 x Crimea: $p_{25} = 0.994$).

female	male	temp.	rep.	not-dysgenic	dysgenic
Urbana-S	Lausanne-S	25°C	1	17	0
Urbana-S	Lausanne-S	25°C	2	13	0
Urbana-S	Lausanne-S	25°C	3	13	0
Urbana-S	Lausanne-S	29°C	1	12	1
Urbana-S	Lausanne-S	29°C	2	18	0
Urbana-S	Lausanne-S	29°C	3	18	0
Lausanne-S	Urbana-S	25°C	1	13	0
Lausanne-S	Urbana-S	25°C	2	15	0
Lausanne-S	Urbana-S	25°C	3	15	0
Lausanne-S	Urbana-S	29°C	1	18	0
Lausanne-S	Urbana-S	29°C	2	11	0
Lausanne-S	Urbana-S	29°C	3	15	0
Hikone-R	Canton-S	29°C	1	18	0
Hikone-R	Canton-S	29°C	2	16	0
Hikone-R	Canton-S	29°C	3	16	0
Canton-S	Hikone-R	29°C	1	18	0
Canton-S	Hikone-R	29°C	2	14	0
Canton-S	Hikone-R	29°C	3	16	0
Iso1	Lausanne-S	25°C	1	16	0
Iso1	Lausanne-S	25°C	2	17	0
Iso1	Lausanne-S	25°C	3	20	0
Lausanne-S	Iso1	25°C	1	18	0
Lausanne-S	Iso1	25°C	2	19	0
Lausanne-S	Iso1	25°C	3	19	0
Iso1	Crimea	25°C	1	19	0
Iso1	Crimea	25°C	2	14	0
Iso1	Crimea	25°C	3	19	0
Crimea	Iso1	25°C	1	16	0
Crimea	Iso1	25°C	2	17	0
Crimea	Iso1	25°C	3	19	0

Table 6: Publicly available short read data used in this work

Mackay et al. (2012)	SRR018294, SRR018305, SRR018517, SRR018521, SRR018574, SRR018580, SRR018582, SRR018593, SRR018601, SRR834538, SRR834536, SRR834540, SRR834542, SRR834550, SRR834548, SRR048925, SRR834549, SRR834524, SRR834525, SRR834528, SRR834529, SRR834531, SRR834532, SRR834533, SRR834534, SRR834535, SRR834504, SRR834530, SRR834510, SRR834505, SRR834506, SRR834507, SRR834513, SRR834515, SRR834516, SRR834544, SRR834539, SRR834554, SRR834520, SRR835025, SRR835023, SRR835028, SRR835026, SRR835029, SRR051592, SRR051594, SRR835024, SRR835027, SRR835031, SRR835032, SRR835033, SRR835037, SRR835035, SRR835036, SRR835038, SRR835040, SRR835041, SRR835042, SRR835045, SRR835046, SRR051896, SRR835081, SRR835101, SRR051905, SRR835082, SRR835083, SRR835085, SRR835099, SRR835088, SRR835089, SRR835091, SRR835092, SRR835093, SRR835094, SRR835095, SRR835051, SRR835052, SRR835057, SRR835053, SRR835064, SRR835065, SRR835066, SRR835070, SRR835067, SRR835068, SRR835071, SRR835074, SRR835075, SRR835076, SRR835078, SRR835080, SRR835084, SRR835069, SRR835073, SRR834546, SRR835103, SRR060062, SRR835100, SRR060098, SRR835090, SRR835058, SRR834537, SRR834543, SRR835030, SRR835049, SRR060821, SRR835072, SRR834541, SRR835096, SRR835054, SRR834521, SRR834517, SRR834545, SRR835050, SRR933564, SRR835223, SRR835228, SRR835236, SRR835242, SRR835252, SRR835256, SRR933571, SRR933572, SRR933575, SRR933577, SRR835333, SRR933578, SRR933580, SRR933582, SRR933585, SRR933586, SRR933587, SRR933588, SRR933591, SRR933595, SRR835338, SRR933596, SRR933597, SRR933598, SRR933600, SRR835247, SRR933563, SRR932121, SRR835221, SRR933566, SRR933569, SRR933570, SRR933573, SRR835331, SRR835326, SRR933581, SRR933589, SRR933593, SRR933594, SRR835341, SRR835345, SRR933599, SRR835349, SRR835329, SRR933583, SRR933592, SRR933601, SRR933579, SRR835343, SRR835939,
Lack et al. (2015)	SRR189040, SRR1686796, SRR1688222, SRR189389, SRR306623, SRR306611, SRR203502, SRR306616, SRR306618
Bergland et al. (2014)	SRR1525685, SRR1525694, SRR1525695, SRR1525696, SRR1525697, SRR1525698, SRR1525699, SRR1525768, SRR1525769, SRR1525770, SRR1525771, SRR1525772, SRR1525773, SRR1525774, SRR2006283
Kapun et al. (2020)	SRR5647729, SRR5647730, SRR5647731, SRR5647732, SRR5647733, SRR5647734, SRR5647735, SRR5647736, SRR5647737, SRR5647738, SRR5647739, SRR5647740, SRR5647741, SRR5647742, SRR5647743, SRR5647744, SRR5647745, SRR5647746, SRR5647747, SRR5647748, SRR5647749, SRR5647750, SRR5647751, SRR5647752, SRR5647753, SRR5647754, SRR5647755, SRR5647756, SRR5647757, SRR5647758, SRR5647759, SRR5647760, SRR5647761, SRR5647762, SRR5647763, SRR5647764, SRR5647765, SRR5647766, SRR5647767, SRR5647768, SRR5647769, SRR5647770, SRR5647771, SRR5647772, SRR5647773, SRR5647774, SRR5647775, SRR5647776
Machado et al. (2019)	SRR3590550, SRR3590551, SRR3590554, SRR3590555, SRR3590556, SRR3590557, SRR3590558, SRR3590559, SRR3590560, SRR3590561, SRR3590562, SRR3590563, SRR3939042, SRR3939043, SRR3939044, SRR3939045, SRR3939046, SRR3939047, SRR3939048, SRR3939049, SRR3939050, SRR3939051, SRR3939052, SRR3939054, SRR3939056, SRR3939057, SRR3939058, SRR3939059, SRR3939076, SRR3939077, SRR3939078, SRR3939080, SRR3939081, SRR3939082, SRR3939083, SRR3939084, SRR3939085, SRR3939086, SRR3939087, SRR3939088, SRR3939089, SRR3939091, SRR3939092, SRR3939093, SRR3939094, SRR3939095, SRR3939096, SRR3939097, SRR3939098, SRR3939099, SRR3939100, SRR3939101, SRR3939102, SRR3939103, SRR3939104, SRR8061818, SRR8061819
Grenier et al. (2015)	SRR1663528, SRR1663529, SRR1663530, SRR1663531, SRR1663532, SRR1663533, SRR1663534, SRR1663535, SRR1663536, SRR1663537, SRR1663538, SRR1663539, SRR1663540, SRR1663541, SRR1663542, SRR1663543, SRR1663544, SRR1663545, SRR1663546, SRR1663547, SRR1663548, SRR1663549, SRR1663550, SRR1663551, SRR1663552, SRR1663553, SRR1663554, SRR1663555, SRR1663556, SRR1663557, SRR1663558, SRR1663559, SRR1663560, SRR1663561, SRR1663562, SRR1663563, SRR1663564, SRR1663565, SRR1663566, SRR1663567, SRR1663568, SRR1663569, SRR1663570, SRR1663571, SRR1663572, SRR1663573, SRR1663574, SRR1663575, SRR1663576, SRR1663577, SRR1663578, SRR1663579, SRR1663580, SRR1663581, SRR1663582, SRR1663583, SRR1663584, SRR1663585, SRR1663586, SRR1663587, SRR1663588, SRR1663589, SRR1663590, SRR1663591, SRR1663592, SRR1663593, SRR1663594, SRR1663595, SRR1663596, SRR1663597, SRR1663598, SRR1663599, SRR1663600, SRR1663601, SRR1663602, SRR1663603, SRR1663604, SRR1663605, SRR1663606, SRR1663607, SRR1663608, SRR1663609, SRR1663610, SRR1663611
Wierzbicki et al. (2020)	SRR11460805, SRR11460802, SRR11460799
Jakšić et al. (2017)	SRR5851905, SRR5851906
Garrigan et al. (2014)	SRR1560275
Turissini et al. (2015)	SRR1774232
Hill et al. (2016)	SRR3113258
Schrider et al. (2018)	SRR5860571, SRR5860572, SRR5860576, SRR5860577, SRR5860582
Lanno et al. (2019)	SRR5860615, SRR5860617, SRR5860619, SRR5860620, SRR5860621, SRR5860622
Miller et al. (2018)	SRR6425993
Kang et al. (2019)	SRR6714726
Melvin et al. (2018)	SRR7698174
Meany et al. (2019)	SRR8834567, SRR8834568, SRR8834569
Cooper et al. (2019)	SRR8840592
Garrigan et al. (2012)	SRR9030358, SRR9030360
Rogers et al. (2014)	SRR9699990, SRR9700000
Stewart and Rogers (2019)	SRR9700024, SRR9700028, SRR9700036, SRR9700038, SRR9700041, SRR9700049, SRR9700063

Table 7: Matches between the consensus sequence of Tirant and a long-read based assembly of *D. simulans* (strain w^{XD1}) (Chakraborty et al., 2020). Consecutive matches (having the same ID) from position 1(2) to 8,526 (i.e. the length of Tirant) of the query represent full-length insertions of Tirant. . For each hit we show the divergence in percent (div.) and the position in the reference genome and in the query sequence (chr, chromosome). The average divergence of the three reported insertions from the consensus sequence of Tirant is d_1 = 1.97%, d_2 = 1.56%, d_3 = 1.60;

		reference genome			query (Tirant)	
ID	div.	chr.	start	end	start	end
1	1.4	2R	3,374,847	3,376,593	8,526	6,797
1	1.3	2R	3,376,594	3,376,982	6,735	6,347
1	1.8	2R	3,376,967	3,379,835	3,818	948
1	3.8	2R	3,379,538	3,380,475	934	2
2	1.3	X	21,362,067	21,363,796	8,526	6,797
2	1.3	X	21,363,793	21,365,739	6,739	4,793
2	1.5	X	21,365,737	21,369,723	4,760	744
2	2.6	X	21,369,245	21,370,341	1,116	1
3	1.6	Y_7	26,081	27,810	8,526	6,797
3	1.3	Y_7	27,811	33,493	6,735	1,050
3	3.7	Y_7	33,298	34,131	832	2

References

Bergland, A. O., Behrman, E. L., O'Brien, K. R., Schmidt, P. S., and Petrov, D. A. (2014). Genomic Evidence of Rapid and Stable Adaptive Oscillations over Seasonal Time Scales in *Drosophila*. *PLoS Genetics*, 10(11).

Chakraborty, M., Chang, C.-H., Khost, D., Vedanayagam, J., Adrion, J. R., Liao, Y., Montooth, K. L., Meiklejohn, C. D., Larracuente, A. M., and Emerson, J. J. (2020). Evolution of genome structure in the *Drosophila simulans* species complex. *bioRxiv*.

Chakraborty, M., Emerson, J. J., Macdonald, S. J., and Long, A. D. (2019). Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nature Communications*, 10(1):419275.

Cooper, J. C., Guo, P., Bladen, J., and Phadnis, N. (2019). A triple-hybrid cross reveals a new hybrid incompatibility locus between *D. melanogaster* and *D. sechellia*. *bioRxiv*, page 590588.

Drosophila 12 Genomes Consortium (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, 450(7167):203–18.

Garrigan, D., Kingan, S. B., Geneva, A. J., Andolfatto, P., Clark, A. G., Thornton, K. R., and Presgraves, D. C. (2012). Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. *Genome Research*, 22(8):1499–1511.

Garrigan, D., Kingan, S. B., Geneva, A. J., Vedanayagam, J. P., and Presgraves, D. C. (2014). Genome diversity and divergence in *Drosophila mauritiana*: Multiple signatures of faster X evolution. *Genome Biology and Evolution*, 6(9):2444–2458.

Grenier, J. K., Roman Arguello, J., Moreira, M. C., Gottipati, S., Mohammed, J., Hackett, S. R., Boughton, R., Greenberg, A. J., and Clark, A. G. (2015). Global diversity lines-a five-continent reference panel of sequenced *Drosophila melanogaster* strains. *G3: Genes, Genomes, Genetics*, 5(4):593–603.

Hill, T., Schlötterer, C., and Betancourt, A. J. (2016). Hybrid Dysgenesis in *Drosophila simulans* Associated with a Rapid Invasion of the P-Element. *PLoS Genetics*, 12(3):1–17.

Hoskins, R. A., Carlson, J. W., Wan, K. H., Park, S., Mendez, I., Galle, S. E., Booth, B. W., Pfeiffer, B. D., George, R. A., Svirskas, R., et al. (2015). The Release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome research*, 25(3):445–458.

Jakšić, A. M., Kofler, R., and Schlötterer, C. (2017). Regulation of transposable elements: Interplay between TE-encoded regulatory sequences and host-specific trans-acting factors in *Drosophila melanogaster*. *Molecular Ecology*, 26(19):5149–5159.

Kang, L., Rashkovetsky, E., Michalak, K., Garner, H. R., Mahaney, J. E., Rzigalinski, B. A., Korol, A., Nevo, E., and Michalak, P. (2019). Genomic divergence and adaptive convergence in *Drosophila simulans* from Evolution Canyon, Israel. *Proceedings of the National Academy of Sciences*, 116(24):11839 – 11844.

Kapun, M., Barrón, M. G., Staubach, F., Obbard, D. J., Wiberg, R. A. W., Vieira, J., Goubert, C., Rota-Stabelli, O., Kankare, M., Bogaerts-Márquez, M., et al. (2020). Genomic Analysis of European *Drosophila melanogaster* Populations Reveals Longitudinal Structure, Continent-Wide Selection, and Previously Unknown DNA Viruses. *Molecular Biology and Evolution*, 37(9):2661–2678.

Lack, J. B., Cardeno, C. M., Crepeau, M. W., Taylor, W., Corbett-Detig, R. B., Stevens, K. A., Langley, C. H., and Pool, J. E. (2015). The *Drosophila* genome nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. *Genetics*, 199(4):1229–1241.

Lanno, S. M., Shimshak, S. J., Peyser, R. D., Linde, S. C., and Coolon, J. D. (2019). Investigating the role of Osiris genes in *Drosophila sechellia* larval resistance to a host plant toxin. *Ecology and Evolution*, 9(4):1922–1933.

Lindsley, D. H. and Grell, E. H. (1968). *Genetic variations of Drosophila melanogaster*. Carnegie Institute of Washington Publication.

Machado, H. E., Bergland, A. O., Taylor, R., Tilk, S., Behrman, E., Dyer, K., Fabian, D. K., Flatt, T., González, J., Karasov, T. L., et al. (2019). Broad geographic sampling reveals predictable, pervasive, and strong seasonal adaptation in *Drosophila*. *bioRxiv*, page 337543.

Mackay, T. F., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., Han, Y., Magwire, M. M., Cridland, J. M., et al. (2012). The *Drosophila melanogaster* genetic reference panel. *Nature*, 482(7384):173–178.

Meany, M. K., Conner, W. R., Richter, S. V., Bailey, J. A., Turelli, M., and Cooper, B. S. (2019). Loss of cytoplasmic incompatibility and minimal fecundity effects explain relatively low Wolbachia frequencies in *Drosophila mauritiana*. *Evolution*, 73(6):1278–1295.

Melvin, R. G., Lamichane, N., Havula, E., Kokki, K., Soeder, C., Jones, C. D., and Hietakangas, V. (2018). Natural variation in sugar tolerance associates with changes in signaling and mitochondrial ribosome biogenesis. *eLife*, 7:e40841.

Miller, D. E., Staber, C., Zeitlinger, J., and Hawley, R. S. (2018). Highly Contiguous Genome Assemblies of 15 *Drosophila* Species Generated Using Nanopore Sequencing . *G3; Genes|Genomes|Genetics*, 8(10):3131–3141.

Obbard, D. J., Maclennan, J., Kim, K.-W., Rambaut, A., O'Grady, P. M., and Jiggins, F. M. (2012). Estimating Divergence Dates and Substitution Rates in the *Drosophila* Phylogeny. *Molecular Biology and Evolution*, 29(11):3459–3473.

Riddle, N. C., Minoda, A., Kharchenko, P. V., Alekseyenko, A. A., Schwartz, Y. B., Tolstorukov, M. Y., Gorchakov, A. A., Jaffe, J. D., Kennedy, C., Linder-Basso, D., et al. (2011). Plasticity in patterns of histone modifications and chromosomal proteins in *Drosophila* heterochromatin. *Genome Research*, 21(2):147–163.

Rogers, R. L., Cridland, J. M., Shao, L., Hu, T. T., Andolfatto, P., and Thornton, K. R. (2014). Landscape of standing variation for tandem duplications in *Drosophila yakuba* and *Drosophila simulans*. *Molecular biology and evolution*, 31(7):1750–1766.

Schrider, D., Ayroles, J., Matute, D., and Kern, A. (2018). Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. *PLOS Genetics*, 14(4):170670.

Stewart, N. B. and Rogers, R. L. (2019). Chromosomal rearrangements as a source of new gene formation in *Drosophila yakuba*. *PLOS Genetics*, 15(9):e1008314.

Turissini, D. A., Liu, G., David, J. R., and Matute, D. R. (2015). The evolution of reproductive isolation in the *Drosophila yakuba* complex of species. *Journal of Evolutionary Biology*, 28(3):557–575.

Wierzbicki, F., Schwarz, F., Cannalonga, O., and Kofler, R. (2020). Generating high quality assemblies for genomic analysis of transposable elements. *bioRxiv*.