Novel computational strategies for the analysis of transposable elements in *Drosophila* cell culture genomes

by

Shunhua Han

(Under the Direction of Casey Bergman)

Abstract

Cell culture systems are widely used in molecular biology, yet studies using cultured cells suffer from irreproducible outcomes. One leading cause of irreproducible research comes from the misidentification of cell lines. Genotyping protocols have been established to authenticate human cell lines but are lacking for cell cultures derived from many important model species, including *Drosophila melanogaster*. Irreproducible research on cell lines could also arise by different mechanisms of genome evolution that are poorly understood, including copy number changes, structural variations, and TE amplification. My work focuses on developing new computational strategies to understand TE dynamics and evolution in the *Drosophila* cell culture system.

First, I utilized the classical observation that TEs can somatically proliferate in *Drosophila* cell culture to develop a novel framework that uses genome-wide TE insertion profiles to identify cell line origin and reveals the relationship among different *Drosophila* cell lines and sub-lines. Using this framework, I found that several *Drosophila* cell lines (Sg4, mbn2, and OSS_E) were misidentified, and that a subset of LTR retrotransposons is sufficient for cell line authentication. I also developed a short-read-based TE detection approach called ngs_te_mapper2, which provided the first evidence that loss of heterozygosity is a mechanism of shaping genome evolution and TE profiles in *Drosophila* cell culture.

Next, I used genome-wide TE profiles for multiple S2 sub-lines to understand whether TE amplification in cell culture is due to an initial burst of transposition after cell line establishment or ongoing transposition during routine cell culture. My results provided strong evidence for ongoing transposition model in cell culture and revealed extensive copy number diversity among S2 sub-lines. This work suggests that TE and copy number variations could lead to genomic changes in commonly used cell lines, which may significantly impact functional studies.

Finally, I developed a long-read-based TE detection approach called TELR and applied it to a tetraploid *Drosophila* cell line called S2R+. My results revealed many TEs that somatically inserted after S2R+ cells became tetraploid. Phylogenomic analysis of *in vitro* TE insertions also revealed that the TE amplification in cell culture could arise from a single or multiple source lineages.

INDEX WORDS: Transposable element, Cell culture, *Drosophila*, Genome evolution, Dissertations, Theses (academic)

Novel computational strategies for the analysis of transposable elements in *Drosophila* cell culture genomes

by

Shunhua Han

B.S., East China University of Science and Technology

China, July 1st, 2015

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2021

Novel computational strategies for the analysis of transposable elements in *Drosophila* cell culture genomes

by

Shunhua Han

Approved:

Major Professor: Casey Bergman

Committee: Jim Leebens-Mack
Liang Liu
Kelly Dyer

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
December 2021

Acknowledgments

I want to express my immense gratitude to Dr. Casey Bergman, who provided me with exceptional guidance on genomics research. His high standard and cautious attitude toward science have significantly impacted me on the path to becoming a bioinformatics scientist. I also want to thank members of the Bergman lab that I have worked with over the past five years: Dr. Guilherme Dias, Preston Basting, and Jingxuan Chen. They have been my great colleagues and collaborators.

For chapter two, I want to thank Nancy Go and Sharon Gorski (Canada's Michael Smith Genome Sciences Centre, BC Cancer) and Michael Strand (University of Georgia) for supplying samples of mbn2 cells; Noah Workman and Magdy Alabady at the University of Georgia Genomics and Bioinformatics Core for assistance with Illumina library preparation and sequencing; Dan Hultmark (Umeå University), Julius Brennecke (Institute of Molecular Biotechnology, Vienna) and Nelson Lau (Boston University) for information about the history of mbn2, OSS and OSC cell lines; and Nelson Lau and two anonymous reviewers for helpful comments on the manuscript.

For chapter three, I want to thank Stacey Holden and Andy Hayes (University of Manchester Genomic Technologies Core Facility), Marissa Howard, Jeffrey Wagner, Julia Portocarrero, and Magdy Alabady (University of Georgia Genomics and Bioinformatics Core), and Kariena Dill and Shaune Hall (Dovetail Genomics) for assistance with Illumina library preparation and sequencing; Shan-Ho Tsai and Yecheng Huang (University of Georgia) for bioinformatics application support; and the Georgia Advanced Computing Resource Center (University of Georgia) for computing time. We thank members of the Bergman Lab (University of Manchester and University of Georgia), and the Dyer, Hall, Sweigart and White Labs (University of Georgia) for helpful suggestions throughout the project. This work was supported

by Wellcome Trust Award 096602/B/11/Z (MGN), University of Georgia Research Education Award Traineeship (PJB), Human Frontier Science Program grant RGY0093/2012 (CMB), and the University of Georgia Research Foundation (CMB).

For chapter four, I want to thank Stuart Macdonald (University of Kansas) for providing fly stocks; Christina McHenry and Robert Lyons at the University of Michigan Biomedical Research Core Facilities for assistance with PacBio library preparation and sequencing; Noah Workman, Julia Portocarrero and Magdy Alabady at the University of Georgia Genomics and Bioinformatics Core for assistance with 10x Genomics library preparation and Illumina sequencing; the Georgia Advanced Computing Resource Center for computing time; members of the Bergman Lab for helpful comments throughout the project. This work was supported by the Human Frontiers of Science and Georgia Research Foundation (C.M.B.) and the Howard Hughes Medical Institute (N.P.).

My scientific career would not have been possible without the full support from my parents, Jie Zhou and Zhaowen Han. I am also grateful to my girlfriend, Xinya Qiao, who has always been by my side through all my ups and downs. Finally, I want to thank my friend Horace Zeng, who made my life in Athens a lot more fun.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 *Drosophila* cultured cell lines

1.1.1 Cell culture

Cell culture refers to the process in which cells are removed from an animal or plant and subsequently grown in a controlled condition outside of their native environment. After cells are isolated from the tissue and start proliferating using appropriate substrate and nutrient conditions, this is called primary cell culture. Cells during primary culture grow in a standard pattern that starts with a slow growth rate (lag phase) and later proliferates in exponential rate (log phase). Once cells fully cover the plate or their density exceeds the capacity of the medium or substrate, then a fraction of the multiplying cells have to be passaged (also referred to as subcultured) to a new environment for continued growth. Once the primary cell culture is successfully passaged, it becomes a cell line. Cell lines derived from primary cell culture are finite, meaning that they can only divide a limited number of times before the proliferate is ceased by a genetic process called cellular senescence [45].

Compared to primary cell lines, a continuous cell line (also referred to as immortalized cell line) is a permanently established cell culture which, due to spontaneous or artificially induced mutation, can escape normal cellular senescence that constraints the cell cycle and can proliferate indefinitely given appropriate fresh media and space [91]. As cultured cells are passaged, cells with highest growth capacity predominate, which results in relatively homogeneous genome content and eliminates almost all genomic and phenotypic variations introduced by meiotic recombination in the whole flies. Using continuous cell lines allows

researchers to bypass ethical concerns associated with the use of animal or human tissues [109]. These characteristics of continuous cell lines also allows researchers to utilize an unlimited amount of effectively homogeneous cellular materials for long-term reproducible research [70].

The first immortal human cell line is known as the "HeLa" cell line, which was derived from cervical cancer cells obtained during the treatment of Henrietta Lacks in 1951 [211]. Since its establishment, it turned out to be highly durable and prolific and has become the most widely used human cell line for biological and biomedical research. Nowadays, there are more than 3600 cell lines from over 150 different species that are cataloged by the American Type Culture Collection (ATCC) Cell Biology Collection [109]. Cell lines are now one of the major tools used in cellular and molecular biology with over 1 million articles published that use these *in vitro* systems. For example, cell lines become used as a model system to study the physiology and biochemistry of cells, the effects of drugs or chemical compounds on cells in biomedical research [24], and as cellular factories to manufacture valuable biological compounds such as vaccines or therapeutic proteins [148].

1.1.2 *Drosophila* CELL LINES

Compared to mammalian cell culture, *Drosophila* cell lines offer several advantages for biological and biomedical research. The molecular processes and biochemistry of many *Drosophila* genes are well characterized and it has a smaller genome, which means the cost of sequencing can be drastically reduced [144]. Research using *Drosophila* cell lines has been made increasing useful with a variety of technologies, which include: 1) targeted mutagenesis on *Drosophila* cell lines for genome-wide functional studies, which can be done using transposons such as *P* element [99], RNAi [71, 169], or CRISPR/Cas9 system [21, 236]; 2) transformation of *in vitro* cultured *Drosophila* cells for the overproduction of specific gene products (e.g., regulatory proteins or protein of commercial interest) [55]; and 3) using *Drosophila* cell culture as testing system of transgenic constructs before making transgenic

fly strains for studying *in vivo* gene function and regulation [144]. The increased popularity of *Drosophila* cell lines is primarily contributed by the *Drosophila* Genomics Resource Center (DGRC) that hosts over 150 diverse *Drosophila* cell lines that are available to the public [54, 144], and Model Organism Encyclopedia of DNA Elements (modENCODE) project, which led to the creation of a large amount of high-quality genomics data on *Drosophila* cell lines [48].

Now *Drosophila* cell lines can be made from primary cell culture [69, 213, 70] or from stable transformation of existing cell lines [144]. The majority of *Drosophila* cell lines currently hosted in DGRC were originally established from *D. melanogaster* strains with different tissues of origin. The most widely used *D. melanogaster* cell lines include embryonic Schneider Line 2 (S2) [213], embryonic Kc line [69] and imaginal disc lines from the Milner lab [233, 61]. Other popular *Drosophila* cell lines include fGS/OSS line derived from *bam* adult ovaries [177], the mbn2 line reportedly derived from tumorous blood cells [82], and the larval central nervous system (CNS)-derived ML-BG3-c2 line [144].

The S2 line is one of the most widely-used non-mammalian cell culture systems. They were derived from primary cultures of late-stage embryonic tissue from an unmarked stock of Oregon-R flies in December 1969 [213]. Two other cell lines, S1 (August 1969) and S3 (February 1970) were established from the same ancestral fly stock [213]. Since the original establishment of the S2 cell line, it has been distributed widely and grown more extensively than S1 and S3 cells [129, 144]. S2 has also been stably transformed into different sub-lines to produce large numbers of variants suitable for a variety of specialized purposes [144]. Transformed S2 cells that express GFP fusion proteins have been used to study protein function in mitosis [200] to identify a set of proteins involved in actin dynamics during lamella formation in *Drosophila* S2 cells [201]. Many sub-lines of S2 cells created by different labs in the *Drosophila* community have been donated back to DGRC for maintenance and distribution [144].

Several S2 sub-lines from DGRC have been extensively distributed and used by the *Drosophila*community. These include S2-DGRC, which were one of the isolates that are transferred from the Cherbas lab during DGRC establishment (see https://dgrc.bio.indiana.edu/cells/S2Isolates); and S2-DRSC, which were donated by Norbert Perrimon, initially used for RNAi-based screens at the *Drosophila* RNAi Screening Center (DRSC) and later used by the modENCODE project [48]. Another one of the most widely used S2 sub-lines is S2R+, which are reported to be derived from S2 cells and were frozen for over 25 years [245]. S2R+ cells are distinct from other S2 cell sub-lines in that they express the Dfrizzled-1 and Dfrizzled-2 membrane proteins and are more adherent to surfaces in tissue culture. This unique characteristic of S2R+ contributes to its popularity among *Drosophila* community and suggests the presence of genome evolution in S2 cell culture that affects cellular phenotypes. Nowadays, S2R+ cells are increasingly used for high-throughput RNAi and CRISPR screens, such as combining CRISPR and RNAi to identify highly reproducible and conserved synthetic lethal interactions [96] and using genome-wide CRISPR knockout screen to examine context-specific gene fitness [236]. In general, the relationship among different S2 sub-lines and the extent of their genomic or phenotypic diversity are unknown.

Extensive transcriptional and whole-genome data are available for many of the *Drosophila* cell lines [48, 127, 253, 56, 241, 129, 228], which complement genome-wide functional studies and provide resources for genome biological research. Transcriptomic analyses based on microarray and RNA-seq data on diverse panels of *Drosophila* cell lines revealed that different *Drosophila* cell lines or even sub-lines within the same cell line have distinct expression profiles [127, 56, 241, 228]. Copy number analyses using WGS data on S2 and other *Drosophila* cell lines revealed that a large portion of genome in *Drosophila* cultured cells are aneuploid [253, 129].

1.1.3 Misidentification issue in *Drosophila* cell lines

Despite continuous cell lines being widely used in biological and biomedical research, experiments on cell lines often show non-reproducible outcomes. One primary source of irreproducible research on cell lines is cross-contamination or mislabelling, which can be collectively referred to as "misidentification". Since the establishment of the first continuous cell line in the 1950s, more than 400 widely used cell lines have been reported to be misidentified, in which HeLa is the most commonly misidentified line due to its popularity [46, 142].

Cell line cross-contamination happens when foreign cells or microorganisms are accidentally introduced into the current cell culture environment [46]. With the ability to proliferate rapidly, the foreign cell line could outcompete the current cell line and take over the culture entirely within a few passages. The end results of cross-contamination is that cells with a different type of the same species or from another species are unintentionally being used in the study [62, 80, 176, 150, 97]. Another source of cell line misidentification is mislabelling due to human errors, which could occur at any stage from cell line establishment, cell line freezing, cell line distribution, to cell line cataloging. Both the issues of cross-contamination and mislabeling require cost-effective and reliable cell line authentication protocols.

Substantial effort has been invested in raising awareness of cell line misidentification, finding solutions for cell line authentication, and establishing good practice for using cell lines in the research [98, 46, 142]. For human cell lines, the most widely used technique to authenticate cell lines is short tandem repeat (STR) profiling [157, 17, 6], which has limitations and can not eradicate the human cell line misidentification problem [183]. There are other profiling-based approaches such as SNP profiling [248, 168]. However, alternative approaches have yet to be adopted as standards to authenticate human cell lines. For most non-human species, including *Drosophila*, cell line authentication protocols remain to be established [144], which contributed to the lack of misidentified non-human cell line evidence reported in ATCC. For *Drosophila*, the STR-based approach has not been used in part because of the low mutation rate of STRs in fruit flies [214].

1.1.4 Genome evolution in long-term *Drosophila* cell culture

In addition to cell line misidentification, another issue of working with continuous cell lines is that long-term cell culture may lead to genomic changes during routine passaging [98]. Previous studies have shown that cell culture may exhibit aneuploidy or heteroploidy [205, 79, 180, 13]. More recent studies using DNA sequencing approach have shown that segmental aneuploidy and other structural variations are present in cell culture genomes [253, 166, 2, 129, 172, 24, 257, 256, 141]. These genomic changes can be explained by cells undergoing crisis in the *in vitro* environment thus having to adapt to the selective pressure of growing conditions by altering genomic content and evolving advantageous genotype [129, 231]. Genome evolution in long-term cell culture could also lead to both genomic and functional diversity among sub-lines of the same cell line [24, 141].

For *Drosophila* cell culture, S2 cells have undergone polyploidization after cell line establishment from a diploid lab strain [213], and display substantial small and large-scale segmental aneuploidy in their polyploid genomes [253, 129]. S2 and other *Drosophila* cell lines also exhibit a higher abundance of TEs compared to whole flies [189, 100, 193], with TE families that are abundant in S2 being different from those amplified in other *Drosophila* cell lines [70, 193, 88, 155]. These genome content changes in *Drosophila* cell lines compared to whole flies suggest that *Drosophila* cultured cells may have different phenotypic and functional properties compared to *in vivo* cells, which could impact the interpretability and reproducibility of functional studies on animal cell culture. Therefore, it is crucial to understand genome evolution in long-term *Drosophila* cell culture.

1.2 Transposable element in *Drosophila* cell culture

1.2.1 Types of transposable element

Transposable elements (TEs) are DNA sequences that can mobilize and proliferate in the genome through transposition. TEs were first discovered by Barbara McClintock in the maize

genome during the 1940s and 1950s [159]. They are found in almost all eukaryotic genomes at significant proportions [59]. For example, more than 45% of the human genome [124] and 85% of the maize genome [212] are made up of TEs. In general, TEs can be divided into two major classes (Class I and Class II) based on their transposition mechanisms, and each major class can be sub-categorized into multiple subclasses. TEs can also be classified as autonomous or non-autonomous elements depending on whether they have open reading frames (ORFs) that encode proteins required for transposition. Autonomous TEs can transpose independently, while non-autonomous TEs rely on the machinery of autonomous TEs to mobilize in the genome.

Class I TEs are also known as retrotransposons that transpose via RNA intermediate. There are three major subclasses of retrotransposon: long terminal repeat (LTR) retrotransposons, long interspersed elements (LINEs), and short interspersed elements (SINEs) [59].

Autonomous retrotransposons including LTR retrotransposons and LINEs have *gag* and *pol* genes that are required for transposition. The RNA-intermediate-based transcription mechanism used by retrotransposons was first demonstrated for the Ty1 element in yeast [36]. In order to transpose from one genome location to another, retrotransposons are first transcribed into RNA intermediate by the RNA polymerase II. The *gag* and *pol* genes of retrotransposon mRNAs are then translated into a protein in the cytoplasm. The protease (PR) of the Pol polyprotein cleaves the peptide into integrase (IN) and reverse transcriptase (RT) enzymes. The Gag protein forms the virus-like particle (VLP). The RT, retrotransposon mRNA, and IN are then packaged into VLPs, in which the retrotransposon mRNA is reverse transcribed into cDNA using RT. The VLP is then imported into the nucleus, where the retrotransposon cDNA is incorporated into a new location in the genome with the help of IN [64, 111, 90, 25]. This "copy-and-paste" transposition mechanism used by autonomous retrotransposons allows them to quickly propagate in the eukaryotic genomes. The integration process of LTR retrotransposons involves having staggered strand transfer reactions in the target location of the genome, which creates double-strand breaks with short gaps on

both sides of the integrated TE. The host cell would then repair the double-strand breaks that gives rise to target site duplications (TSDs) [59, 243, 174, 139]. For LTR elements, the length of the TSD sequence is typically fixed, TE family-specific and generally less than 10bp [139].

Although both LTR retrotransposons and LINEs share similar core functional ORFs, the LTR retrotransposons surround *gag* and *pol* genes with LTR sequences while LINEs do not. The terminal repeat structure of LTR retrotransposons also allows homologous recombination events between LTR sequences from the same isoform that result in formation of solo-LTR element [237]. Some LTR retrotransposons, including *gypsy* and *ZAM* also include an ORF to encode envelope (env)-like protein that allows virus particles to infect other cells or organisms by horizontal gene transfer. There are two major superfamilies for LTR retrotransposons: Ty3/gypsy and Ty1/copia. Both superfamilies occur in virtually all major groups of eukaryotes [153]. The differences between Ty3/gypsy and Ty1/copia are largely on the order of functional ORFs [59].

LINEs and SINEs together are often referred to as non-LTR retrotransposon as they do not contain LTR sequences [38]. These non-LTR retrotransposons make up majority of TEs in the human genome [124], in which LINE-1 (L1) is the most abundant family [223, 110, 222], and is the only fully autonomous TEs that can actively transpose in the human genome today. LINE elements, due to their transposition mechanism, are often present in a truncated form on the 5' end that result in TE losing the ability to transpose [143]. In the *Drosophila* genome, at least 21% of LINE-like elements are full-length and potentially active [106]. SINEs are non-autonomous elements that do not encode any functional proteins for transposition, and they require LINEs to mobilize in the genome. Unlike LINEs, SINEs are not present in *Drosophila* genome and thus will not be discussed further here. The transposition mechanism of non-LTR retrotransposons is similar to LTR retrotransposons on using RNA intermediate and reverse transcription. However, non-LTR retrotransposons use a process called target primed reverse transcription (TPRT) to integrate new TE copies in the genome [89, 143].

Class II TEs are known as DNA transposons. A DNA transposon typically encodes a single transposase gene [191], which is flanked by terminal inverted repeats (TIRs). The transposase gene in autonomous DNA transposons is responsible for TE excision from its original genome location and insertion into a new location. The majority of DNA transposons that include TIRs use a "cut-and-paste" transposition mechanism [204] while some others that lack TIRs (like Helitrons) use a different transposition mechanism [107]. The "cut-and-paste" transposition starts with DNA transposons folding inward so that TIRs on both sides of the transposase gene would bind together and dimerize. The TE is then cleaved and excised from the original location then integrated into a new target site, creating a target site duplication that flanks the inserted element. Although DNA transposons use a "cut-and-paste" transposition mechanism, the copy number of TEs can still increase because the donor sites can be regained from homologs or sister chromatids.

1.2.2 TEs in *Drosophila melanogaster*

Most TEs in the human genome are inactive or extinct (except for the L1 elements). However, a significant portion of TEs in *D. melanogaster* genome are full length and potentially active [38, 106, 19], which makes it an excellent model organism to study TE transposition activity and regulation. LTR retrotransposon is the most abundant type of TE in *D. melanogaster*. It makes up ∼2.6% of the *D. melanogaster* euchromatin, which is more than the remaining TE classes combined (0.9% for LINE-like elements and 0.3% for TIR elements) [106]. A total of 128 TE families (60 families of LTR retrotransposons, 41 non-LTR retrotransposons, and 24 DNA transposons) are present in *D. melanogaster* with over 5,390 instances in the reference genome [29]. The majority of TEs are found in the non-coding regions [106, 115], which could be explained by negative selection against TE insertions in functional coding regions [106, 10, 60]. The distribution and dynamics of TEs could be different between lab strains and natural populations [106]. For example, the *P* element is absent from many lab strains, but it is abundant in natural populations due to an invasion in the mid-20th

century [113, 8]. Unlike species such as *Saccharomyces cerevisiae* that have a large portion of LTR retrotransposon as solo-LTRs, only a small portion of LTR retrotransposons in *D. melanogaster* are solo-LTRs [106].

TE insertions are also highly polymorphic among individual flies [52], meaning that a large amount of TEs insertions are not fixed in the population. Large number of *de novo* TE insertions were previously reported in the deep-sequencing of inbred fly lines from the *Drosophila* Genetic Reference Panel (DGRP) [139, 149, 60, 259]. The low frequency of *Drosophila* TEs can be explained by multiple non-mutually exclusive models including purifying selection model [84, 115, 60, 35], transposition burst model [27, 35], and ectopic recombination model [170, 186]. Specialized tools have been developed to study population dynamics of TE in *D. melanogaster* by identifying presence and absence of known reference TEs and estimating their allele frequencies [77, 76].

TEs in *D. melanogaster* germline cells are tightly regulated by piRNA-mediated pathway [234, 41, 221]. piRNAs are derived from TE copies that accumulated in heterochromatic regions called piRNA clusters [41]. In contrast, the retrotransposons in *D. melanogaster* somatic cells are regulated by siRNA-mediated pathway [221]. siRNAs are generated from Dcr2 cleavage of long dsRNA precursors derived from retrotransposons [206]. These cellular pathways together with population genetic forces could determine the overall copy number of TEs in the *Drosophila* genome.

TEs are major sources of genetic variation in *D. melanogaster* [19] and could impact genome evolution. Even though a majority of TEs are considered to be either deleterious or neutral [35], some insertions can be adaptive [115, 35]. The adaptive mechanism of TE insertions could involve inactivating or duplicating genes, adding or removing regulatory regions, introducing alternative splicing, and affecting nearby gene expression. TE insertions were also detected in the mature transcript of genes in the sequenced *D. melanogaster* genome from wild populations [140], suggesting that TE sequences could contribute to genetic novelty and gene structure.

1.2.3 TE activity in *Drosophila* cell culture

Previous studies have shown that TEs are amplified in the *Drosophila* cell culture relative to whole flies [189, 100, 193]. The increase of TE abundance in the *Drosophila* cell culture was first detected by Potter *et al.* 1979 and Ilyin *et al.* 1980 for multiple TE families using in-situ hybridization approach. Recent study using next-generation sequencing (NGS) technique detected between ∼800 to ∼3000 non-reference TE insertions among *Drosophila* cell lines, with LTR retrotransposons making up the bulk of these new insertions [193]. In comparison, only ∼200 to ∼1400 non-reference TE insertions can be detected among *Drosophila* lab strains [193].

Potter *et al.* 1979 [189] found that the new TEs are in different locations compared to the initial distribution of TEs in the fly strains where the cell culture was derived from, providing evidence that the amplified TEs are due to new transposition instead of tandem duplication in cell culture. The transposition-based TE amplification in cell culture could be explained by a burst of transposition during the initial establishment of cell lines, by ongoing TE insertions during long-term cell culture, or a combination of both processes [70].

Potter *et al.* 1979 suggested that the TE transposition rate in *Drosophila* cell culture is uniform, favoring the ongoing TE transposition hypothesis. Later work using Southern blotting suggested that TE copy number remains relatively stable in long-term Kc cell cultures [105]. In contrast, TE copy number was observed to be amplified during the cell line establishment [105, 63], which favors the hypothesis that TE amplification occurred during the initial period of the culture. Di Franco *et al.* 1992 contrasted the stability of TE profiles among sub-lines of one of the oldest *Drosophila* cell lines (Kc) [105] with elevated TE abundance in a newly-established cell line (inb-c). The authors concluded that the increased TE abundance in *Drosophila* cell lines resulted from a transposition burst during the establishment of a new cell line, with relative stasis thereafter. However, comparing old and new cultures from different cell lines is not a definitive test of whether ongoing TE proliferation occurs during routine culture because of differences in the founder genotypes and cell

type of independently established cell lines. More recently, Sytnikova *et al.* 2014 provided evidence for transposition after initial cell line establishment in *Drosophila* by showing an increase in abundance of the *ZAM* element in a continuously cultured sub-line of the OSS cell line (OSS_C) relative to a putative frozen progenitor sub-line (OSS_E) [231]. This observation suggests that the TE transposition activity may not follow the early burst model. However, the data from Sytnikova *et al.* 2014 also suggests that the OSS_E sub-line shares more insertion with OSC lines than the OSS_C sub-line, requiring more investigation and caution for data interpretation (the current understanding is that OSS_E sub-line is actually a mislabeled version of the OSC line, see details in Chapter 2). Overall, there is a lack of understanding on whether the amplification of TEs in the long-term *Drosophila* cell culture is due to the initial burst of TEs during cell line establishment or ongoing transposition. If the latter hypothesis is accurate, it would profoundly impact our understanding of cell culture genome evolution and functional studies using cell lines since genomic variations among sub-lines could lead to phenotypic changes.

Previous studies also showed that TE families that are amplified in cell culture vary among *Drosophila* cell lines and even sub-lines of the same cell line [70, 193], leading to unique TE landscapes in different *Drosophila* cell lines and sub-lines [193]. In addition, Sytnikova *et al.* 2014 [231] showed that sub-lines of the same cell line often share a higher proportion of TE insertions relative to distinct cell lines and that some sub-lines of the same cell line might have different TE landscape. These observations suggests that TE insertions can potentially be used as markers to discriminate different cell lines established from distinct *D. melanogaster* donor genotypes (e.g., S2 vs. Kc cells) and possibly also from the same donor genotype, including divergent sub-lines of the same cell line (e.g., S2 vs. S2R+ cells) [69, 213, 245]. Also, the TE insertion profiles can potentially be used to understand the evolutionary relationship between different *Drosophila* cell lines or sub-lines.

1.3 Detecting transposable element in *Drosophila* genome

As mentioned in Section 1.2.2, TE insertions are highly polymorphic among individual flies [52]. The increased somatic transposition activity in *Drosophila* cell culture should create more polymorphic loci that are not fixed in samples of cultured cells and are not present in the reference genome (so called "non-reference" TEs). In order to comprehensively understand TE distribution and transposition activity in *Drosophila* cell culture genome and test whether polymorphic TEs provide sufficient signals for *Drosophila* cell line authentication, it is crucial to accurately detect non-reference TEs in cultured cell lines.

Traditionally, TE content in *Drosophila* genome can be analyzed using experimental approaches like Southern blotting [34], in-situ hybridization [52, 32, 33, 31] and PCR [170, 52]. These traditional approaches can only indirectly estimate the content of a particular TE sequence [20]. In addition, techniques such as in-situ hybridization can not provide a high resolution map of the insertion sites and may miss TEs in partial forms. The development of NGS technologies enables detection of TEs in high sensitivity, specificity and resolution [73].

In the NGS era, there are four main strategies for discovering and annotating TEs in a given sample based on whether a whole-genome assembly (WGA) of the sample and a TE library are being used. WGA refers to the computational process in which an organism's chromosomes are reconstructed from fragmented DNA reads. The TE library includes consensus sequences of the active TE copies, often generated from a multiple sequence alignment (MSA) of TE copies from the same TE family with slight variations due to accumulated mutations. Some commonly used databases storing TE consensus sequences in eukaryotic genomes include RepBase [15] and Dfam [242].

The first strategy uses sample's WGA and a TE library, in which direct annotation of known TE families can be done on the WGA. One of the most widely used tools is RepeatMasker (`http://www.repeatmasker.org/`), which screens any input DNA sequences for interspersed repeats and low complexity DNA sequences. RepeatMasker is often used

in combination with RepBase and Dfam to search for TE sequences that are homologous to ones present in the database. It can also be used with a custom TE library with a more species-specific focus such as for *D. melanogaster* (https://github.com/bergmanlab/transposons). In addition, tools such as Assemblytics [173] and SVMU [51] can be used to specifically detect non-reference insertions that are present in the sample genome but not in the reference genome. The TE annotation performance using this strategy is dependent on the availability and quality of the WGA and the sensitivity and specificity of the TE library.

The second strategy uses sample's WGA but does not use a TE library, in which *de novo* TE discovery using WGA can be done using tools such as RECON [16], RepeatScout [190], RepeatModeler (http://www.repeatmasker.org/RepeatModeler/), and RepeatModeler2 [78]. Similar to the first strategy, comparison between sample's WGA and a reference genome assembly allows the detection of non-reference insertions.

The third strategy uses a TE library but does not use the sample's WGA. An alignment-based approach that involves aligning raw reads to the reference genome is typically used in this strategy to detect non-reference TE insertion candidates. This strategy is commonly adopted in many short-read- (Table 1.1) and long-read- (Table 1.2) based TE detection tools.

The fourth strategy does not rely on the target sample's WGA or a TE library, in which *de novo* TE discovery can be made directly from raw reads using tools such as RepeatExplorer [178] and dnaPipeTE [85] on short-read data and RepLong on long-read data [87]. This strategy allows the possibility of annotating and studying TEs when a genome lacks reliable corresponding reference genome assemblies.

D. melanogaster has one of the best sequenced, assembled, and annotated eukaryotic genomes [160] and has high quality curated TE library (https://github.com/bergmanlab/transposons), which means the third strategy based on raw reads and a TE library can be adopted for non-reference TE detection. Thus, here I only focus on methods that can detect non-reference TEs using WGS and long-read data with a TE library and a reference genome for *D. melanogaster*.

Of all genome variants, TEs are one of the most challenging variants to detect from NGS data due to its repetitive nature, which would often result in ambiguous read alignment for reads that come from TE loci [73]. Due to challenges and the need to accurately detect TEs from NGS data, many methods have been developed over the years [73, 175] (Table 1.1). However, gold standard methods are yet to be established, and detecting non-reference TE insertions in NGS data remains an active research area. The development of long-read sequencing technologies such as Pacbio and Nanopore probably alleviates some of the problems with ambiguous short read alignment to some extent since the raw reads could be long enough to cover the entire TE breakpoint. Because the computational workflows differ, I will separately discuss short-read-based and long-read-based TE detection approaches using unassembled NGS data.

1.3.1 Detection of non-reference TEs from short read WGS data

There are two main approaches for detecting non-reference TEs using WGS data, the first approach is split-read-based and the second is discordant-read-pair-based [19, 73, 175]. The current TE detection tools usually include either of the two approaches or use a combination of both approaches in their workflows (Table 1.1). Important features from all short-read-based non-reference TE detection tools have been summarized in Table 1.1.

The split-read-based approach utilizes reads that span the junctions between 5' or 3' end of non-reference TEs and unique flanking regions in the sample genome. A portion of a junction read can be mapped to the unique region in the reference genome that flanks the insertion breakpoint, while the remaining portion can be mapped to the TE library (Figure 1.1A). Typically, there should be clusters of 5' and 3' junction reads to serve as TE-supporting reads for a non-reference TE insertion locus. In addition, clusters of junction reads on 5' and 3' end should overlap, and the maximum overlap size should approximate the size of the TSD. This is usually the way to detect the precise breakpoint of insertion and thus is important to be implemented in the TE detection algorithm. Several factors need to

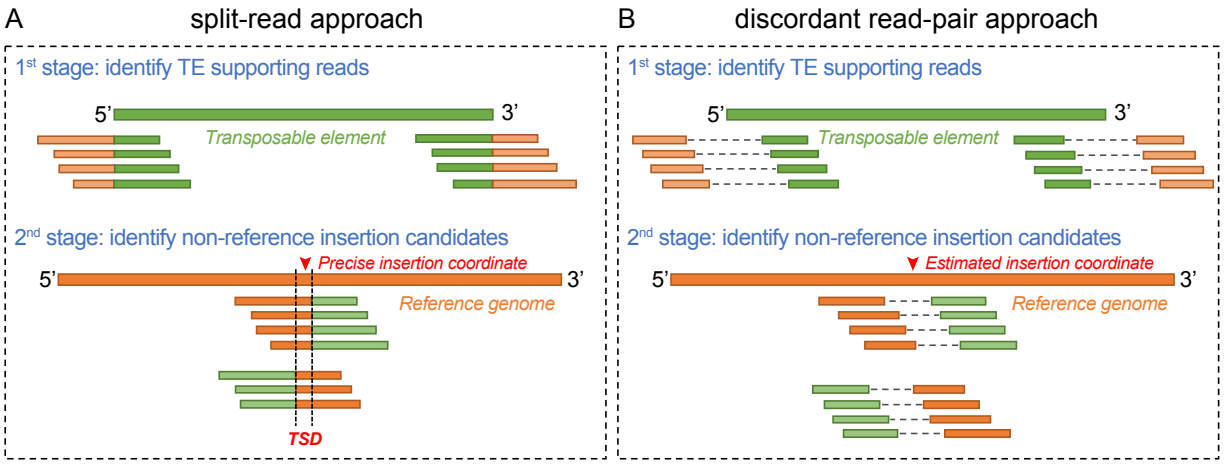

Figure 1.1: **Main approaches for detecting non-reference TEs using WGS data.** Left panel: The split-read-based approach identifies insertion signal by finding reads that partially map to unique region of reference genome and partially map to TE consensus sequence. Non-reference TE insertion candidate locus can be identified if alignments of TE-supporting reads on 5' and 3' end of TE are nearby each other. The small overlap between 5' and 3' TE-supporting reads usually represent the TSD sequence, which can be used to infer the precise insertion coordinate. Right panel: The discordant-read-pair-based approach identifies insertion signal by finding read pairs in which one read can map to unique region of reference genome while its mate maps to TE consensus sequence. This approach doesn't require split-mapping of a single read. However, it doesn't allow the detection of TSD and precise insertion coordinate.

be considered when using a split-read-based approach for TE detection. First, the aligners should be able to partially align reads and generate a recognizable signature like CIGAR strings that TE detection tools can parse. Most short-read-based TE detection tools use BWA MEM [133] or bowtie [125], both of which can report split-mapped reads in CIGAR string format. Second, the read length affects the sensitivity of this strategy. If the read size is too small, the aligner may not confidently map different segments of a junction read to the reference genome or TE, leading to ambiguous or false predictions.

The discordant-read-pair-based approach utilizes mate-paired short reads in which one read maps to 5' or 3' end of non-reference TEs while its mate mapping to unique flanking

regions in the sample genome (Figure 1.1B). This approach does not require the aligners to report split-mapped reads. However, detection of the TSD sequence and getting precise insertion coordinates are difficult using discordant-read-pair-based approach due to the lack of junction-read information. In some cases the insertion coordinate is estimated approximately based on the median insert size [115].

Certain limitations need to be considered when using short-read-based non-reference TE detection tools. First, the current methods skew toward more recent and non-nested TEs that are in normal recombination regions with low overall repetitiveness [19]. Second, each tool requires a unique set of software dependencies, has different and often non-standard input and output format, and often lacks evaluations that are fair to other systems [73, 175]. Given that different species have wildly diverse TE structures and spectrum, it is important to use TE detection tools that perform reliably on the genome of interested species. Third, identifying non-reference TE insertions from NGS data is computationally challenging, mainly because the short read size limits the ability to resolve the multi-mapping situation and identify exact insertion or deletion breakpoints, resulting in a considerable amount of false-positive and false-negative predictions. Ewing *et al.* 2015 and Nelson *et al.* 2017 showed that a major portion of TE predictions is program-specific among several somatic insertion detection methods [73]. Rahman *et al.* 2015 [193] used PCR to validate TE predictions from TIDAL and revealed substantial false-positive rates. A previous study that benchmark TE prediction tools on rice genome revealed precision on non-reference insertions to be around 80-90% [235]. Finally, the size of short reads is generally much smaller than a typical full-length TE insertion in *D. melanogaster* genome (6kb+), limiting the ability to reconstruct non-reference TE insertion sequence [5, 188].

1.3.2 Detection of non-reference TEs from long-read sequencing data

As mentioned in the previous section, short-read-based methods are cost-efficient for detecting non-reference TEs from WGS data. However, the nature of short read means

Table 1.1: **Software for detecting non-reference TEs using WGS data.** TAF refers to whether the software reports TE allele frequency in the output file. Split-read and Read-pair refer to the approaches used to predict non-reference TEs (see details in the main text). Finally, Conda refers to whether the software is made available as a package in the conda package management system.

Name of the tool	TAF	Split-read	Read-pair	Date Published	Language	Conda
PoPoolationTE	Yes	No	Yes	1/26/12	Perl	No
ngs_te_mapper	No	Yes	No	2/9/12	R	No
TE-Locate	No	No	Yes	9/12/12	Perl	No
Retroseq	Yes	No	Yes	2/1/13	Perl	No
RelocaTE	No	Yes	No	6/1/13	Perl	No
Grasper	No	No	Yes	3/7/14	Java	No
TIF	No	Yes	No	3/14/14	Perl	No
TEMP	Yes	Yes	Yes	4/21/14	Perl	No
Mobster	Yes	Yes	Yes	10/28/14	Java	Yes
TE-Tracker	No	No	Yes	11/19/14	Perl	No
Discord-retro	No	Yes	Yes	1/18/15	Python	No
Tangram	No	Yes	Yes	2/12/15	C/C++	No
ITIS	Yes	Yes	Yes	3/5/15	Perl/R	No
Jitterbug	Yes	No	Yes	10/12/15	Python	No
TIDAL	No	Yes	No	12/15/15	C/Perl/R/Bash	No
SPLITREADER	Yes	Yes	No	6/3/16	Bash	No
PoPoolationTE2	Yes	No	Yes	8/2/16	Java	No
TEPID	Yes	Yes	Yes	12/2/16	Python	Yes
RelocaTE2	No	Yes	Yes	1/26/17	Python	Yes
TEFLON	Yes	No	Yes	5/1/17	Python	No
STEAK	No	Yes	Yes	8/21/17	C++	No
MELT	Yes	Yes	Yes	8/30/17	Java	No
TEbreak	Yes	Yes	Yes	5/1/18	Python	No
TranSurVeyor	No	Yes	Yes	11/16/18	C++	No
iMGEins	Yes	Yes	Yes	12/18/18	Java	No
Trackposon	Yes	No	Yes	1/3/19	Bash/R/Perl	No
ERVcaller	No	Yes	Yes	3/21/19	Perl	No
MGEfinder	Yes	Yes	No	12/17/19	Python	No
TEMP2	Yes	Yes	Yes	1/28/21	Perl	No
TEFinder	Yes	No	Yes	1/31/21	Bash	No
xTEA	Yes	Yes	Yes	6/22/21	Python	No

it is not sensitive for detecting full-length or complex TE insertions in *D. melanogaster*, the size of which are typically much longer than what a single short read can cover. In

comparison, long-read sequencing technologies such as Pacific Biosciences (PacBio) [158] and Oxford Nanopore [11] can produce longer reads (≥15kbp) that may cover the entire TE insertions (typically less than 10kbp) plus their associated flanking genomic sequences in the target genome. Long reads enable more accurate read alignment and assembly, thus provide an opportunity for developing approaches that have higher sensitivity for non-reference TE detection [68, 188, 167]. Indeed, recent studies in *Drosophila* have shown that 38% of TE insertions in *D. melanogaster* chromosome arm 2L in A4 relative to ISO1 detected from long-read-based approach are missed by short-read-based approach [51].

Currently, there are two main approaches for detecting non-reference TEs from long-read data. The first approach is based on *de novo* WGA. WGA-based SV detection has been demonstrated on an inbred *D. melanogaster* strain [51] and in human [4, 101, 232]. Tools such as wtdbg2 [203] and Flye [117] can be used to produce *de novo* WGAs from long-read data, the WGAs of sample genome and a reference genome can then be used as input to specialized tools such as Assemblytics [173] and SVMU [51] to generate non-reference insertion candidate based on WGA-to-reference comparison. However, identifying non-reference TEs in diploid heterozygous genome using the WGA-based approach requires not only contiguous but haplotype-resolved and complete representation of the sample genome [151], which makes it impractical for polyploid heterozygous genomes like *Drosophila* cell culture.

The other approach is based on alignment of raw reads to the reference genome. The insertion signal is generally picked up based on read split-mapping and abnormal coverage [216, 151]. It is worth noting that the discordant-read-pair-based approach used in short-read data does not apply in long-read data because long reads are single-ended. In addition, the high sequencing error rate of long reads need to be taken into consideration using this alignment-based approach since it will affect both the alignment and SV calling steps. Specialized methods have been developed to align long reads such as BLASR [49], Minimap2 [136], and NGMLR [216]. The identification of SVs from long-read sequencing data can be done using methods such as Sniffles [216], SVIM [92], pbsv (`https://github.com/`

PacificBiosciences/pbsv), cuteSV [104], and PBHoney [72]. Sniffles is one of the well-recognized methods in this field, which can report low-frequency SVs and should be useful for detecting polymorphic and mosaic variations like non-reference TEs in heterozygous genomes.

Another important feature of long-read TE detection is the ability to output partial or complete TE sequences. This can be done by 1) extracting soft-clipped sequences from TE-supporting reads [68], 2) from WGA [167], or 3) from localized *de novo* assembly after SV candidate is identified using an alignment-based approach [151, 161, 58]. The third option should be particularly useful when the sample's WGA is not available. The local contig assembly may also resolve complex TE insertion events given that the assembled contig size is much longer than a single read [151].

There are seven methods currently available that can detect non-reference TEs using long-read data including LoRTE [68], TLDR [74], PALMER [258, 161], xTea [58], TrEMOLO [167], TrEMOLO [167], rMETL [103], and the workflow developed and used in Siudeja *et al.* 2021. A brief review on each of the seven methods is provided as follows (Table 1.2).

LoRTE [68] is one of the earliest tools that are capable of detecting polymorphic non-reference TEs using Pacbio data. Its TE detection workflow includes two major stages. In the first stage, sequences that flank each annotated TE on the 5' and 3' sides in the reference genome are extracted and aligned to raw reads. LoRTE then labels a reference TE insertion as "present" if both 5' and 3' flanks are uniquely mapped and a TE sequence can be found in between both flanking sequence alignments. In the second stage, all reference TE sequences identified in the first stage are removed in the raw reads. LoRTE then aligns TE consensus sequences from a TE library to raw reads to identify non-reference TE insertions. If a read can be aligned to the TE library, sequences that flank the aligned portion in the read are extracted and aligned to the reference genome. If the gaps between two flanking sequence alignments are less than 50bp, LoRTE then labels the gap region as candidate locus for a non-reference TE insertion. Finally, all reads that support the same insertion event are

Table 1.2: **Software for detecting non-reference TEs using long-read data.** TE sequence refers to whether the software can report non-reference TE insertion sequence and, if so, how the sequence is generated. TAF refers to whether the software reports TE allele frequency in the output file. Finally, Conda refers to whether the software is made available as a package in the conda package management system.

Name of the tool	Data Type	Predict TSD	TE sequence	TAF	Genotype	Conda	Language	PMID
LoRTE	Pacbio/Nanopore	No	Clipped sequence from one read	No	Yes	No	Perl	28405230
PALMER	Pacbio/Nanopore	Yes	Local assembly on corrected reads	No	No	No	C++	31853540
TLDR	Pacbio/Nanopore	Yes	Consensus	No	No	No	Python	33186547
xTEA	Pacbio/Nanopore	Yes	Unpolished local assembly	No	No	Yes	Python	34158502
rMETL	Pacbio/Nanopore	No	No	No	Yes	Yes	Python	30759188
TrEMOLO	Pacbio/Nanopore	Yes	from WGA	Yes	No	No	Python	32722451
somatic-transposition-fly-intestine	Nanopore	Yes	No	No	No	No	Python	33634906

clustered for final insertion calls. LoRTE has been demonstrated to work on *Drosophila* datasets. However, it cannot predict TSD sequences, estimate TAF, and provide assembled TE sequences that are essential features for studying TE dynamics and sequence evolution. In addition, LoRTE performs the read alignment step on a single read basis using NCBI BLAST+ [44], making the workflow inefficient for high coverage datasets.

TLDR [74] was designed to detect non-reference TE insertions and resolve full-length insertion sequences from Nanopore data. The TE insertion signal is detected using a split-read-based approach. The soft-clipped sequences from split-mapped reads are aligned to a TE library to identify TE insertion candidate loci. Importantly, TLDR can reconstruct TE insertion sequence from the consensus of multiple sequence alignment (MSA) of TE-supporting reads. The consensus sequence is then refined through realignment of TE-supporting reads. TLDR is also able to predict TSD and determine precise insertion coordinate. Unlike LoRTE, TLDR is designed to focus on detecting TEs in the human genome and has not been tested on *Drosophila* datasets. Also, TLDR does not provide TAF for the insertion.

PALMER [258, 161] was designed to detect non-reference L1 insertions in the human genome from Pacbio and Nanopore reads. PALMER has a similar computational workflow compared to LoRTE but the former method includes more features. The raw-reads-to-reference alignment and reference TE annotation are used as input to the PALMER workflow. It first masks out reference TE sequences in the reads that can map to annotated TEs in the reference genome. The remaining reads are aligned with a TE library to identify putative insertion sequences. PALMER then searches hallmarks of human mobile elements such as TSDs, transductions, and poly(A) tract sequences in the putative insertion sequences. Finally, PALMER clusters TE-supporting reads for the final call, in which a minimum number of supporting reads is required. Importantly, PALMER can report the TSD sequence and reconstruct the TE consensus sequence from clustered and corrected raw reads. Like TLDR, PALMER has not been tested on *Drosophila* datasets, and it does not provide TAF for the insertion.

xTea [58] was designed to identify TE insertions using data from multiple sequencing technologies. The long-read mode of xTea detects non-reference TE insertion signals using a split-read-based approach. For each TE candidate locus, xTea does local assembly on all clipped reads around the estimated insertion site using wtdbg2 [203]. The 5' and 3' flanking sequences around the estimated insertion site in the reference genome are then aligned to the assembled local contig to identify insertion sequences. Finally, each insertion sequence is aligned to a TE library to annotate family information. The assembled local contig is also realigned to the reference genome to refine the insertion site. xTea is designed to detect TEs in human and has not been tested on *Drosophila* datasets. Also, xTea does not provide polished assembly and estimate TAF for a predicted TE insertion.

TrEMOLO [167] uses both WGA-based and alignment-based approaches for non-reference TE detection using long-read data. The WGA-based module in TrEMOLO uses Assemblytics [173] to detect non-reference insertions by comparing sample's WGA with a reference genome assembly. The insertion sequences identified by Assemblytics are then aligned to a TE library for identiying TE insertion candidates. The alignment-based module in TrEMOLO then aligns raw reads to sample's WGA using minimap2 [136] and detects SVs using Sniffles [216]. Long insertion sequences detected by Sniffles are then aligned to a TE library for detecting minor/heterozygous TE insertions. Importantly, TrEMOLO can estimate TAF but it does not assemble the minor or heterozygous TE insertions.

rMETL [103] detects insertion signals using a split-read-based approach and clusters split-mapped reads to identify insertion candidates. The clustered reads are then realigned to a TE library for family annotation and final insertion call. rMETL is optimized to detect TEs in human such as Alu, L1, and SVA elements and it has not been tested on *Drosophila* datasets. Also, this tool does not estimate TAF or provide assembled sequences for each TE insertion.

Finally, Siudeja *et al*, 2021 developed a workflow for detecting rare somatic transposition events in the fly intestine (`https://github.com/bardin-lab/somatic-transposition-fly-intestine`)

[1]. This workflow can detect somatic insertions that are fully contained within singleton reads and can predict TSD sequence. The workflow has not been generalized to work for other datasets in *Drosophila*. Also, it does not estimate TAF or provide assembled TE sequences from reads that jointly support the insertion event.

All seven long-read-based TE detection methods mentioned above were developed with specific research purposes (e.g., studying L1 elements in the human genome or somatic TEs in *Drosophila* whole flies). As a result, different methods provide different sets of information regarding the TE insertion event. Key features of all long-read-based TE detection methods have been summarized in Table 1.2. In general, each tool may output one of three types of TE sequences: (1) a soft-clipped sequence from one representative TE-supporting read (LoRTE), (2) a consensus sequence from all reads that support a given insertion event (TLDR), or (3) a TE sequence resulting from local-assembly of TE-supporting reads (xTea, PALMER). Upon investigation, none of the currently available methods meet all following criteria: 1) is compatible on *Drosophila* data or is species-agnostic, 2) can predict TSD sequence, 3) provides assembled and polished TE sequence instead of soft-clipped sequences extracted from one TE-supporting read, and 4) can estimate TAF. Thus, a generalized tool is needed to meet all criteria described above to study the dynamics and sequence evolution of TEs in complex heterozygous genomes like cell lines.

Some factors need to be considered when using or developing a long-read-based non-reference TE detection approach, including 1) the sequence divergence between genomes of the sample and reference. All currently available methods rely on a reference genome. High divergence between sample and reference genomes could result in low TE prediction performance using both WGA or an alignment-based approach; 2) the quality of the TE library. All currently available methods rely on a TE library, which means the detection sensitivity depends on whether the TE consensus sequences in the library can incorporate the TE divergence in the sample genome; 3) method evaluation. Currently, there is no comprehensive evaluation framework that can benchmark the performance of non-reference TE

detection tools on heterozygous *Drosophila* data, which needs to be established; and 4) the high per-base error rate of long reads. To achieve better sequence quality, strategies using polished assembly or consensus (PALMER, TLDR) are generally better than using unpolished assembly (xTea) or clipped sequences from a single representative read (LoRTE).

1.4 Dissertation Outline

The overall theme of this dissertation is to develop novel strategies to detect non-reference transposable elements (TEs) in order to understand the dynamics and evolution of TEs in *Drosophila* cell culture genomes.

In chapter 2, I leveraged the classical observation that TEs proliferate in cultured *Drosophila* cells to demonstrate that genome-wide TE insertion profiles can reveal the identity and provenance of *Drosophila* cell lines. I identified multiple cases where TE profiles clarify the origin of *Drosophila* cell lines (Sg4, mbn2, and OSS_E) relative to published reports and provide evidence that insertions from only a subset of LTR retrotransposon families are necessary to mark *Drosophila* cell line identity. I also developed a new bioinformatics approach to detect TE insertions and estimate intra-sample allele frequencies in legacy WGS data (called ngs_te_mapper2), which revealed loss of heterozygosity as a mechanism shaping the unique TE profiles that identify *Drosophila* cell lines. This work contributes to the general understanding of the forces impacting metazoan genomes as they evolve in cell culture and paves the way for high-throughput protocols that use TE insertions to authenticate cell lines in *Drosophila* and other organisms (see details on the protocol development in Appendix).

In chapter 3, I generated genome-wide TE profiles for 29 sub-lines of S2 cells from WGS data to understand whether TE amplification in cell culture is due to an initial burst of transposition after cell line establishment or ongoing transposition during routine cell culture. I showed that TE insertions provide abundant markers to reconstruct the evolutionary history of S2 sub-lines, and that the major phylogenetic relationships among S2 sub-lines

inferred from TE insertions correlate with genome-wide copy number differences. The evolutionary history of S2 built from TE profiles show that publicly available S2 sub-lines form one monophyletic group including two major subgroups (A and B), and reveal no evidence for widespread cross-contamination of available S2 cultures by other *Drosophila* cell lines. Using ancestral state reconstruction, I inferred that TE insertion has occurred on all internal branches of the S2 phylogeny, but that only a small subset of *D. melanogaster* TE families have proliferated during S2 evolution, most of which are retrotransposons that do not encode a retroviral *envelope* gene. Together, these results support the conclusions that TE insertions provide useful markers of S2 sub-line identity and genome organization and that TE proliferation in *Drosophila* somatic cell culture is an ongoing, cell-autonomous process that does not result from ubiquitous deregulation of global transpositional control mechanisms.

In chapter 4, I used WGS data for the tetraploid *Drosophila* S2R+ cell line from long-read and linked-read technologies to better understand the pattern and process of TE proliferation in *Drosophila* cell culture. WGAs of S2R+ from long- or linked-read data were highly fragmented relative to assembly of similar data from wild-type flies and generated variable estimates of TE content. In order to study TE sequence evolution without depending on WGAs, I developed a novel bioinformatics tool called "TELR" that identifies, locally assembles, and estimates allele frequency of TEs from long-read data (https://github.com/bergmanlab/telr). Application of TELR to a PacBio dataset for S2R+ revealed many haplotype-specific TE insertions that come from somatic transpositions after tetraploidization of the S2 genome after initial establishment of S2 cell line. Local assemblies from TELR also allowed phylogenetic analysis of paralogous TE copies within the S2R+ genome, which revealed that proliferation of different TE families in *Drosophila* cell lines could be driven by single or multiple source lineages. Overall, this work provides a model for the analysis of TEs in complex heterozygous or polyploid genomes that are not amenable to WGA and yields new insights into the mechanisms of TE sequence evolution in animal cell culture.

Chapter 2

Transposable element profiles reveal cell line identity and loss of heterozygosity in *Drosophila* cell culture[1]

[1]Shunhua Han, Preston J Basting, Guilherme B Dias, Arthur Luhur, Andrew C Zelhof, and Casey M Bergman, Transposable element profiles reveal cell line identity and loss of heterozygosity in Drosophila cell culture, *Genetics*, 2021;, iyab113.

2.1 ABSTRACT

Cell culture systems allow key insights into biological mechanisms yet suffer from irreproducible outcomes in part because of cross-contamination or mislabelling of cell lines. Cell line misidentification can be mitigated by the use of genotyping protocols, which have been developed for human cell lines but are lacking for many important model species. Here we leverage the classical observation that transposable elements (TEs) proliferate in cultured *Drosophila* cells to demonstrate that genome-wide TE insertion profiles can reveal the identity and provenance of *Drosophila* cell lines. We identify multiple cases where TE profiles clarify the origin of *Drosophila* cell lines (Sg4, mbn2, and OSS_E) relative to published reports, and also provide evidence that insertions from only a subset of LTR retrotransposon families are necessary to mark *Drosophila* cell line identity. We also develop a new bioinformatics approach to detect TE insertions and estimate intra-sample allele frequencies in legacy whole-genome sequencing data (called "ngs_te_mapper2"), which revealed loss of heterozygosity as a mechanism shaping the unique TE profiles that identify *Drosophila* cell lines. Our work contributes to the general understanding of the forces impacting metazoan genomes as they evolve in cell culture and paves the way for high-throughput protocols that use TE insertions to authenticate cell lines in *Drosophila* and other organisms.

2.2 Introduction

Cultured cell lines play essential roles in biological research, providing model systems to support discovery of basic molecular mechanisms and tools to produce biomolecules with medical and industrial relevance. Despite their widespread use, experiments in cultured cells often show non-reproducible outcomes, and increasing the rigor of cell-line based research is a priority of both funders and journals alike [142]. One major source of irreproducible research comes from mislabelling or cross-contamination of cell lines (collectively referred to here as "misidentification"), resulting in cells of the wrong type or species being used in

a particular study [62, 80, 176, 150, 97]. As such, substantial effort has been invested into minimizing cell line misidentification through genotyping cell lines, cataloguing misidentified lines, standardizing cell line nomenclature, and the use of research resource identifiers [157, 46, 17, 248, 12].

Starting with the first reports on the cell line misidentification problem, a variety of cytological and molecular techniques have been developed to authenticate mammalian cell lines [62, 80, 179, 83, 157, 47]. These efforts culminated in development of short tandem repeats (STRs) as a widely-used standard to authenticate human cell lines at the molecular level [157, 17, 6]. STR-based authentication has mitigated – but not eradicated – the human cell line misidentification problem, in part because of limitations in the stability, measurement, and matching of STRs [183, 7, 248, 94]. More recently, alternative methods for genotyping human cell lines based on single nucleotide polymorphisms (SNPs) have been developed [47, 248, 138, 250, 168], but these methods have not yet been accepted as standards for cell line authentication in humans [6].

For most species beside humans, cell line authentication standards and protocols remain to be established [6]. For example, no protocols currently exist to authenticate cell lines in the fruitfly *Drosophila melanogaster*, despite the existence of over 150 different cell lines for this model animal system [144]. As such, no evidence of misidentified *Drosophila* cell lines have been catalogued to date by the International Cell Line Authentication Committee (v10, `https://iclac.org/databases/cross-contaminations/`). Development of cell line identification protocols and standards for common model organisms like *Drosophila* is an important goal for increasing rigor and reproducibility in bioscience. Achieving this goal for a new species requires an understanding of the genome biology and cell line diversity of that organism, and should ideally take advantage of powerful, cost-effective modern genomic technologies.

Relative to humans, the STR mutation rate is low in *D. melanogaster* [214] and thus the use of STRs for discriminating different *Drosophila* cell lines is likely to be limited. In con-

trast, it is well-established that transposable element (TE) insertions are highly polymorphic among individual flies [52], that TE abundance is elevated in *Drosophila* cell lines relative to whole flies [189, 100, 193], and that TE families amplified in cell culture vary among *Drosophila* cell lines [70, 193]. These properties suggest that TE insertions should be useful markers to discriminate different cell lines established from distinct *D. melanogaster* donor genotypes (e.g., S2 vs Kc cells) and possibly also from the same donor genotype, including divergent sub-lines of the same cell line (e.g., S2 vs S2R+ cells) [69, 213, 245]. Indeed, previous studies have shown that *D. melanogaster* cell lines have unique TE landscapes, and that sub-lines of the same cell line often share a higher proportion of TE insertions relative to distinct cell lines [231, 193].

Here we show that *Drosophila* cell lines can successfully be clustered and identified on the basis of their genome-wide TE profiles using a combination of publicly available paired-end short-read whole genome sequencing (WGS) data from the modENCODE project [129] and new WGS data for eight widely-used *Drosophila* cell lines. Our approach reveals the first examples where the reported provenance of *Drosophila* cell lines – Sg4 [171] and mbn2 [82] – conflicts with identity inferred from genomic data. Importantly, our TE-based clustering approach also allows us to identify which subset of TE families discriminate the most widely used *Drosophila* cell lines, paving the way for development of PCR-based genotyping protocols that can be used for cost-effective *Drosophila* cell line identification.

Additionally, we develop a new tool for detection of TEs in single-end WGS data (called "ngs_te_mapper2") and integrate our new data with legacy data [218, 231] to resolve the history and provenance of the widely-used OSS and OSC ovarian cell lines [177, 208]. Using TE-based clustering, we provide evidence that OSS and OSC cell lines can be discriminated on the basis of the *ZAM* retrotransposon family. We propose that the OSS_E sub-line reported in Sytnikova *et al.* 2014 [231] approximates an ancestral state of the OSC cell line, with contemporary OSC sub-lines having undergone loss of heterozgosity (LOH) in cell culture from an OSS_E-like state. Together, our results show that TE insertions are a powerful

source of genetic markers that can be used for cell line authentication in *Drosophila* and that LOH is an important mechanism driving *Drosophila* cell line genome evolution.

2.3 Materials and Methods

2.3.1 Genome sequencing

Public genome sequencing data for 26 samples of 18 *Drosophila* cell lines were obtained from the modENCODE project [129]. Frozen stocks of eight additional samples from six *Drosophila* cell lines (mbn2, Sg4, ML-DmBG3-c2, ML-DmBG2-c2, OSS, and OSC) were obtained from the *Drosophila* Genomics Resource Center (DGRC), the Gorski lab (Canada's Michael Smith Genome Sciences Centre, BC Cancer) and the Strand lab (University of Georgia). DNA extractions were performed using Qiagen Blood and Tissue kit (Cat# 69504) for the mbn2 sample from the Strand lab and using the Zymo-Quick kit (Cat# D4068) for all other samples. Purified DNA was analyzed by Qubit and Fragment Analyzer to determine the concentration and size distribution, respectively. Samples were normalized to the same concentration before preparing libraries with the KAPA Hyper Prep Kit (Cat# KK8504). During library prep, DNA was fragmented by acoustic shearing with Covaris E220 Evolution before end repair and A-tailing. Single indices were ligated to DNA fragments. Libraries were purified and cleaned with Solid Phase Reversible Immobilization (SPRI) beads before PCR amplification. Final libraries underwent an additional round of bead cleanup before being assessed by Qubit, qPCR (KAPA Library Quantification Kit Cat# KK4854), and Fragment Analyzer. Libraries were then sequenced in paired-end 150bp mode on an Illumina NextSeq500 high output flowcell and demultiplexed using bcl2fastq. Metadata, sequencing statistics, and SRA accession numbers for all cell line DNA-seq samples used in this study can be found in Table 2.4.

2.3.2 Detection of non-reference TE insertions using paired-end sequencing data

Paired-end sequencing data from the modENCODE project [129] and our study was used as input to seven methods designed to detect non-reference TE insertions in *Drosophila* [139, 115, 259, 116, 3, 249] using McClintock (revision 40863acf11052b18afb4cdcd7b1124de48cba397; options: -m "trimgalore, popoolationte, popoolationte2, temp, temp2, teflon, ngs_te_mapper, ngs_te_mapper2") [175]. Additionally, we predicted non-reference TE insertions using a version of TIDAL1.2 [193, 246] that was modified to output results in a format compatible with results from McClintock (https://github.com/bergmanlab/TIDAL, revision 2d110b17b3b287dbc1ceb67c87fe171d15095c84). The reference genome for these analyses was comprised of the major chromosome arms from the *D. melanogaster* dm6 assembly (chr2L, chr2R, chr3L, chr3R, chr4, chrM, chrY, and chrX) and the TE library was the Berkeley *Drosophila* Genome Project canonical TE dataset v10.1 (`https://github.com/bergmanlab/transposons/blob/master/releases/D_mel_transposon_sequence_set_v10.1.fa`; revision f94d53ea10b95c9da99258ac2336ce18871768e9).

Paired-end samples analyzed here vary substantially in read length (50-151bp) and depth of coverage (5X-136X) (Table 2.4). We chose not to normalize input datasets by downsampling to the lowest read length and coverage to avoid reducing sensitivity of non-reference TE detection methods for higher quality samples. Using complete samples allowed us to observe that the number of non-reference TE predictions per sample (Table 2.5) showed a strong dependence on read length (Figure 2.3) or coverage (Figure 2.4) for all methods besides TEMP [259]. Thus, we used TEMP predictions with default McClintock filtering (retain only 1p1 predictions with >0.1 intra-sample allele frequency cutoff) for the global analysis of the modENCODE-only and expanded (modENCODE plus new samples) datasets. To resolve details of the relationship among mbn2 sub-lines, we used read length and coverage normalized mbn2 samples with relaxed filtering criteria for TEMP predictions (retain all 1p1/2p/singlton predictions with no intra-sample allele frequency cutoff).

2.3.3 Detection of non-reference TE insertions using single-end sequencing data

Single-end sequencing data for OSS and OSC cell line samples from two previous studies [218, 231] and forward reads from our paired-end samples were used to predict non-reference TE insertions using ngs_te_mapper2 (https://github.com/bergmanlab/ngs_te_mapper2) in McClintock (revision 40863acf11052b18afb4cdcd7b1124de48cba397; options: -m "trimgalore, coverage, ngs_te_mapper2, map_reads") [175]. ngs_te_mapper2 is a re-implementation of the non-reference TE detection method initially reported in Linheiro *et al.* 2012 [139] that improves speed and sensitivity and has been extended to estimate TE allele frequency (see Section 2.4.1 for details). Reference genome and TE library files used for McClintock runs on single-end sequencing data were the same as used above for paired-end sequencing data. Because ngs_te_mapper2 detection rates and allele frequency estimates are sensitive to read length and depth of coverage (see Section 2.4.1), reads from single-end sequencing data and the forward read of our paired-end sequencing data were normalized by trimming all reads to 100bp using fastp (v0.20.1) [53] and downsampling to the lowest coverage sample (14X) using seqtk (v1.3) [134].

2.3.4 Classification of intra-sample TE insertion allele frequency

To predict whether TE insertions within OSS and OSC cell line samples were heterozygous or homozygous, we built a classifier that uses allele frequencies estimated by ngs_te_mapper2 from single-end sequencing data as input. A non-reference TE insertion was predicted to be heterozygous if the intra-sample allele frequency estimated by ngs_te_mapper2 is between 0.25 to 0.75 and predicted to be homozygous if the intra-sample allele frequency is greater than or equal to 0.95. TE insertions with intra-sample allele frequencies outside these ranges were considered unclassified. The classifier was benchmarked using synthetic homozygous and heterozygous WGS datasets created with wgsim (v0.3.1-r13) using the ISO1 (dm6) and A4 (GCA_003401745.1) [51] genome assemblies as input. The classifier yields >91% precision

using input from the results of ngs_te_mapper2 applied to the simulated datasets (see Section 2.4.1 for details).

2.3.5 Identification of orthologous TE insertions

Because positional resolution of non-reference TE predictions is inexact [175], we identified a high-quality set of orthologous non-reference TE insertion loci as follows. Genome-wide non-redundant BED files of non-reference TE predictions generated by McClintock were filtered to exclude TEs in low recombination regions using boundaries defined by Cridland *et al.* 2013 [60] lifted over to dm6 coordinates. Normal recombination regions included in our analyses were defined as chrX:405967–20928973, chr2L:200000–20100000, chr2R:6412495–25112477, chr3L:100000–21906900, chr3R:4774278–31974278. We restricted our analysis to normal recombination regions, since low recombination regions have high reference TE content which reduces the ability to predict non-reference TE insertions [29, 154]. We also excluded *INE-1* family from our analysis, as this family is reported to be inactive for millions of years [220, 238]. Non-reference TE predictions in high recombination from all samples were then clustered into orthologous loci using BEDtools cluster (v2.26.0) enforcing predictions within each cluster to be on the same strand (option -s) [192]. Orthologous loci were then filtered using the following criteria: 1) retain only a single TE family per locus; 2) retain only a single TE prediction per sample per locus; and 3) retain TE predictions only from long-terminal repeat (LTR) retrotransposon, LINE-like retrotransposon or DNA transposon families. For clustering of paired-end samples, we imposed the additional filtering requirement that all clusters include at least sample per locus with a TEMP 1p1 prediction.

2.3.6 Clustering and identification of cell line samples using TE insertion profiles

Non-reference TE predictions at orthologous loci were then converted to a binary presence/absence matrix in order to cluster cell lines on the basis of their TE insertion profiles.

Cell line clustering was performed using Dollo parsimony in PAUP (v4.0a168) [230]. Dollo parsimony analyses were conducted using heuristic searches with 50 replicates. A hypothetical ancestor carrying the assumed ancestral state for each locus (absence) was included as a root in the analysis [23]. "DescribeTrees chgList=yes" option was used to assign character state changes to branches in the tree. Node support for the most parsimonious tree was evaluated by integrating 100 bootstrap replicates generated by PAUP using SumTrees (v4.5.1) [229].

Identification of a cell line sample was performed by adding its TE profile to a binary presence/absence matrix of "primary replicates" of 22 non-redundant *Drosophila* cell line samples and performing cell line clustering using the same approach mentioned above. A phylogenetic tree of the 22 non-redundant primary *Drosophila* cell line samples was used as a backbone topological constraint during a heuristic searches for the most parsimonious tree that included one additional "secondary replicate". Node support for the most parsimonious tree was evaluated by integrating 100 bootstrap replicates without topological constraints.

2.3.7 B-allele frequency and copy number analysis

BAM files generated by McClintock were used for variant calling using bcftools (v1.9) [132]. Indels were excluded from variant calling, leaving only single-nucleotide polymorphisms (SNPs) in the VCF file. For a given SNP, the B-allele frequency (BAF) was determined as the coverage of reads supporting non-reference allele divided by total coverage at that position using the DP4 field.

BAM files generated by McClintock were also used to generate copy number variant (CNV) profiles for non-overlapping 10kb windows of the dm6 genome using Control-FREEC (v11.6) [37]. Windows with less than 85% mappability were excluded from the analysis based on mappability tracks generated by GEM (v1.315 beta) [65]. The baseline ploidy was determined by normalized DNA read density of 10 kb windows following the method used in Lee *et al.* 2014 [129]. The sex information was determined from relative read density

between chromosome X and autosomes. The minimum and maximum expected value of the GC content was set to be 0.3 and 0.45, respectively.

2.3.8 Clustering of cell line samples based on transcriptomes

Total RNA sequencing samples for 17 *Drosophila* cell lines with 100bp paired-end reads were obtained from Stoiber *et al.* 2016 [228] and from the modENCODE *D. melanogaster* transcriptome sequencing project [42]. SRA accession numbers for all cell line RNA-seq samples used in this analysis can be found in Table 2.6. Transcript abundances for protein-coding genes were quantified in unit of transcripts per million (TPM) using kallisto quant (v0.46.2) [40] using the release 6.32 version of the *D. melanogaster* transcript coding sequences corresponding to Ensembl genes from Ensembl release 103 (http://ftp.ensembl.org/pub/release-103/fasta/drosophila_melanogaster/cds/Drosophila_melanogaster.BDGP6.32.cds.all.fa.gz) [247]. Transcript-level abundance estimates were summarized into gene-level abundance estimates using the release 6.32 version of the *D. melanogaster* gene annotation from Ensembl release 103 (http://ftp.ensembl.org/pub/release-103/gtf/drosophila_melanogaster/Drosophila_melanogaster.BDGP6.32.103.gtf.gz) using tximport (v1.18.0) [225]. The summarized gene-level abundance matrix was log transformed and visualized using the Rtsne package (v0.15) [120] with following parameters: perplexity=1, theta=0.0, max_iter=5000, check_duplicates=FALSE.

2.4 Results and Discussion

2.4.1 Description of the ngs_te_mapper2 method for detecting non-reference TE insertions in single-end whole genome shotgun data

As is required by this study to resolve the history and provenance of the widely-used OSS and OSC ovarian cell lines [177, 208], we developed a new tool (called "ngs_te_mapper2") for detecting transposable element (TE) insertions from single-end next-generation sequencing (NGS) data and estimate intra-sample TE allele frequency. ngs_te_mapper2 (https:

//github.com/bergmanlab/ngs_te_mapper2) is a re-implementation of the method for detecting non-reference TE insertions in single-end whole genome shotgun sequence data initially reported in Linheiro *et al.* 2012 [139]. ngs_te_mapper2 uses a three-stage procedure to annotate non-reference TEs as the span of target site duplication (TSD) (Figure 2.1), following the annotation framework described in Bergman *et al.* 2012 [26]. In the first stage, whole genome shotgun (WGS) reads are mapped to a library of TE sequences to identify "junction reads" that span the start/end of TE and genomic flanking sequences are retained. Such reads are often referred as "split reads", although in reality these reads are not split in the resequenced genome.

In the second stage, junction reads from each side of TE insertion identified in the first stage are separately mapped to a reference genome that has been hard-masked with RepeatMasker (v4.0.7; http://www.repeatmasker.org/) using the same TE library from stage one (Figure 2.1). Genome-wide coverage profiles are computed using samtools (v1.9) [137] and genomic intervals with enriched coverage from junction read clusters on the 5' and 3' side of TEs are annotated in bed format. Regions of overlap between intervals of junction read clusters from the 5' and 3' side of TEs in the resequenced genome define the locations of TSDs for predicted non-reference TE insertions. The strand of non-reference TE predictions is determined from the relative orientation of alignments of the junction reads to the reference genome and TE library.

In the third stage, all reads from the original whole genome shotgun sequence data are mapped against the same hard-masked reference genome as in stage two (Figure 2.1). This additional mapping step is necessary to obtain all reads that span the TE-flank junctions, as well as identify if any reads are present for the alternative "reference" haplotype that does not carry the non-reference TE insertion. For each candidate non-reference TE insertion site, the number of junction reads covering 5' and 3' side of each candidate TE insertion are estimated as the number of soft-clipped reads overlapping a 10bp window on the 5' and 3' side of the TSD, respectively ($\text{Count}_{junction5'}$ and $\text{Count}_{junction3'}$). The number of non-reference reads

($\text{Count}_{non-ref}$) is estimated as max($\text{Count}_{junction5'}$, $\text{Count}_{junction3'}$). The number of reference reads (Count_{ref}) is estimated as number of non-soft-clipped reads spanning the TSD with at least 3bp extension on both sides of the TSD. The allele frequency for non-reference TEs is heuristically estimated as $\text{Count}_{non-ref}/(\text{Count}_{non-ref} + \text{Count}_{ref})$.

2.4.2 Evaluation of ngs_te_mapper2 performance

To evaluate the prediction performance of ngs_te_mapper2 and ngs_te_mapper under ideal conditions (one homozygous non-reference TE insertion with a known location), we created artificial ISO1 (dm6) genomes that each contain a single synthetic transposon insertion from one of the 125 TE families (excluding *INE-1*) in the Berkeley *Drosophila* Genome Project canonical TE dataset v10.1 (https://github.com/bergmanlab/transposons/blob/master/releases/D_mel_transposon_sequence_set_v10.1.fa). Insertion sites were selected at random in regions of normal recombination that were more than 500 bp from a reference TE in the *D. melanogaster* release 6.38 genome annotation (http://ftp.flybase.net/releases/FB2021_01/dmel_r6.38/gff/dmel-all-r6.38.gff.gz). After selecting an insertions site, a 5bp target site duplication was created and the full length canonical TE sequences was inserted into an otherwise unmodified dm6 genome sequence.

Ten synthetic genomes were created for each family in the *D. melanogaster* TE set, excluding the inactive *INE-1* family, leading to total of 1250 synthetic genomes, each with a single non-reference TE insertion. 625 synthetic genomes contained a non-reference TE insertion of the TE canonical sequence (positive strand insertions), and 625 contained a non-reference TE insertion of the reverse complement of the TE canonical sequence (negative strand insertions). For each synthetic genome, 100bp paired-end reads were simulated at 14X, 25X, 50X, and 100X coverage using wgsim (v0.3.1-r13; -e 0.01 -d 500) [135]. The forward reads of each simulated read pair, the unmodified dm6 reference genome, and the Berkeley *Drosophila* Genome Project canonical TE dataset v10.1 were used as input for ngs_te_mapper

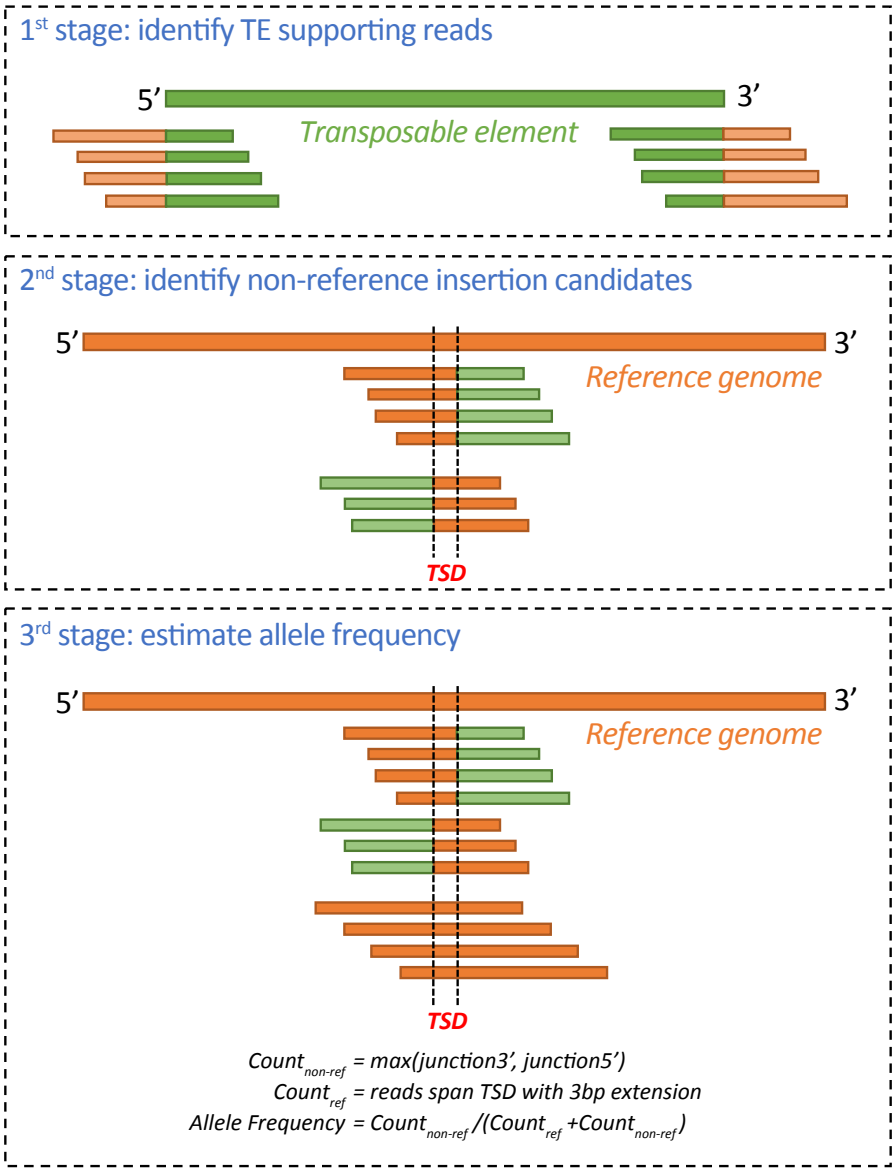

Figure 2.1: **ngs_te_mapper2 workflow for predicting non-reference TE insertions.** In the first stage, raw reads are mapped to the TE consensus sequences. Reads that partially map to TEs are extracted as putative TE supporting reads. In the second stage, putative TE supporting reads are mapped to reference genome that has been hard-masked with RepeatMasker using input TE library. Non-reference TE insertion candidates are identified if alignments of TE supporting reads on 5' and 3' end of TE overlap. In the third stage, raw reads are mapped to unmodified reference genome for estimating intra-sample insertion allele frequency. See details in Section 2.4.1.

and ngs_te_mapper2 to detect non-reference TE insertions using McClintock (revision 40863acf11052b18afb4cdcd7b1124de48cba397; options: -m ngs_te_mapper, ngs_te_mapper2). Non-reference insertion predictions from ngs_te_mapper and ngs_te_mapper2 were considered a true positive if they occurred within 5bp of the actual synthetic insertion location and have the same TE family. Benchmark results under the single homozygous insertion scenario are summarized in Table 2.1. Under ideal conditions, the recall for ngs_te_mapper2 is high ($\geq$91.7%) and far exceeds that of ngs_te_mapper for all coverage levels. Likewise, in this idealized simulation setting the precision for ngs_te_mapper2 is $\geq$97.0% and the same as or better than ngs_te_mapper at all coverage levels.

Table 2.1: **ngs_te_mapper2 performance benchmark using single insertion synthetic data.** ngs_te_mapper [139] and ngs_te_mapper2 were benchmarked by creating single synthetic TE insertions in the ISO1 (dm6) genome assembly, simulating reads from these modified assemblies under different coverages, then generating insertion predictions using unmodified assembly as reference genome and comparing predictions with expected insertion annotations. 10 single synthetic insertion simulation experiments were performed for each of the 125 TE families in *D. melanogaster* (excluding *INE-1*), making up 1250 total simulations each with one synthetic insertion. "Total" represents the total number of predictions from all 1250 experiments after filtering (see filtering criteria in Section 2.4.2). "True Positives" and "False Positives" represent the number of predictions that match or don't match expected insertion annotations, respectively (see matching criteria in Section 2.4.2). "False Negatives" represent the number of expected insertion annotations that are not predicted by the TE detection method. "Precision" represents the number of true positives divided by total number of predictions. "Recall" represents the number of true positives divided by total number of expected insertions (1250).

Method	Coverage	Total	True Positives	False Positives	False Negatives	Precision	Recall
ngs_te_mapper	14	487	481	6	769	98.8%	38.5%
ngs_te_mapper	25	525	519	6	731	98.9%	41.5%
ngs_te_mapper	50	534	527	7	723	98.7%	42.2%
ngs_te_mapper	100	533	524	9	726	98.3%	41.9%
ngs_te_mapper2	14	1149	1146	3	104	99.7%	91.7%
ngs_te_mapper2	25	1181	1173	8	77	99.3%	93.8%
ngs_te_mapper2	50	1189	1174	15	76	98.7%	93.9%
ngs_te_mapper2	100	1211	1175	36	75	97.0%	94.0%

Simulation of single homozygous insertion in unique regions of the dm6 reference genome provides a benchmark of ngs_te_mapper2 under ideal conditions, but does not incorporate the

reality that TEs can insert into more complex regions of the genome, can exist in heterozygous state and are multiple TEs are predicted simultantously in real samples. To model both homozygous and heterozygous non-reference TE insertions and evaluate ngs_te_mapper2 under a more realistic setting, we created synthetic datasets using reads simulated from the ISO1 (dm6) and A4 (GCA_003401745.1) [51] genome assemblies. In theory, a good predictor should be able to accurately predict "non-reference" insertions that are present in genome 1 (e.g., ISO1) but absent from genome 2 (e.g., A4) using reads simulated from genome 1 mapped to genome 2. We therefore simulated 100bp synthetic paired-end sequencing data from the ISO1 genome assembly under 14X, 25X, 50X, 100X coverages using wgsim (v0.3.1-r13; -e 0.01 -d 500) [135] to model homozygous insertions. Additionally, we simulated synthetic paired-end sequencing data by combining equal numbers of reads from both ISO1 and A4 genome assemblies to model heterozygous insertions. The synthetic datasets were used as input to ngs_te_mapper2 to detect non-reference TE insertions using McClintock (revision 40863acf11052b18afb4cdcd7b1124de48cba397; options: -m "trimgalore, ngs_te_mapper2, map_reads"). The A4 assembly was used as the reference genome and the Berkeley *Drosophila* Genome Project canonical TE dataset v10.1 were used for these analyses.

As ground truth for evaluating ngs_te_mapper2 performance, curated TE annotations from the release 6.38 version of *D. melanogaster* genome (`http://ftp.flybase.net/releases/FB2021_01/dmel_r6.38/gff/dmel-all-r6.38.gff.gz`) were lifted over to A4 genome assembly. After excluding *INE-1* insertions and TE insertions in low recombination regions, 627 curated TEs in ISO1 could be lifted over to A4 on the basis of their flanking regions. ngs_te_mapper2 predictions were considered true positives if the predicted TE insertion coordinates were within a 5bp window of a lifted over ISO1 TE annotation and if the predicted TE family was the same as the lifted over annotation. The final benchmark results for ngs_te_mapper2 applied to simulated real genomes are summarized in Table 2.2. Similar to single synthetic insertion simulations above, ngs_te_mapper2 has high precision ($\geq$95.0%) at all coverage levels in simulations designed to model genome-wide TE prediction.

In contrast, recall for ngs_te_mapper2 under a more realistic setting was much lower than in single synthetic insertion simulations, especially at low coverage levels, and was lower for heterozygous insertions than homozygous insertions at all coverage levels. These results indicate that the TE insertion predictions ngs_te_mapper2 makes are accurate but that the method has an appreciable false negative rate on low coverage samples.

Table 2.2: **ngs_te_mapper2 performance benchmark using genome-wide synthetic data from ISO1 and A4 genome assemblies.** Non-reference TE insertion predictions made by ngs_te_mapper2 using the A4 genome assembly as reference were evaluated against curated TE annotations in ISO1 lifted over to A4 coordinates (see Section 2.4.2 for details). Zygosity represents whether simulated reads were generated from both ISO1 and A4 (heterozygous) or ISO1 only (homozygous). "True Positives" and "False Positives" represent the number of predictions that match and doesn't match with lifted over insertion annotations, respectively. "False Negatives" represent the number of lifted over non-reference TE insertion annotations that are not predicted by ngs_te_mapper2. "Precision" represents the number of true positives divided by total number of predictions. "Recall" represents the number of true positives divided by total number of lifted over non-reference TE insertion annotations.

Zygosity	Coverage	Total	True positives	False positives	False negatives	Precision	Recall
heterozygous	14	346	336	10	285	97.1%	53.8%
heterozygous	25	424	412	12	209	97.2%	66.0%
heterozygous	50	476	462	14	159	97.1%	74.0%
heterozygous	100	482	464	18	157	96.3%	74.4%
homozygous	14	437	424	13	197	97.0%	67.9%
homozygous	25	473	461	12	160	97.5%	73.9%
homozygous	50	482	465	17	156	96.5%	74.5%
homozygous	100	516	490	26	131	95.0%	78.5%

2.4.3 Evaluation of a classifier for predicting homozygous or heterozygous TE insertion in single-end WGS data

To fill a gap in tools available to analyze intra-sample TE allele frequencies in single-end WGS data, we developed a classifier to determine whether a TE insertion predicted by ngs_te_mapper2 is homozygous or heterozygous. Our model classifies a TE insertion as homozygous if the intra-sample allele frequency is $\geq$0.95, as heterozygous if the allele frequency is between 0.25 and 0.75, and is considered unclassified if neither of these conditions

are met. To evaluate this approach we used ngs_te_mapper2 predictions made from the simulated paired-end sequencing data generated from ISO1 and A4 genome assemblies described in the previous section. We evaluated the classifier as follows: if the simulated reads were generated from ISO1 only, then all all non-reference TE insertions were expected to be homozygous and the precision was calculated as $\text{Count}_{homozygous}$ / Count_{all}. If the simulated data were a combination of reads from both ISO1 and A4, then all non-reference TE insertions were expected to be heterozygous and the precision is $\text{Count}_{heterozygous}$ / Count_{all}. The final benchmark results were summarized in Table 2.3. Our classifier had ≥91.3% precision at all coverage levels and never falsely classified a heterozygous TE insertions as homozygous, and is thus conservative for the purposes of detecting loss of heterozygosity.

Table 2.3: **Performance benchmark for intra-sample TE insertion zygosity classifier.** ngs_te_mapper2 predictions on synthetic data from ISO1 and A4 genome assemblies were used as input for the classifier. Zygosity represents whether the simulated reads were generated from both ISO1 and A4 (heterozygous) or ISO1 only (homozygous). Precision represents the proportion of predictions being correctly classified as heterozygous or homozygous by the classifier.

Zygosity	Coverage	Total	Homozygous count	Heterozygous count	Unclassified count	Precision
heterozygous	14	346	0	326	20	94.2%
heterozygous	25	424	0	419	5	98.8%
heterozygous	50	476	0	473	3	99.4%
heterozygous	100	482	0	477	5	99.0%
homozygous	14	437	399	4	34	91.3%
homozygous	25	473	438	3	32	92.6%
homozygous	50	482	456	3	23	94.6%
homozygous	100	516	489	9	18	94.8%

2.4.4 Clustering of cell lines using TE insertions reveals rare cases of mismatch with expected provenance

The development of ngs_te_mapper2 enabled us to identify TE insertions from WGS data for cell line identification. We reasoned that TE insertions would be favorable genetic markers for cell line identification in *Drosophila* because the joint processes of germline transposition in

whole flies and somatic transposition in cell culture together would create unique TE profiles, both for cell lines derived from distinct *D. melanogaster* donor genotypes and for sub-lines of cells derived from the same original donor genotype (Figure 2.2). Furthermore, we posited that shared presence or absence of TE insertions at orthologous loci would allow the identity or similarity among cell line samples to be assessed based on a clustering approach.

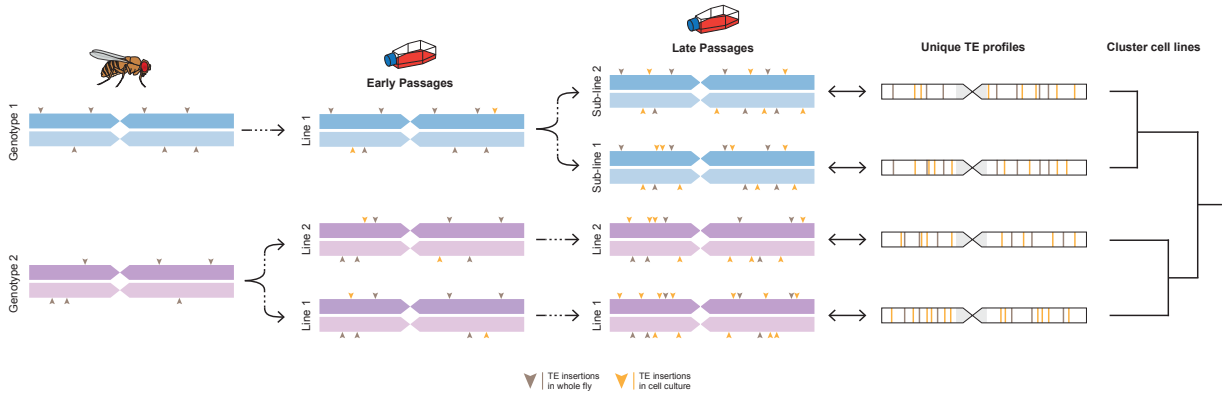

Figure 2.2: **Germline and somatic transposition jointly can create unique TE profiles in *Drosophila* cell line genomes.** A homologous pair of chromosomes is shown for two donor fly genotypes used to establish two distinct cell lines. TE profiles initially differ because transposition events in whole flies (grey arrowheads) are maintained at low population frequencies by purifying selection [52]. After establishment of distinct cell lines, ongoing transposition in cell culture (orange arrowheads) leads to increased TE abundance relative to whole flies [189, 100, 193] and further differentiates TE profiles, both for distinct cell lines derived from the same or different donor genotypes as well as for sub-lines of the same cell line. Ultimately these processes lead to unique TE profiles that can identify cell lines and allow them to be clustered based on shared presence or absence of TE insertions at orthologous loci. The model depicts a simplified case of diploidy, when in reality cell culture genomes can have complex genome structure due to polyploidy and segmental aneuploidy.

We initially investigated the possibility of TE-based cell line identification in *Drosophila* using public genome sequences for 26 samples from 18 cell lines generated by the modENCODE project [129] (Table 2.4). Paired-end Illumina WGS sequences were used to predict non-reference TEs using TEMP [259], which showed the least dependence on read length (Figure 2.3) or coverage (Figure 2.4) out of eight non-reference TE detection methods tested on the data used in this study. We clustered cell lines on the basis of their TE profiles using Dollo parsimony, which accounts for the virtually homoplasy-free nature of TE insertions

within species [23, 194], the ancestral state of TE absence at individual loci [23] and false negative predictions inherent in non-reference TE detection software [175, 198, 235]. Use of Dollo parsimony for clustering cell line samples also allows ancestral states to be reconstructed, facilitating inference of which TE families diagnostically identify individual cell lines or groups of cell lines. We note that we do not attempt to interpret the clustering relationships among distinct cell lines in an evolutionary context, however our approach does provide insight into the evolutionary history of clonally-evolving sub-lines established from the same original cell line.

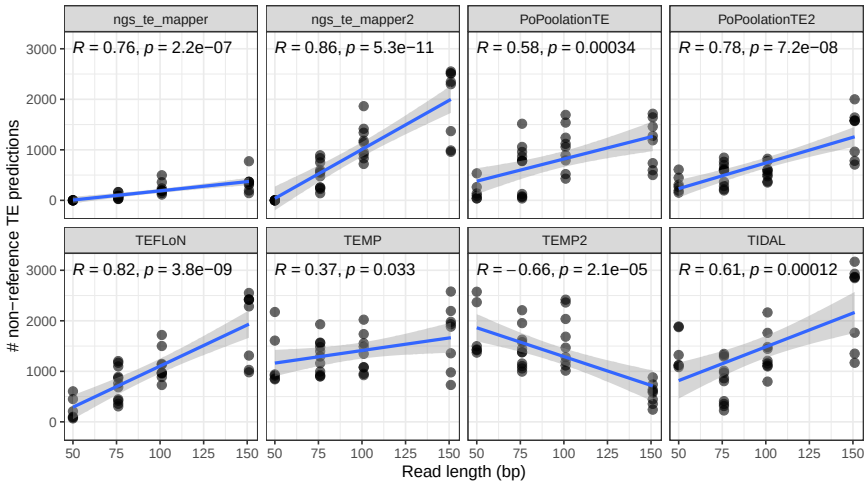

Figure 2.3: **Relationship between read length and number of non-reference TE predictions for the expanded dataset of 34 *Drosophila* cell line samples.** Each panel represents predictions from one of the eight component methods designed for detection of TE insertions in *Drosophila* that is included in McClintock. The X-axis represents read length in base pairs (bp) and the Y-axis represents the number of non-reference TE predictions. The best fit line and 95% CI were included using linear method. Pearson correlation coefficient with p-values are shown on the top of each panel.

We predicted between 730 and 2579 non-reference TE insertions in euchromatic regions of *Drosophila* cell line samples from the modENCODE project (Table 2.5). As reported previously for human cancer cell lines [251], each *Drosophila* cell line sample had a unique profile of TE insertions. The most parsimonious clustering of *Drosophila* cell lines using TE profiles revealed several expected patterns that indicate TE insertions reliably mark the identity of *Drosophila* cell lines (Figure 2.5A). First, replicate samples of the same cell line

Table 2.4: **Metadata and sequencing information for 34 paired-end whole genome shotgun sequencing samples from 22 *Drosophila* cell lines used in this study.** Samples indicated by an asterisk were generated in the current study, while other samples were generated by the modENCODE project [129]. *Drosophila* Genomics Resource Center (DGRC) cell line names and stock identifiers are given for all cell line samples except two mbn2 samples obtained from the Gorski lab (Canada's Michael Smith Genome Sciences Centre, BC Cancer) and the Strand lab (University of Georgia), respectively. For DGRC cell lines, the donor lab represents the lab who donated the stock to the DGRC. Ancestral genotypes represents the genotype of flies from which the cell lines were established. Inferred ploidy represents the ploidy estimated by analyzing DNA density of whole genome data using the method of Lee *et al.* 2014 [129]. Inferred sex represents the sex of the cell line inferred by analyzing DNA density of whole genome data and analysis of sex determination gene expression based on Lee *et al.* 2014 [129]. Coverage represents the average mapped depth of coverage after quality and adaptor trimming. N.A. indicates that this information is not available.

Cell line	DGRC ID	FlyBase ID	Donor lab	Lab origin	Ancestral genotype	Inferred ploidy	Inferred sex	SRA	Read length	Coverage	Primary replicate
1182-4H	DGRC-177	FBtc0000177	Debec	Debec	mh	2	female	SRR497717	101	26.46	yes
CME-L1	DGRC-156	FBtc0000156	Cottam & Milner	Milner	Oregon-R	2	male	SRR497712	101	62.17	yes
CME-W1-Cl.8+	DGRC-151	FBtc0000151	Cottam & Milner	Milner	Oregon-R	2	male	SRR612105	50	10.99	no
CME-W1-Cl.8+	DGRC-151	FBtc0000151	Cottam & Milner	Milner	Oregon-R	2	male	SRR612106	50	10.05	no
CME-W1-Cl.8+	DGRC-151	FBtc0000151	Cottam & Milner	Milner	Oregon-R	2	male	SRR497726	76	18.14	yes
CME-W2	DGRC-155	FBtc0000155	Cottam & Milner	Milner	Oregon-R	2	male	SRR497730	76	31.15	yes
Kc167	DGRC-1	FBtc0000001	Cherbas	Echalier	e/se	4	female	SRR612107	50	15.01	yes
Kc167	DGRC-1	FBtc0000001	Cherbas	Echalier	e/se	4	female	SRR612109	50	10.82	no
mbn2	DGRC-147	FBtc0000147	Werner & Hultmark	Gateff	l(2)mbn	4	male	SRR497728	76	18.38	no
mbn2 (*)	DGRC-147	FBtc0000147	Werner & Hultmark	Gateff	l(2)mbn	4	male	SRR13360020	151	112.69	yes
mbn2 (Gorski) (*)	N.A.	N.A.	Gorski	Gateff	l(2)mbn	4	male	SRR13360019	151	130.60	no
mbn2 (Strand) (*)	N.A.	N.A.	Strand	Gateff	l(2)mbn	4	male	SRR13360018	151	136.06	no
ML-DmBG2-c2 (*)	DGRC-53	FBtc0000053	Ueda & Ui-Tei	Miyake	$y^1v^1f^1mal^{F1}$	2	male	SRR13360022	151	127.03	yes
ML-DmBG3-c2 (*)	DGRC-68	FBtc0000068	Ueda & Ui-Tei	Miyake	$y^1v^1f^1mal^{F1}$	2	male	SRR13360021	151	130.59	yes
ML-DmD16-c3	DGRC-97	FBtc0000097	Ueda & Ui-Tei	Miyake	$y^1v^1f^1mal^{F1}$	4	female	SRR497715	76	10.73	no
ML-DmD16-c3	DGRC-97	FBtc0000097	Ueda & Ui-Tei	Miyake	$y^1v^1f^1mal^{F1}$	4	female	SRR497710	101	48.55	yes
ML-DmD17-c3	DGRC-107	FBtc0000107	Ueda & Ui-Tei	Miyake	$y^1v^1f^1mal^{F1}$	4	female	SRR497725	101	55.02	yes
ML-DmD20-c2	DGRC-109	FBtc0000109	Ueda & Ui-Tei	Miyake	$y^1v^1f^1mal^{F1}$	2	male	SRR497724	76	26.93	yes
ML-DmD20-c5	DGRC-112	FBtc0000112	Ueda & Ui-Tei	Miyake	$y^1v^1f^1mal^{F1}$	2	male	SRR497718	76	6.24	no
ML-DmD20-c5	DGRC-112	FBtc0000112	Ueda & Ui-Tei	Miyake	$y^1v^1f^1mal^{F1}$	2	male	SRR497723	101	15.42	yes
ML-DmD4-c1	DGRC-126	FBtc0000126	Ueda & Ui-Tei	Miyake	$y^1v^1f^1mal^{F1}$	2	male	SRR497716	76	34.62	yes
ML-DmD8	DGRC-92	FBtc0000092	Ueda & Ui-Tei	Miyake	$y^1v^1f^1mal^{F1}$	2	female	SRR497729	76	29.34	yes
ML-DmD9	DGRC-85	FBtc0000085	Ueda & Ui-Tei	Miyake	$y^1v^1f^1mal^{F1}$	4	female	SRR497714	76	8.89	no
ML-DmD9	DGRC-85	FBtc0000085	Ueda & Ui-Tei	Miyake	$y^1v^1f^1mal^{F1}$	4	female	SRR497711	101	40.28	yes
OSC (*)	DGRC-288	FBtc0000288	Saito & Siomi	Niki	w1118	2	female	SRR13360016	151	131.31	yes
OSS (*)	DGRC-190	FBtc0000190	Niki	Niki	w1118	2	female	SRR13360017	151	117.27	yes
S1	DGRC-9	FBtc0000009	Cherbas	Schneider	Oregon-R	2	male	SRR497713	76	30.39	yes
S2-DRSC	DGRC-181	FBtc0000181	Perrimon & Mathey-Prevot	Schneider	Oregon-R	4	male	SRR612111	50	15.45	no
S2-DRSC	DGRC-181	FBtc0000181	Perrimon & Mathey-Prevot	Schneider	Oregon-R	4	male	SRR612112	50	22.10	yes
S2R+	DGRC-150	FBtc0000150	Wheeler	Schneider	Oregon-R	4	male	SRR497722	76	5.32	no
S2R+	DGRC-150	FBtc0000150	Wheeler	Schneider	Oregon-R	4	male	SRR497719	101	10.66	yes
S3	DGRC-5	FBtc0000005	Cherbas	Schneider	Oregon-R	4	male	SRR497721	101	14.99	yes
Sg4	DGRC-179	FBtc0000179	Pirrotta	Schneider	Oregon-R	4	male	SRR497720	101	26.91	no
Sg4 (*)	DGRC-179	FBtc0000179	Pirrotta	Schneider	Oregon-R	4	male	SRR13360015	151	131.41	yes

cluster most closely with one another with 100% bootstrap support in all seven cases where data is available (S2, S2R+, CME-W1-Cl.8+, ML-DmD9, ML-DmD16-c3, ML-DmD20-c5, and Kc167). Second, different cell lines created in the same lab (presumably from the same

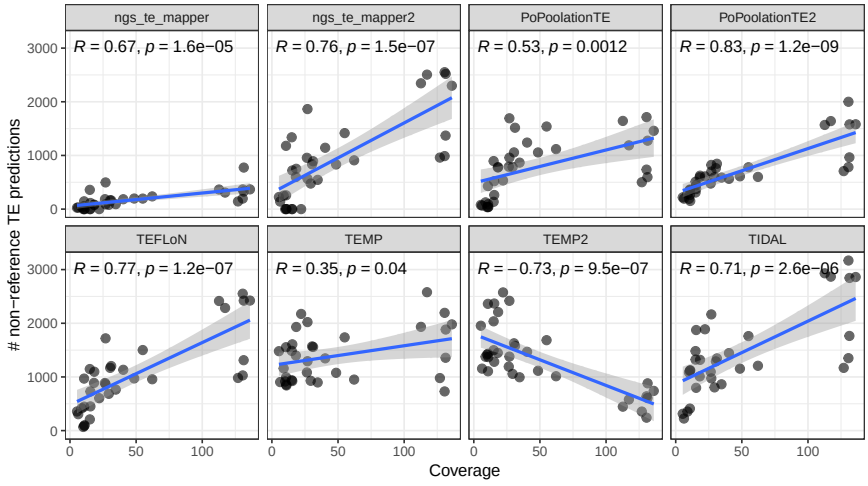

Figure 2.4: **Relationship between average genome coverage and number of non-reference TE predictions for the expanded dataset of 34 *Drosophila* cell line samples.** Each panel represents predictions from one of the eight component methods designed for detection of TE insertions in *Drosophila* that is included in McClintock. The X-axis represents the average genome coverage computed by McClintock and the Y-axis represents the number of non-reference TE predictions. The best fit line and 95% CI were included using linear method. Pearson correlation coefficient with p-values are shown on the top of each panel.

ancestral fly genotype) cluster with each other before they cluster with cell lines generated in other labs, or with cells lines having different ancestral genotypes. Third, we observe that divergent sub-lineages of the same cell line (i.e., S2 and S2R+) cluster closely together [213, 245]. We also find weak evidence for clustering of cell lines generated in different labs (Schneider, Milner) that are derived from the same putative ancestral fly stock (Oregon-R). However, we caution against over-interpretation of this result, given previous reports for substantial genetic diversity among common lab stocks like Oregon-R [193, 227]. Also, cell lines derived from the Schneider and Milner labs have distinct B-allele frequency (BAF) profiles, suggesting different ancestral Oregon-R genotypes (Figure 2.6B).

Overall, clustering patterns based on TE profiles suggest that misidentification is rare among the panel of cell lines sequenced by modENCODE. However, we observed two cases

Table 2.5: **Summary of predictions generated by eight non-reference TE insertion detection methods for 34 *Drosophila* cell line samples.** Numbers of non-reference TE insertion predictions are based on default settings for TIDAL [193, 246] and default McClintock [175] settings for all other methods. *INE-1* and non-reference TE insertion predictions in low recombination regions were excluded from all methods. New sequence data from this study are indicated by asterisks.

Cell line	SRA	TEMP	TEMP2	PoPoolationTE	PoPoolationTE2	TEFLoN	ngs_te_mapper	ngs_te_mapper2	TIDAL
1182-4H	SRR497717	1084	1192	790	476	887	185	956	1096
CME-L1	SRR497712	951	1013	1118	600	957	237	909	1208
CME-W1-Cl.8+	SRR612105	841	1370	40	152	100	0	1	1125
CME-W1-Cl.8+	SRR612106	915	1418	37	192	68	0	0	1083
CME-W1-Cl.8+	SRR497726	1398	1448	776	594	889	83	599	1010
CME-W2	SRR497730	1559	1578	1516	845	1203	163	892	1345
Kc167	SRR612107	939	1498	136	309	209	0	0	1323
Kc167	SRR612109	856	1432	91	266	88	0	0	1119
mbn2	SRR497728	1931	2209	781	623	1097	77	751	1318
mbn2 (*)	SRR13360020	1933	446	1643	1568	2415	365	2344	2931
mbn2 (Gorski) (*)	SRR13360019	2194	639	1714	2000	2551	366	2551	3169
mbn2 (Strand) (*)	SRR13360018	1979	740	1459	1581	2423	368	2299	2862
ML-DmBG2-c2 (*)	SRR13360022	979	355	501	708	981	143	960	1169
ML-DmBG3-c2 (*)	SRR13360021	730	241	736	782	1028	196	988	1350
ML-DmD16-c3	SRR497715	995	1105	37	365	446	57	257	410
ML-DmD16-c3	SRR497710	1077	1115	1056	611	973	197	832	1154
ML-DmD17-c3	SRR497725	1737	1685	1539	780	1501	196	1416	1761
ML-DmD20-c2	SRR497724	1293	1379	958	699	864	89	566	973
ML-DmD20-c5	SRR497718	904	1155	65	195	306	30	141	225
ML-DmD20-c5	SRR497723	924	1282	520	374	729	114	720	797
ML-DmD4-c1	SRR497716	897	994	866	588	764	92	545	863
ML-DmD8	SRR497729	928	1057	783	596	685	80	476	805
ML-DmD9	SRR497714	1160	1376	126	283	416	34	252	354
ML-DmD9	SRR497711	1346	1468	1240	564	1133	184	1143	1444
OSC (*)	SRR13360016	1357	607	596	964	1312	327	1370	1764
OSS (*)	SRR13360017	2579	580	1188	1640	2285	308	2506	2868
S1	SRR497713	1569	1627	1057	763	1169	165	835	1290
S2-DRSC	SRR612111	1608	2368	263	447	451	0	0	1874
S2-DRSC	SRR612112	2174	2575	534	609	604	0	0	1888
S2R+	SRR497722	1480	1954	78	219	357	26	225	313
S2R+	SRR497719	1554	2362	429	353	969	145	1179	1114
S3	SRR497721	1486	2036	895	509	1147	359	1338	1482
Sg4	SRR497720	2022	2418	1692	823	1719	497	1864	2165
Sg4 (*)	SRR13360015	1883	881	1275	1579	2421	774	2519	2844

where the similarity of cell lines based on genome-wide TE profiles conflicted with expectations based on reported provenance. First, we unexpectedly found that the Sg4 cell line (originally called Sf4 by its maker Donna Arndt-Jovin) clusters most closely with S3 cells, although the DGRC and FlyBase currently consider Sg4 to be a variant of S2 cells (http://flybase.org/reports/FBrf0205934.html; http://flybase.org/reports/FBtc0000179; https://dgrc.bio.indiana.edu/cells/S2Isolates). More strikingly, we also observed that the mbn2 cell line originally reported by Gateff *et al.* 1980 [82] to be derived from the l(2)mbn stock was placed inside a well-supported cluster containing cell lines (S1, S2,

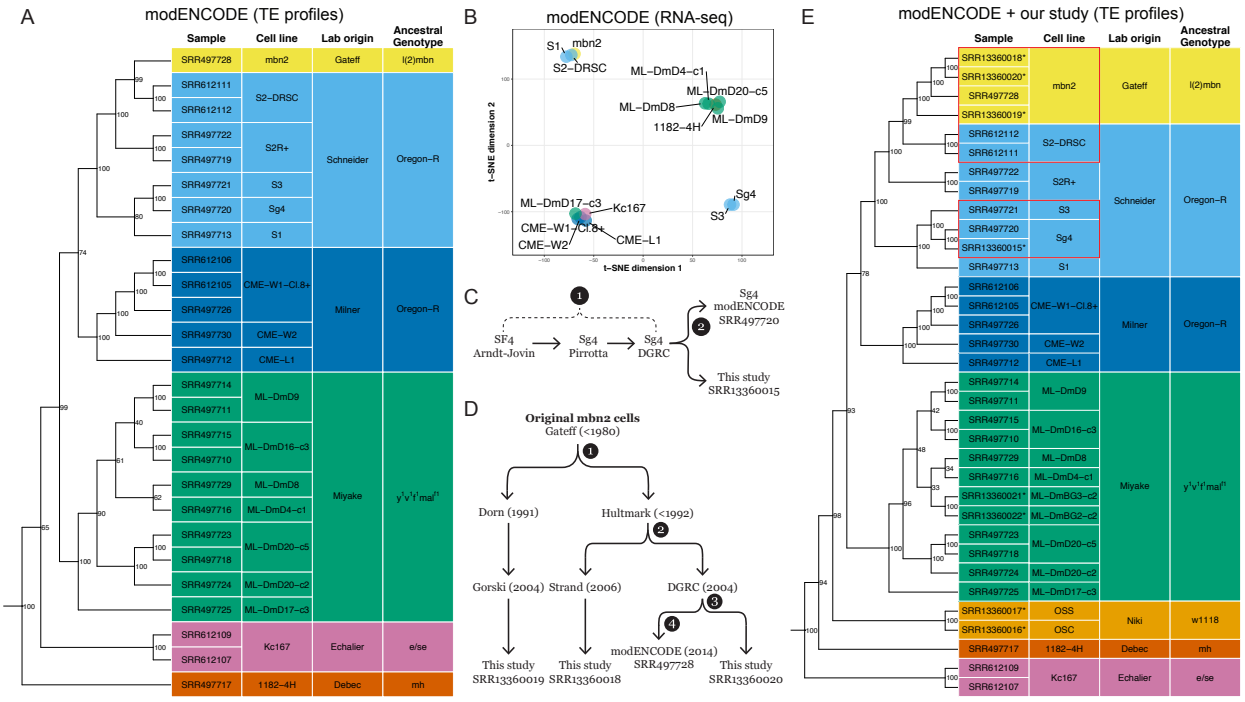

Figure 2.5: **TE insertion profiles cluster *Drosophila* cell lines by lab origin and reveal unexpected placement of the Sg4 and mbn2 cell lines.** (A) Clustering of *Drosophila* cell line samples from the modENCODE project was constructed using Dollo parsimony based on non-reference TE insertions. Samples are colorized by the lab origin based on the first publication reporting the original variant of the cell line. Ancestral genotype is based on the *D. melanogaster* stock reported to create the original variant of the cell line. (B) t-SNE visualization of 15 *Drosophila* cell line samples using transcriptomic data in Stoiber *et al.* 2016 [228]. Samples are colorized by the lab origin of cell lines. (C) Key events in the history of the Sg4 cell line creation and distribution. (D) Key events in the history of the mbn2 cell line distribution. Node labels in panels C and D represent timepoints in the past that potential cell line misidentification events could have occurred. (E) Clustering of *Drosophila* cell line samples from the modENCODE project plus new data reported here (indicated by asterisks in panel E) was constructed using Dollo parsimony based on non-reference TE insertions. Numbers beside nodes in panels A and E indicate percent support based on 100 bootstrap replicates. Red boxes in panel E highlight cases where the reported provenance of *Drosophila* cell lines conflicts with identity inferred from genomic data.

S2R+, S3, Sg4) generated by Schneider *et al.* 1972 [213] from an Oregon-R stock. Our clustering of mbn2 cells inside the Schneider cell clade is consistent with a previously unexplained

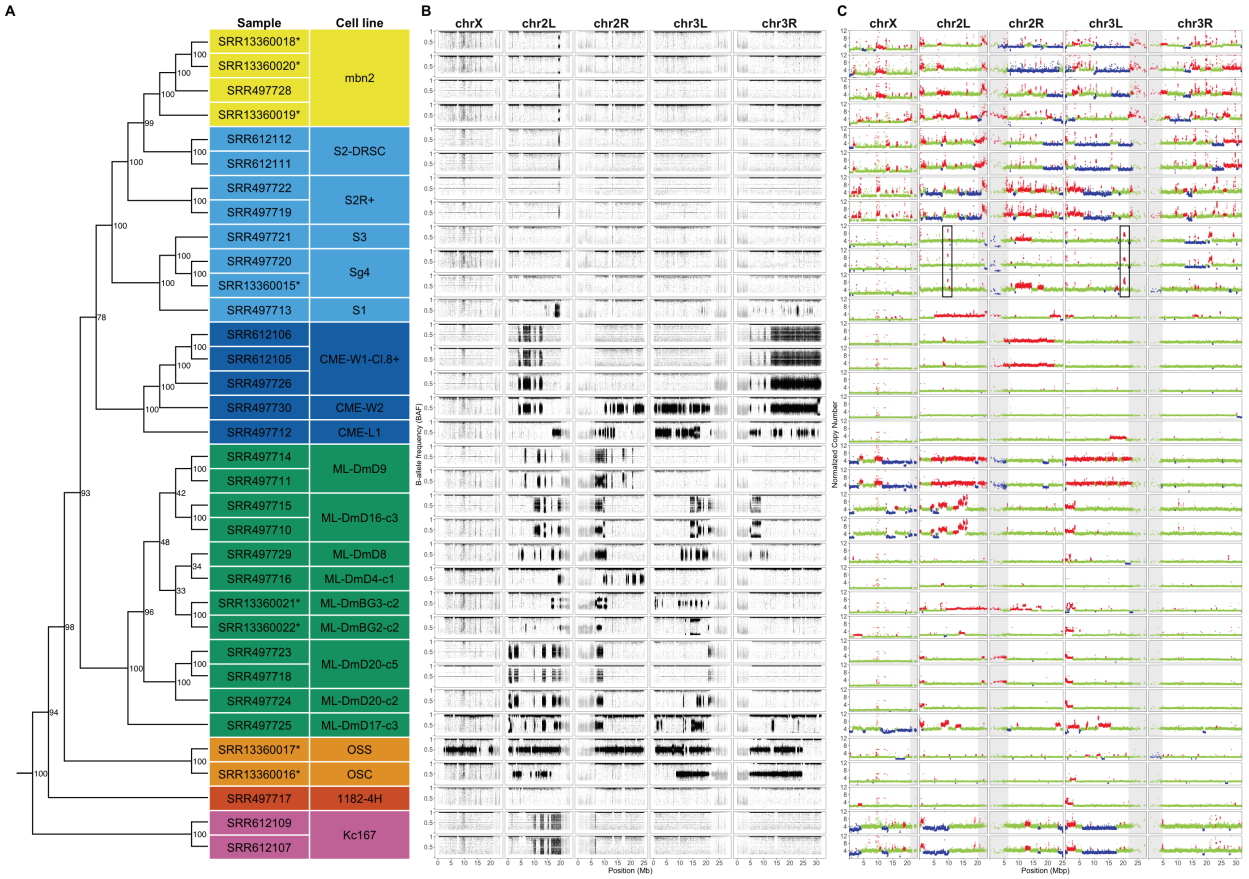

Figure 2.6: **Copy number and B-allele frequency profiles for the expanded dataset of 34 *Drosophila* cell line samples.** (A) Dollo parsimony tree of 34 *Drosophila* cell lines samples (including replicates and sub-lines) based on non-reference TE predictions. Node labels indicate support for each clade based on 100 bootstrap replicates. New sequence data from this study are indicated by asterisks. (B) B-allele frequency profiles for *Drosophila* cell lines on major chromosome arms. For a given SNP, the B-allele frequency (BAF) was determined as the coverage of reads supporting non-reference allele divided by total coverage at that position. SNPs in low recombination regions are plotted in grey. (C) Copy number profiles for *Drosophila* cell lines on major chromosome arms. Each data point represents normalized copy number (ratio*ploidy) for a given 10kb window estimated by Control-FREEC [37]. Data points for each window are colorized by CNV status (red: CNV gain; green: no CNV; blue: CNV loss), which are based on the comparison between normalized copy number for that window and baseline ploidy for the chromosome arm. Black boxes in panel C highlight regions where Sg4 and S3 cell lines share the same copy number gains that are not shared in other cell samples. Low recombination regions are shaded in grey.

observation that mbn2 cells share an unexpectedly high proportion of TE insertions with both S2 and S2R+ cells [193].

Clarification of the provenance of the Sg4 and mbn2 cell lines used by modENCODE is important since many functional genomics resources were generated for these cell lines [202] and over 125 publications involving these cell lines are curated in FlyBase [126]. To cross-validate genomic clustering based on TE profiles and to assess potential functional similarity between Sg4↔S3 and mbn2↔S2 cell lines, we clustered cell lines on the basis of their transcriptomes. Transcriptome-based clustering should reveal similarities among cell types rather than genotypes, and thus is not expected to globally match our TE insertion based clustering. However, both cell type and genotype clustering should support the similarity of pairs of cell lines that are derived from a common ancestral cell line.

Previous transcriptome-based clustering of cell lines based on early whole-genome tiling microarray datasets from the modENCODE project did not reveal similarities among Sg4 and S3 or mbn2 and S2 [56], however clustering of small RNA-seq data did reveal similarities among these cell lines [241]. Using a consistent batch of poly-A RNA-seq samples from a panel of 15 DGRC cells lines with genome data [228] (Table 2.6), we estimated expression levels for protein-coding genes then used T-distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction [147, 146] to visualize similarity of cell lines based on their gene expression profiles. This analysis revealed that gene expression profiles based on transcriptome data support the clustering of Sg4 with S3 and mbn2 with S2 (Figure 2.5B). Transcriptome-based clustering of Sg4 with S3 and mbn2 with S2 is also observed in a different batch of RNA-seq samples generated independently by the modENCODE project [42] (Figure 2.7, Table 2.6). These results provide replicated transcriptomic support for the clustering of Sg4↔S3 and mbn2↔S2 cell lines revealed by TE profiles, and also highlight functional similarities between these pairs of cell lines.

Table 2.6: **Summary of transcriptome data for *Drosophila* cell lines analyzed in this study.** Samples are from two consistent batches of RNA-seq experiments performed on DGRC cell lines with genome data. The first batch is poly-A RNA-seq samples from Stoiber *et al.* 2016 [228] (PRJNA306537) and the other batch is total RNA-seq samples from Brown *et al.* 2014 [42] (PRJNA75285). All samples have 100bp paired end reads.

Cell_line	SRA	Study_accession	Gigabases
1182-4H	SRR1197409	PRJNA75285	10.0
1182-4H	SRR3038250	PRJNA306537	3.7
CME-L1	SRR1197410	PRJNA75285	10.8
CME-L1	SRR3038125	PRJNA306537	4.4
CME-W1-Cl.8+	SRR3038123	PRJNA306537	3.1
CME-W2	SRR1197407	PRJNA75285	10.9
CME-W2	SRR3038127	PRJNA306537	2.6
Kc167	SRR1197456	PRJNA75285	11.6
Kc167	SRR3040509	PRJNA306537	3.4
mbn2	SRR1197406	PRJNA75285	9.3
mbn2	SRR3040560	PRJNA306537	2.7
ML-DmD16-c3	SRR1197401	PRJNA75285	10.1
ML-DmD17-c3	SRR3041988	PRJNA306537	1.8
ML-DmD20-c5	SRR1197396	PRJNA75285	10.4
ML-DmD20-c5	SRR3042157	PRJNA306537	2.8
ML-DmD4-c1	SRR1197397	PRJNA75285	10.3
ML-DmD4-c1	SRR3042204	PRJNA306537	2.7
ML-DmD8	SRR1197284	PRJNA75285	8.0
ML-DmD8	SRR3042539	PRJNA306537	4.3
ML-DmD9	SRR1197283	PRJNA75285	10.3
ML-DmD9	SRR3042543	PRJNA306537	3.4
S1	SRR1197281	PRJNA75285	8.8
S1	SRR3042563	PRJNA306537	3.4
S2-DRSC	SRR1197282	PRJNA75285	9.6
S2-DRSC	SRR3042565	PRJNA306537	3.1
S2R+	SRR1197280	PRJNA75285	9.0
S3	SRR1197277	PRJNA75285	8.9
S3	SRR3042571	PRJNA306537	3.8
Sg4	SRR1197278	PRJNA75285	8.8
Sg4	SRR3042573	PRJNA306537	4.9

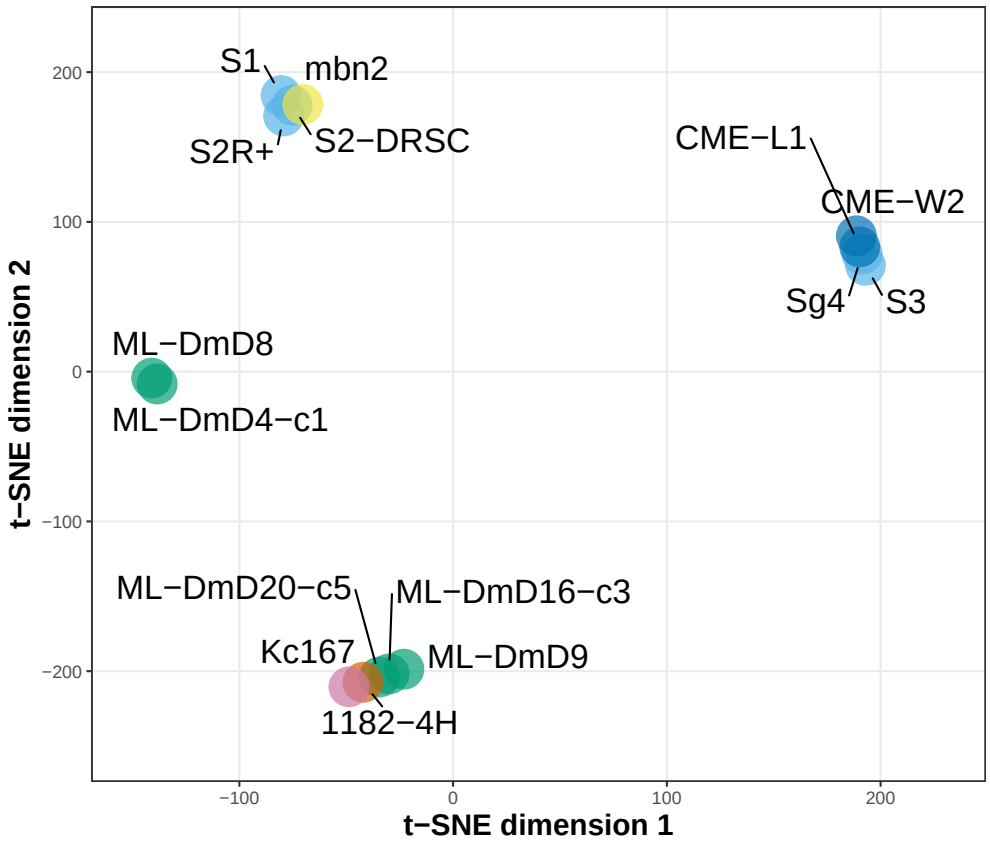

Figure 2.7: **t-SNE visualization of 15 *Drosophila* cell lines using total RNA-seq data from [42].** t-SNE visualization was produced with perplexity=1. Samples are colorized by the lab origin of cell lines.

2.4.5 TE profiles help resolve the provenance of the Sg4 and mbn2 cell lines

To better understand the cause of the surprising patterns of clustering for the Sg4 and mbn2 cell lines in the modeENCODE data, we generated paired-end Illumina WGS sequences for additional samples of Sg4 and mbn2 cells from the DGRC and other sources. In addition, we sequenced several other popular *Drosophila* cell lines (OSS, OSC, ML-DmBG3-c2, and ML-DmBG2-c2) that were not originally sequenced in the modENCODE cell line genome project [129]. To guide sampling and aid the interpretation of the expanded dataset, we reconstructed

key events in the history of the Sg4 (Figure 2.5C) and mbn2 cell lines (Figure 2.5D). We predicted non-reference TE insertions in these additional samples and then reclustered the expanded dataset using the same methods as the modENCODE-only dataset. Inclusion of additional samples altered some details of the clustering relationships among D-series cell lines generated by the Miyake lab and the position of distantly related cell lines with respect to the root (Kc167 and 1182-4H) (Figure 2.5A vs E). However, key aspects of our clustering approach that facilitate cell line identification (replicates clustering most closely, clustering of cell lines from the same lab/ancestral genotype) appear to be robust to the set of cell line samples analyzed.

Clustering TE profiles from this expanded dataset of 34 samples from 22 *Drosophila* cell lines revealed that our resequenced sample of DGRC Sg4 clusters with high support first with the modENCODE sample of DGRC Sg4 then with S3 (Figure 2.5E). This result confirms the reproducibility of the S3↔Sg4 genomic similarity and rejects the possibility of cell line swap during the modENCODE cell line sequencing project (node 2; Figure 2.5C). Additional evidence for the similarity of Sg4 and S3 can be observed in their BAF and CNV profiles. All Sg4 and S3 samples are generally devoid of heterozygosity across their entire genomes, including lacking a small patch of heterozygosity at the base of chromosome arm 2L that is present in all S2 or S2R+ samples (Figure 2.6B). All Sg4 and S3 samples also share CNVs on chromosome arms 2L and 3L that are not present in any S2/S2R+ sample (Figure 2.6C). Together, these data support the conclusion that DGRC Sg4 is a variant of the S3 cell line, not the S2 cell line as currently thought. Presently, we are unable to determine where misidentification of Sg4 as a variant of S2 occurred in the provenance chain from initial development of the cell line by the Arndt-Jovin lab to receipt by the DGRC (Figure 2.5C, node 1). Future analysis of additional Sg4 sub-lines circulating in the research community [171, 215] will be necessary to establish the timing of this event and if the S3↔Sg4 similarity first observed in the DGRC Sg4 sub-line is more widespread.

The second case of unexpected clustering we observed in the modENCODE data involving mbn2 and S2 is potentially more surprising and consequential given that these cell lines are reported to be derived from different ancestral genotypes. mbn2 cells were reportedly derived from a stock carrying l(2)mbn on a 2nd chromosome marked with three visible mutations [81, 82], while S2 cells were derived from a wild-type Oregon-R stock [213]. Unfortunately, the l(2)mbn mutation was never characterized at the molecular level, and no fly stocks carrying l(2)mbn currently exist in public stock centers that could be sequenced and compared with the mbn2 cell line. In the absence of external biological resources to verify the identity of an authentic mbn2 cell line, we attempted to infer the timing and extent of the potential mbn2 misidentification event first observed in the modENCODE data by sequencing sub-lines of mbn2 from DGRC and other sources. We resequenced another sample of the DGRC mbn2 sub-line, a sub-line from the Strand lab (University of Georgia) derived from the same donor as the DGRC sub-line (Hultmark lab, Umeå University), and a sub-line from the Gorski lab (Canada's Michael Smith Genome Sciences Centre, BC Cancer) derived from an independent donor (Dorn lab, Johannes Gutenberg-Universität Mainz) (Figure 2.5D). The Hultmark and Dorn labs each report obtaining mbn2 cells directly from the Gateff lab in the early 1990s [210, 196]. This sampling allowed us to infer if potential misidentification occurred during the modENCODE project (node 4), at the DGRC (node 3), in the Hultmark lab (node 2), or in the Gateff lab (node 1) (Figure 2.5D).

Analysis of TE profiles in our expanded dataset revealed that all four samples of mbn2 cluster together as a single, well-supported group that is most similar to a cluster containing S2 cells (Figure 2.5E). The detailed relationships among sub-lines within the mbn2 cluster deviate slightly from expectations based on cell line history (Figure 2.5D), however this discrepancy appears to be caused by differences in read length or coverage between the data from modENCODE and our study (Figure 2.9). All mbn2 samples have the low SNP heterozygosity across most of their genomes that is characteristic of Schneider cell lines, and also share the small patch of heterozygosity at the base of chromosome arm 2L found in

S2 and S2R+ cells (Figure 2.6B). Additionally, all four mbn2 samples share widespread segmental aneuploidy across the entire euchromatin that is a common hallmark of S2 and S2R+ cells, but not other *Drosophila* cell lines (Figure 2.6C). Together, these data support the conclusions that multiple independent sub-lines of mbn2 cells all share a common origin and are likely to descend from a single divergent lineage of S2 cells. Based on these observations, we speculate that currently-circulating mbn2 cells derive from a mislabelling or cross-contamination event with S2 cells in the Gateff lab that occurred prior to distribution to the Hultmark or Dorn labs (Figure 2.5D, node 4). This scenario is consistent with the facts that S2 cells were developed and widely distributed prior to the origin of mbn2 cells [213, 82] and that there was a 12 year gap between the initial report describing mbn2 cells and use in any subsequent publication [82, 210].

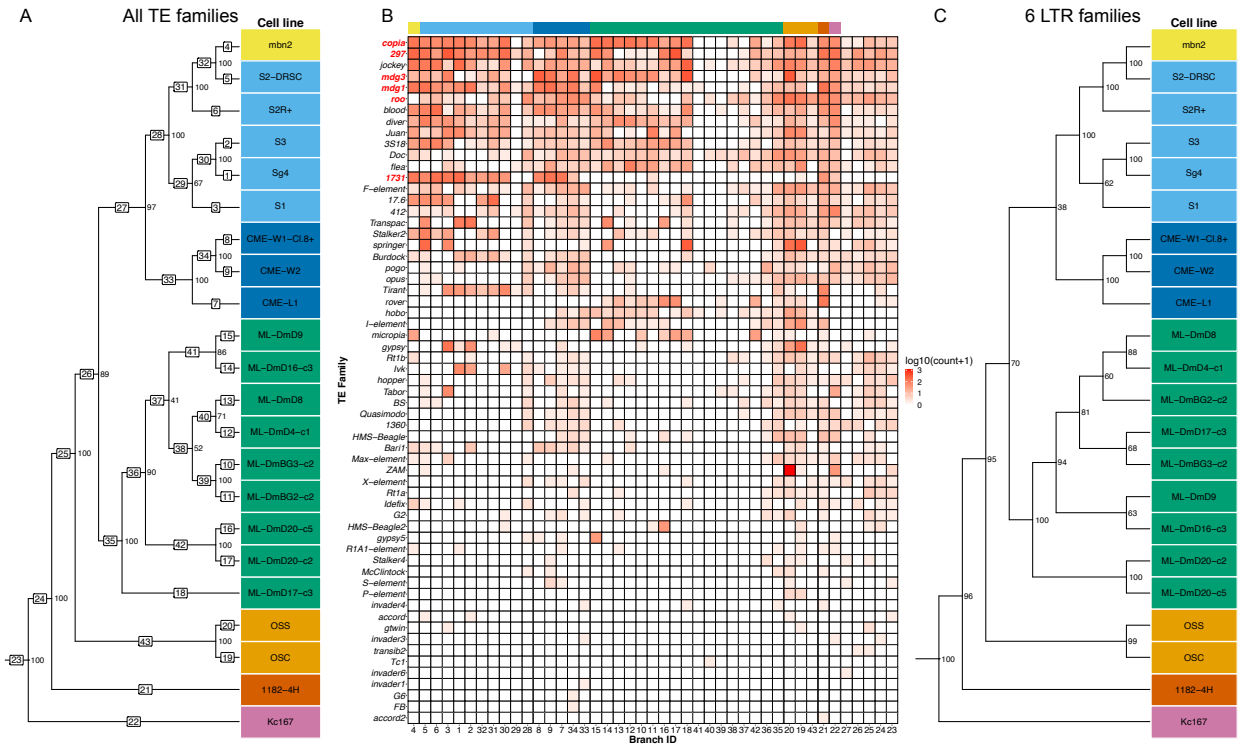

Figure 2.8: **A small subset of LTR retrotransposon families can identify *Drosophila* cell lines.** (A) Dollo parsimony tree of 22 *Drosophila* cell lines (without replicates) based on non-reference TE predictions for all 125 *D. melanogaster* TE families. Samples are colorized by lab origin as in Figure 2.5. Numbers inside boxes on branches indicate branch ID, and numbers beside nodes indicate percent support based on 100 bootstrap replicates. (B) Heatmap showing the number of non-reference TE insertion gain events per family on each branch of the tree in panel (A) based on ancestral state reconstruction using Dollo parsimony. The heatmap is colorized by log-transformed (log10(count+1)) number of gains per family per branch, sorted top to bottom by overall non-reference TE insertion gains per family across all branches, and sorted left to right into clades representing lab origin with lab origin clade color codes indicated at the top of the heatmap. The six diagnostic LTR retrotransposon families used in panel (C) are highlighted in red. (C) Dollo parsimony tree of 22 *Drosophila* cell lines (without replicates) based on non-reference predictions of six LTR retrotransposon families (*297*, *copia*, *mdg3*, *mdg1*, *roo* and *1731*). Numbers beside nodes indicate percent support based on 100 bootstrap replicates.

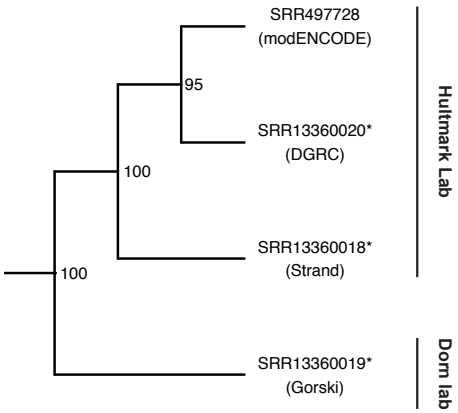

Figure 2.9: **Clustering of normalized mbn2 cell line genome samples from the modENCODE project plus this study.** Clustering was performed on TE insertions generated using mbn2 samples that were normalized by trimming read lengths to 76bp and down-sampling to 19X depth. For this analysis, we also relaxed TEMP filtering to include more weakly-supported predictions at otherwise high-quality loci because of the lower overall coverage in all samples. Numbers beside nodes indicate percent support based on 100 bootstrap replicates. Tip labels include SRA run identifiers and source lab for samples (in parentheses). New sequence data from this study are indicated by asterisks. Clade annotations indicate the donor lab from which the source lab obtained their sub-line of mbn2 cells.

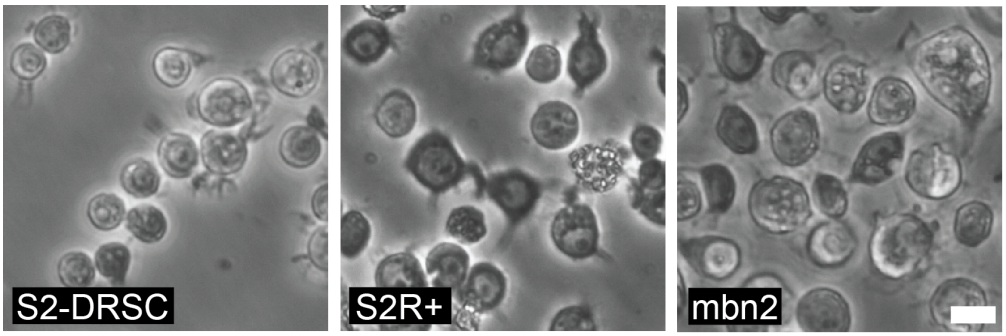

Figure 2.10: **Morphology of S2, S2R+, and mbn2 cell lines.** Phase-contrast micrographs of S2-DRSC (DGRC-181), S2R+ (DGRC-150), and mbn2 (DGRC-147). Scale bar is 10 microns.

The possibility that mbn2 is a divergent lineage of S2 is plausible given that both cell lines are thought to have a hemocyte-like cell type [82, 56, 144]. Furthermore, it is known that different lineages of *bona fide* S2 cells vary substantially in their morphology and gene expression, some of which share properties with mbn2 cells [210, 245, 56] (Figure 2.10). Under phase-contrast microscopy, canonical S2 cells represented by the S2-DRSC sub-line

are generally a mix of loosely adherent spherical cells and simple round flat cells. In contrast, live S2R+ cells can be characterized by many "phase dark" cells that attach to the growth substrate, which can flatten out to exhibit both polygonal and "fried egg" morphology. S2R+ cells that are loosely attached to the growth surface are generally spherical with fine cell protrusions. Like S2R+ cells, mbn2 cells are characterized by a mix of flattened phase dark cells that assume the polygonal and fried egg morphology, as well as loosely adhering spherical cells. However, loosely adherent mbn2 cells have a bigger diameter relative to S2-DRSC and S2R+ cells. Recognition of mbn2 as a potentially divergent S2 lineage suggests that there is more phenotypic diversity among different S2 lineages than previously recognized.

2.4.6 A subset of LTR retrotransposon families are sufficient to identify *Drosophila* cell lines

Our analysis has thus far provided evidence that TE insertion profiles of commonly used *Drosophila* cell lines based on whole-genome sequences can be used to cluster cell lines and uncover cases of cell line misidentification. However, for these results to form the foundation for a *Drosophila* cell line authentication protocol, it is necessary to show that a cell line sample can successfully be identified on the basis of its TE profile. Furthermore, it is important to explore if whole-genome data is required for TE-based cell line identification in *Drosophila* since the cost of WGS could preclude its routine application by many labs. Therefore, we next investigated whether a subset of *Drosophila* TE families could potentially be sufficient for *Drosophila* cell line identification, with the aim of guiding development of a cost-effective targeted PCR-based enrichment protocol that could be used more widely by the research community.

To investigate this possibility, we first clustered a non-redundant dataset of one "primary" replicate from each of the 22 *Drosophila* cell lines in the expanded dataset based on their whole-genome TE profiles (Figure 2.8A), which resulted in a similar clustering to the same sample of 22 cell lines including all replicates (Figure 2.5E). Replicates with the longest

read length or depth of coverage were chosen as the primary replicate in the non-redundant dataset (Table 2.4). We then took advantage of the ability of Dollo parsimony to reconstruct ancestral states and map the gain of TE insertions on each branch of the most parsimonious tree. TE insertions were then aggregated into families on each branch of the tree to visualize family- and branch-specific TE insertion profiles. This analysis revealed that a subset of 60 out of the 125 curated TE families in *D. melanogaster* are informative for *Drosophila* cell line clustering using TEMP predictions (Figure 2.8B). Within the set of clustering-informative TE families, we observed that some TE families are broadly represented across many cell lines with different origins (e.g., *copia*, *297*, *jockey*, *mdg3*, *mdg1*, and *roo*), although the quantitative abundance of these TE families varies across cell lines. Other TE families appear to be represented in only one cell line or a subset of cell lines from the same lab origin (e.g., *ZAM*, *Tabor*, *HMS-Beagle2*, *gypsy5*, *1731*, *17.6*, *springer*, *Tirant*, *rover*, and *micropia*). These results provide systematic genome-wide evidence for the classical observation that proliferation of different TE families in cultured *Drosophila* cells is cell-line dependent [70]. Additionally, these patterns of cell-line specific TE proliferation provide further support for the conclusions that the DGRC Sg4 cell line is a lineage of S3 cells (all share *Ivk* proliferation), and that mbn2 cell lines are a divergent lineage of S2 cells (all share *1731* proliferation) (Figure 2.8B).

Based on these results, we next evaluated whether a small, experimentally-tractable subset of TE families is sufficient to cluster and identify *Drosophila* cell lines. For this analysis, we focused on LTR retrotransposon families since this type of TE inserts with intact termini and therefore provide reliable 5' and 3' junctions for targeted PCR-based enrichment protocols [224]. We used the pattern of family- and branch-specific TE insertion to heuristically guide selection of a subset of six LTR retrotransposon families (*copia*, *297*, *mdg3*, *mdg1*, *roo*, and *1731*; TE family names highlighted in red in Fig 2.8B), which defined unique TE profiles for each cell line and generated the same major patterns of *Drosophila* cell line clustering as the genome-wide dataset of all 125 TE families (Figure 2.8C). Finally,

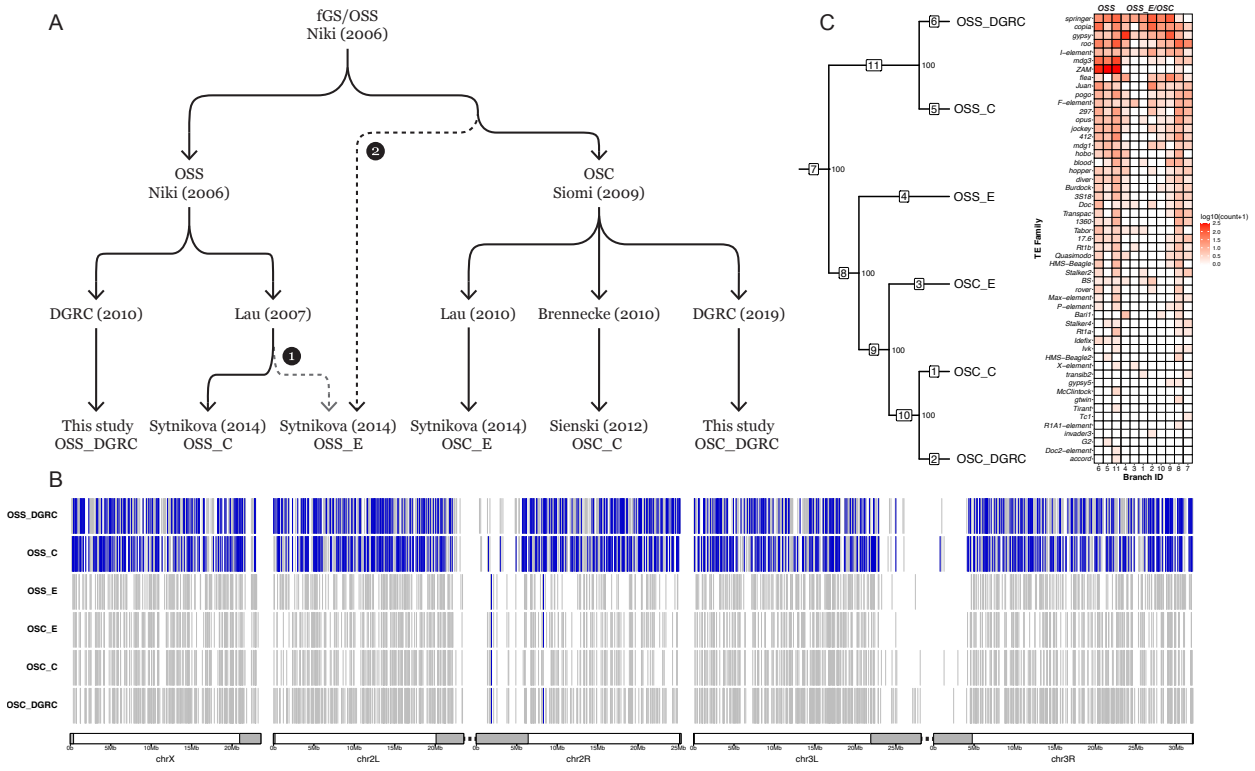

Figure 2.11: **ZAM proliferation reveals OSS cell line identity.** (A) Key events in the history of OSS and OSC cell line creation and distribution. Dotted lines represent alternative hypotheses for the identity of OSS_E. Branch 1 represents the reported provenance that hypothesizes OSS_E is an early diverging OSS sub-line; branch 2 hypothesizes that OSS_E approximates an ancestral state of the OSC cell line. (B) Genome-wide non-reference TE insertion data for six ovarian cell lines with *ZAM* insertions highlighted in blue and all other TE families in grey. (C) Dollo parsimony tree of ovarian cell lines based on all non-reference TE predictions. Numbers inside boxes on branches indicate branch ID, and numbers beside nodes indicate percent support based on 100 bootstrap replicates. (left). Heatmap showing the number of non-reference TE insertion gain events per family on each branch of the tree based on ancestral state reconstruction using Dollo parsimony. The heatmap is colorized by log-transformed (log10(count+1)) number of gains per family per branch, sorted top to bottom by overall non-reference TE insertion gains per family across all branches and sorted left to right into the *bona fide* OSS and OSS_E/OSC clusters (right).

we tested whether a cell line sample (not used in the tree construction) can be accurately identified on the basis of its six-family TE profile. To do this, we used the six-family TE tree derived from the non-redundant set of primary replicates as a backbone to constrain

Dollo parsimony searches including one additional "secondary" replicate for each of the 12 secondary replicates from the nine cell lines in the expanded dataset with secondary replicates. In 100% of cases (12/12), the additional secondary replicate clustered most closely with the primary replicate from the same cell line (Figure 2.12). In 10/12 cases, the bootstrap support for the clustering of replicates was 100%, and the remaining two cases (both for CME-W1-Cl.8+) had lower bootstraps (≥64%) presumably because of the short read length for these secondary replicates (50bp). This proof-of-principle analysis indicates that TE insertions from a small subset of LTR retrotransposon families can accurately identify *Drosophila* cell line samples, and that only a subset of "diagnostic" TE families are needed to develop a *Drosophila* cell line authentication protocol. Based on these results, we have developed and validated a PCR enrichment-based NGS protocol that generates genome-wide TE profiles using this subset of LTR retrotransposon families and can be used to authenticate *Drosophila* cell lines at lower cost than WGS analysis (D. Mariyappa, D.B. Rusch, S. Han, A. Luhur, D. Overton, D.F.B. Miller, C.M. Bergman, A.C. Zelhof; preprint in https://www.biorxiv.org/content/10.1101/2021.08.16.456580v1).

2.4.7 TE profiles provide insight into *Drosophila* ovarian cell line history

The observation that different TE families are amplified in distinct *Drosophila* cell lines raises the question of whether a single TE family could diagnostically mark the identity of a *Drosophila* cell line or sub-line. One such candidate for this possibility is the retroviral-like LTR retrotransposon *ZAM* in the closely related OSS and OSC ovarian somatic cell lines [177, 208]. As shown above, we observed a massive increase in *ZAM* insertions in OSS cells relative to the OSC cell line (branches 19 and 20 in Figure 2.8A, B), supporting previous findings by Sytnikova *et al.* 2014 [231]. However, Sytnikova *et al.* 2014 [231] also reported that *ZAM* amplification did not occur in all OSS sub-lines, only in a contemporary sub-line of OSS cells (called OSS_C), but not in a putatively early passage sub-line of OSS cells (called OSS_E).

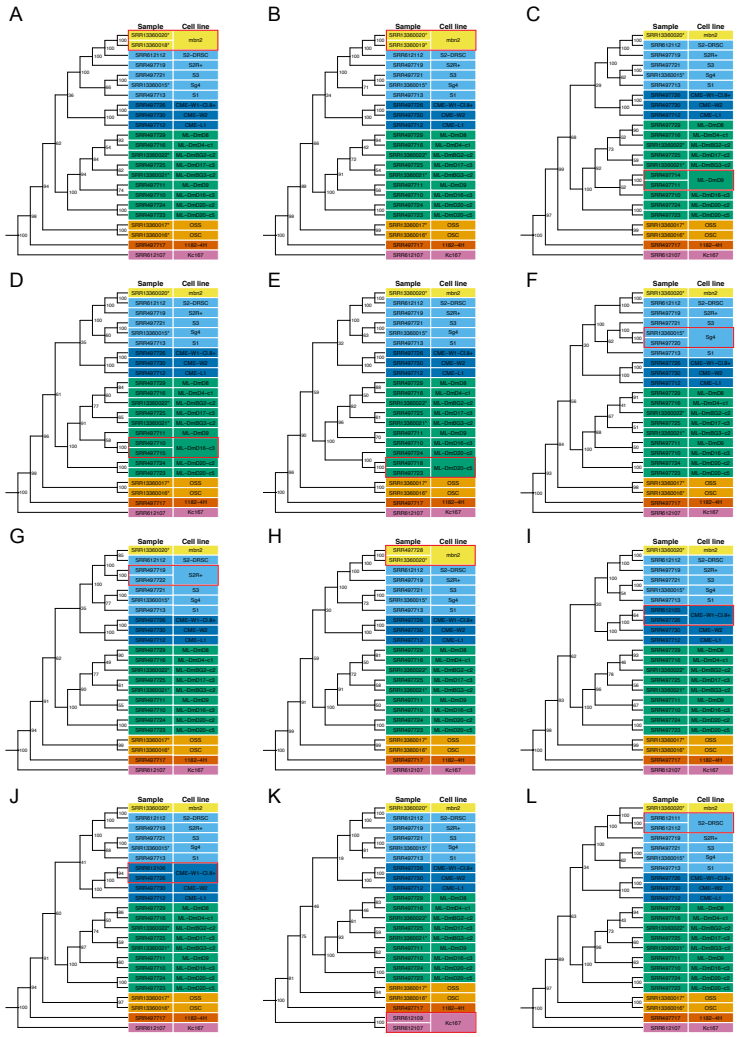

Figure 2.12: ***Drosophila* cell line samples can be identified using TE profiles from a diagnostic set of six LTR retrotransposon families.** Panels represent Dollo parsimony trees of a common set of 22 *Drosophila* cell line primary replicates plus one additional secondary replicate, one tree for each of the 12 secondary replicates from the nine cell lines in the expanded dataset with secondary replicates. Dollo parsimony trees were constructed using non-reference TE predictions for six *D. melanogaster* LTR retrotransposon families (*297*, *copia*, *mdg3*, *mdg1*, *roo* and *1731*). Samples are colorized by lab origin. Cell lines with secondary replicates are highlighted in red boxes. Node labels indicate support for each clade based on 100 bootstrap replicates. New sequence data from this study are indicated by asterisks.

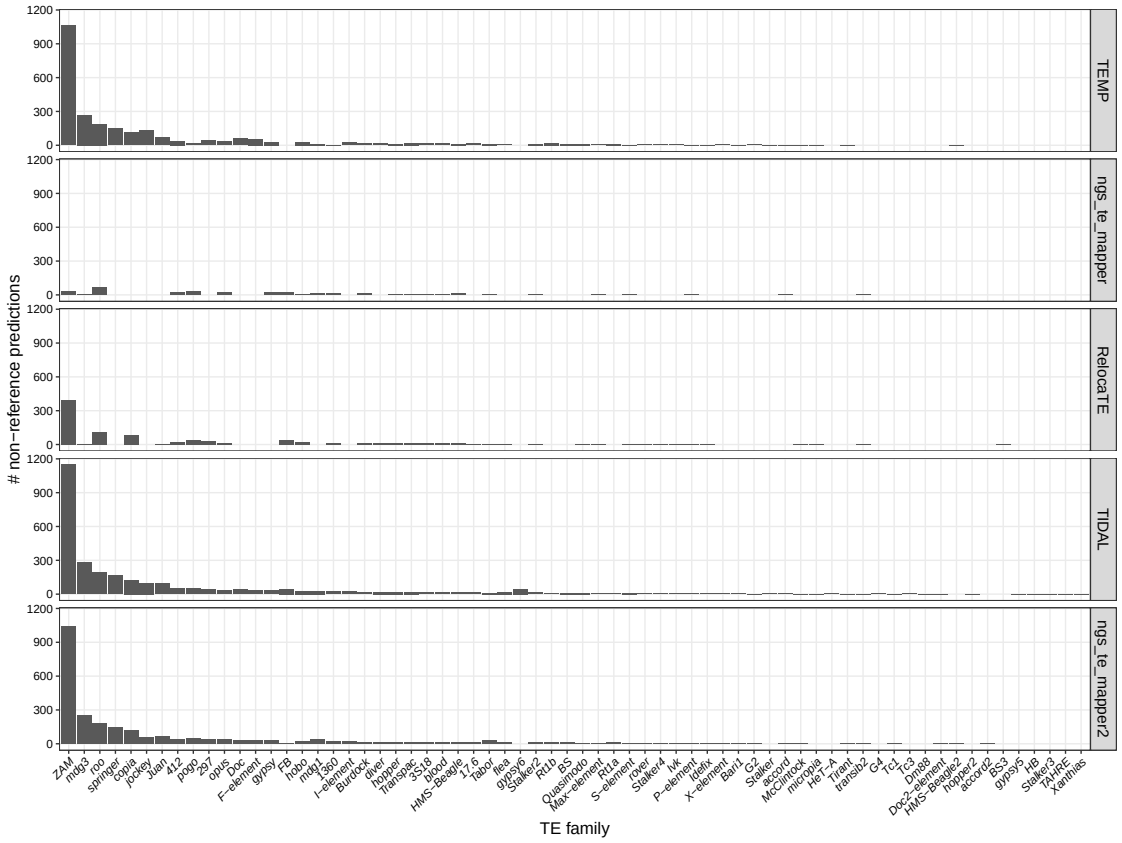

Figure 2.13: **Number of non-reference TE predictions on OSS_DGRC using five TE detection methods.** Paired-end sequencing data for OSS_DGRC was used as input for TEMP, ngs_te_mapper, RelocaTE, TIDAL and ngs_te_mapper2 to detect non-reference TE insertions using McClintock. *INE-1* and insertion predictions in low recombination regions were excluded from all panels.

To address whether *ZAM* proliferation is restricted to a subset of OSS sub-lines or is in fact a specific marker for all OSS sub-lines, we performed an integrated analysis of TE predictions in WGS data from six OSS and OSC samples from our and two previous studies [218, 231]. To formulate alternative hypotheses and guide interpretation of our results, we first compiled the reported provenance of these six OSS and OSC cell line samples. As shown in Figure 2.11A, the ultimate ancestor of all OSS and OSC cell lines is a cell line composed of germline and somatic ovarian cell types called fGS/OSS [177]. fGS/OSS cells

were subsequently selected in the Niki lab to remove germline-marked stem cells to create the ancestor of the OSS (ovarian somatic sheet) cell line. The Niki lab sent two batches of OSS cells to the Lau lab in 2007 (Nelson Lau, personal communication): one was expanded and continuously cultured to become the OSS_C sub-line; the other was briefly cultured and stored as a cryopreserved culture for many years, then thawed and sequenced in 2013 creating the OSS_E sample [231]. Our sample of OSS cells comes from an independent sub-line donated by the Niki lab to the DGRC in 2010 (OSS_DGRC). The Niki lab also sent fGS/OSS cells to the Siomi lab, who independently selected against germline cells to create another somatic cell line called OSC (ovarian somatic cells) [208]. OSC cells were sent by the Siomi lab in 2010 separately to the Lau (OSC_C) and Brennecke (OSC_E) labs, and were later donated by the Siomi lab to the DGRC in 2019 (OSC_DGRC).

Because WGS data from Sienski *et al.* 2012 [218] and Sytnikova *et al.* 2014 [231] is single-ended, integrated analysis of ovarian cell lines required a different TE prediction strategy than the one used for analysis of the paired-end datasets above. Preliminary analyses revealed that some single-end TE predictors (e.g., ngs_te_mapper and RelocaTE) [139, 199] severely under-predicted insertions specifically for the *ZAM* family in the DGRC OSS sample relative to TEMP results based on paired-end data (Figure 2.13). Additionally, our analysis of OSS and OSC samples ultimately required tracking intra-sample TE allele frequencies, which is not available in other TE predictors that use single-end data (e.g., TIDAL) [193]. Thus, we developed a new implementation of the single-end TE predictor originally described in in Linheiro *et al.* 2012 [139] called ngs_te_mapper2 (`https://github.com/bergmanlab/ngs_te_mapper2`) that improves speed and sensitivity relative to the original version and has been extended to estimate intra-sample TE allele frequencies (Figure 2.1; Table 2.1 and 2.2; see Section 2.4.1 and 2.4.2 for details).

Using datasets normalized to the same read length and coverage in order to optimize resolution of closely related sub-lines, we predicted non-reference TE insertions in all OSS and OSC sub-lines with ngs_te_mapper2. These results revealed that *ZAM* has proliferated

massively in the OSS_DGRC and OSS_C sub-lines (553 and 630 copies, respectively, in euchromatic regions), but is present in only one or two copies in OSS_E and all OSC sub-lines (Figure 2.11B). The abundance of *ZAM* in these ovarian cell lines is more than 10-fold higher than fly strains where *ZAM* has been mobilized because of deletions in the *flamenco* piRNA locus [128, 252] or because of multigenerational knockdown of the piRNA effector protein *piwi* [18, 167].

Under the "reported provenance" hypothesis that OSS_E and OSS_C share a more recent common ancestor than they do with OSS_DGRC (Fig 2.11A, branch 1), this pattern of *ZAM* abundance can only be explained by unlikely scenarios such as a massive loss of *ZAM* insertions on the branch leading to OSS_E, or independent parallel amplifications of *ZAM* on the OSS_C and OSS_DGRC sub-lines. An alternative hypothesis to explain the pattern of *ZAM* abundance is motivated by another observation made by Sytnikova *et al.* 2014 that OSS_E shares more TE insertions in common with OSC sub-lines (OSC_E and OSC_C) than it does with a contemporary OSS sub-line (OSS_C) [231]. This pattern is not expected under the reported provenance hypothesis and suggests that OSS_E may in fact be an OSC-like lineage, rather than an early passage OSS sub-line. Under this alternative "uncertain provenance" hypothesis (Fig 2.11A, branch 2), the only *bona fide* OSS sub-lines would be OSS_C and OSS_DGRC, and *ZAM* proliferation could truly be a diagnostic marker of OSS cell line identity.

To test these alternative hypotheses, we used ngs_te_mapper2 predictions as input to cluster OSS and OSC sub-lines using Dollo parsimony. We found two highly supported clusters, one containing only the OSS_C plus OSS_DGRC sub-lines and the other containing OSS_E plus all OSC sub-lines (Figure 2.11C). Ancestral state reconstruction clearly demonstrated that high *ZAM* abundance is restricted to the cluster containing OSS_C and OSS_DGRC sub-lines. The only two *ZAM* insertions that are found in OSS_E and OSC sub-lines are both shared by multiple sub-lines and therefore likely inserted in a common ancestor of the entire clade (Figure 2.11B). We verified that the clustering relationships among OSS

and OSC sub-lines were not solely driven by the ZAM amplification by repeating our clustering analysis excluding *ZAM* insertions, obtaining the same topology as in the complete dataset (Figure 2.14A).

Further support for the hypothesis that OSS_E is an OSC-like lineage can be found in patterns of SNP and CNV variation in these cell line genomes (Figure 2.14B, C). OSS_C and OSS_DGRC have essentially identical BAF profiles across the entire genome (Figure 2.14B). In contrast, OSS_E and OSC sub-lines share a BAF profile everywhere but the distal regions on chromosome arms 2L, 3L and 3R (Figure 2.14B; Figure 2.15A). BAF profiles on all of chromosome X and arm 2R clearly differentiate OSS_C and OSS_DGRC (heterozygous) from OSS_E and OSC sub-lines (homozygous) (Figure 2.14B). Likewise, CNV profiles support the clustering of OSS_C with OSS_DGRC and OSS_E with the OSC sub-lines. OSS_C and OSS_DGRC share a large deletion on chromosome X not found in OSS_E plus OSC sub-lines, and OSS_E plus the OSC sub-lines share a smaller deletion on chromosome arm 3L not found in OSS_C or OSS_DGRC (Figure 2.14C). Based on these results, we conclude that OSS_E is a divergent lineage of OSC cells rather than early passage OSS cells, that *ZAM* amplification truly marks *bona fide* OSS cell lines (include the OSS line distributed by the DGRC), and that ngs_te_mapper2 TE predictions based on single-end WGS data can be effectively used to cluster *Drosophila* cell lines and reveal aspects of cell line history.

2.4.8 Loss of heterozygosity impacts TE profiles in *Drosophila* cell culture

Re-interpreting OSS_E as a divergent lineage of OSC cells requires explaining both the similarity and distinctness of its TE, BAF and CNV profiles from other OSC sub-lines. Two observations led us to hypothesize that OSS_E approximates an ancestral state of current OSC sub-lines. First, OSS_E occupies a basal position in the OSS_E plus OSC cluster based on TE profiles (Figure 2.11C). Second, the BAF profile for OSS_E shows heterozygosity that extends in the distal regions of chromosome arms 2L, 3L and 3R relative to OSC sub-lines (Figure 2.15A, green shading). We propose that differences in BAF profiles in these distal

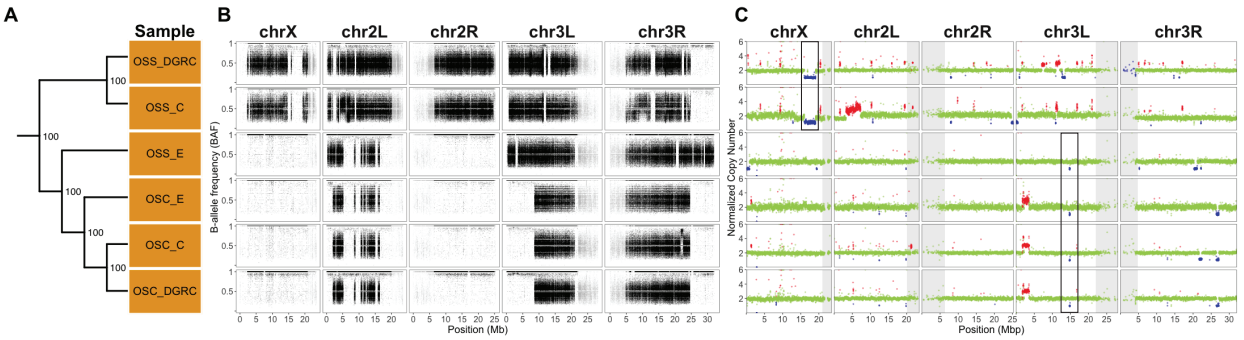

Figure 2.14: **Copy number and B-allele frequency profiles for six ovarian cell line samples.** (A) Dollo parsimony tree of six ovarian cell line samples based on non-reference TE predictions excluding *ZAM* insertions using single-end WGS data. Node labels indicate support for each clade based on 100 bootstrap replicates. (B) B-allele frequency profiles for ovarian cell line samples on major chromosome arms. For a given SNP, the B-allele frequency (BAF) was determined as the coverage of reads supporting non-reference allele divided by total coverage at that position. SNPs in low recombination regions are plotted in grey. (C) Copy number profiles for ovarian cell line samples on major chromosome arms. Each data point represents normalized copy number (ratio*ploidy) for a given 10kb window estimated by Control-FREEC [37]. Data points for each window are colorized by CNV status (red: CNV gain; green: no CNV; blue: CNV loss), which are based on the comparison between normalized copy number for that window and baseline ploidy for the chromosome arm. Black boxes in panel C highlight regions where cell lines share the same copy number loss events that are not shared in other cell samples. Low recombination regions are shaded in grey.

regions are caused by loss of heterozygosity (LOH) that occurred in an ancestor of all OSC sub-lines after divergence from the lineage leading to OSS_E. We infer that these large-scale distal LOH events were caused by a copy-neutral process such as mitotic recombination, since the baseline copy number in these distal LOH regions is the same in OSS_E and OSC sub-lines (Figure 2.15B; Figure 2.14C). Similar to previous reports in a primate cell line [182], we do observe smaller copy number gain and loss events, respectively, within the large regions of LOH on chromosome arms 2L and 3R. However, these copy number events are much smaller and fully contained within the larger LOH regions and therefore unlikely to be the cause of the large-scale distal LOH events. Despite previous evidence for putatively polyploid cells in the OSC_E sub-line [231], we infer a diploid baseline copy number in the "stem line" that

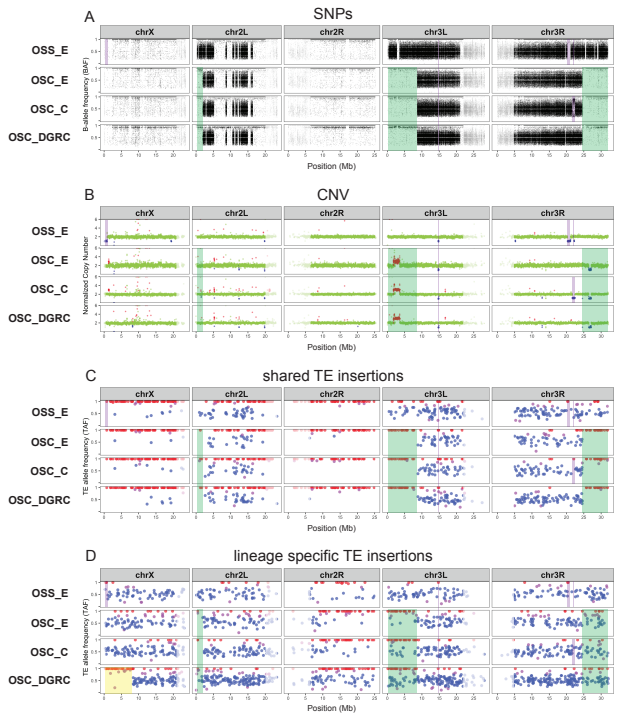

Figure 2.15: **Loss of heterozygosity, copy number evolution and ongoing transposition shape TE profiles in *Drosophila* ovarian somatic cell lines.** Genome-wide profiles for OSS_E and OSC sub-lines of (A) intra-sample allele frequency based on SNP variants, (B) copy number, (C) intra-sample allele frequency based on TE insertions shared by OSS_E and OSC sub-lines, and (D) intra-sample allele frequency based on lineage specific TE insertions restricted to only OSS_E or the OSC sub-lines. SNPs and TE insertions in highly-repetitive low recombination regions are shaded in grey. For SNP profiles, the B-allele frequency (BAF) was determined as the coverage of reads supporting the non-reference allele divided by total coverage at that variant positions; regions of heterozygosity in a diploid genome are shown in BAF profiles where clusters of SNPs have allele frequencies centered around 0.5. For copy number profiles, each data point represents normalized copy number (ratio*ploidy) for a given 10kb window estimated by Control-FREEC [37]; data points for each window are colorized by CNV status (red: CNV gain; green: no CNV; blue: CNV loss), which are based on the comparison between normalized copy number for that window and a baseline ploidy of 2n. For TE profiles, TE insertions are classified as being homozygous (red), heterozygous (blue), or undefined (purple) based on intra-sample allele frequencies estimated by ngs_te_mapper2. Green shading indicates copy-neutral LOH regions defined by more extensive patterns of SNP heterozygosity in OSS_E relative to OSC sub-lines that are putatively caused by mitotic recombination. Yellow shading indicates copy-neutral LOH regions based on runs of homozygous TE insertions in OSC_DGRC relative to other OSC sub-lines that are putatively caused by mitotic recombination. Purple shading indicates LOH regions that are putatively caused by segmental deletion.

leads to the majority of inherited cells in the OSS_E and OSC sub-lines, since BAF profiles (Figure 2.15A) and DNA density profiles (Figure 2.16) of the bulk samples show modal values that together are consistent with diploidy but not any higher ploidy values.

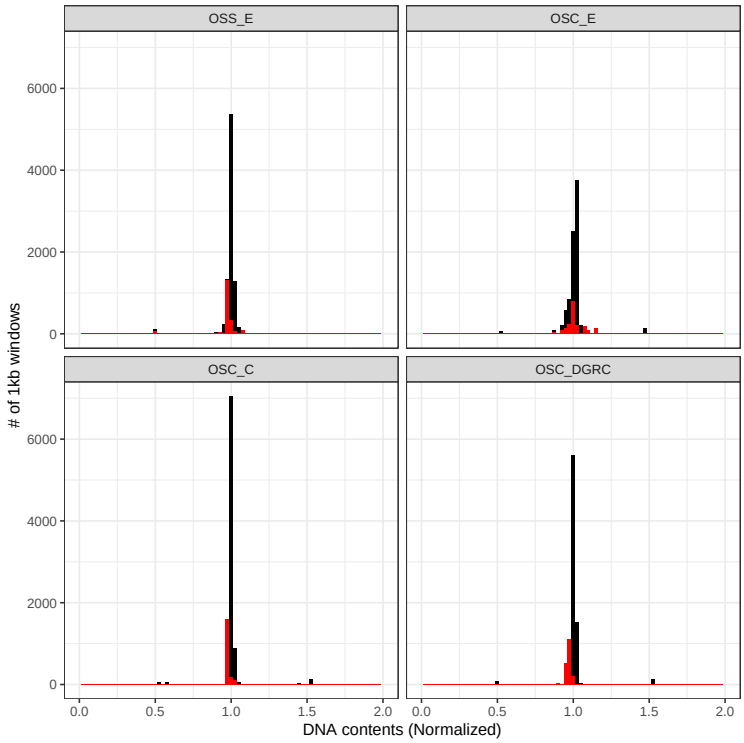

Figure 2.16: **Normalized DNA content for OSS_E and OSC cell lines.** Histograms of normalized DNA read density of 1kb windows using the method described in [129]. Reads mapping to chromosome X are shown in red. Reads mapping to autosomes are shown in black. Peaks at 0.5, 1.0 and 1.5 are consistent with a diploid base copy number for the OSS_E/OSC lineage.

If this evolutionary scenario is correct, shared TEs (which inserted prior to the divergence of OSS_E and OSC sub-lines) that are heterozygous in OSS_E are predicted to be homozygous in OSC sub-lines in distal LOH regions, but should maintain heterozygosity elsewhere in the genome. To test these predictions, we used intra-sample allele frequency estimates from ngs_te_mapper2 to classify the zygosity of TE insertions in OSS_E and OSC sub-lines. Evaluation of our classifier on simulated diploid genomes revealed it had >91% precision and crucially never falsely classified heterozygous insertions as homozygous (Table

2.3), and is thus conservative with respect to detection of LOH using TE insertions. As predicted under our model, we observed that there are many shared TE insertions in distal LOH regions that are heterozygous in OSS_E but virtually all TE insertions in these regions are homozygous in OSC sub-lines (Figure 2.15C, green shading). Outside of distal LOH regions, shared TE insertions that are heterozygous in OSS_E generally retain heterozygosity in OSC sub-lines (Figure 2.15C). In contrast, we observe that many lineage-specific TE insertions (which occurred after the divergence of OSS_E and OSC sub-lines) are heterozygous in OSC sub-lines in distal LOH regions (Figure 2.15D, green shading). Together these results support the inferences that OSS_E approximates an ancestral state of current OSC sub-lines, that LOH events can cause fixation of previously heterozygous TE insertions in *Drosophila* cell lines, and that ongoing transposition in *Drosophila* cell culture can restore genetic variation in regions where previous large-scale LOH events have eliminated ancestral SNP or TE insertion variation.

Contrasting patterns of genetic variation between OSS_E and OSC sub-lines in distal regions of chromosome arms 2L, 3L and 3R provided the initial evidence for LOH as mechanism of genome evolution in *Drosophila* cell culture. Additional evidence for copy-neutral LOH in *Drosophila* cell culture can be found in the lack of SNP heterozygosity on all of chromosome X and arm 2R (Figure 2.15A), which can be explained by whole-arm LOH events caused by centromere-proximal somatic recombination events in the common ancestor of the OSS_E/OSC lineage, assuming that the genome-wide heterozygosity observed in *bona fide* OSS sub-lines is ancestral (Figure 2.14B). Consistent with the prediction of whole-arm LOH in the ancestor of the OSS_E/OSC lineage followed by ongoing transposition in cell culture, we observe that most shared TE insertions on chromosome X and arm 2R are homozygous (Figure 2.15C), while lineage-specific TE insertions are heterozygous (Figure 2.15D). Intriguingly, and in contrast to other OSC sub-lines, we also observe that lineage-specific TE insertions on the distal eight megabases of chromosome X in OSC_DGRC are almost all homozygous (Figure 2.15D, yellow shading). This observation can be explained by a sec-

ondary copy-neutral LOH event in the distal region of chromosome X that occurred recently only in the OSC_DGRC lineage. In this case, heterozygosity restored by ongoing TE insertion in *Drosophila* cell culture allows detection of a subsequent LOH events in the same genomic region that cannot be detected using SNP variation.

In addition to large-scale LOH events affecting distal regions or whole chromosome arms that can be explained by copy-neutral processes such as mitotic recombination, we also observed smaller scale LOH events that can be explained by hemizygosity due to segmental deletion (purple shading in Figure 2.15A, B). For example, we observe a 200kb region on chromosome arm 3L in all OSS_E and OSC sub-lines that lacks heterozygous SNPs which can be explained by a segmental deletion that that occurred in the common ancestor of the OSS_E/OSC lineage (Figure 2.15; Figure 2.17). LOH by segmental deletion is supported by shared TEs in this region being homozygous in all OSS_E and OSC sub-lines. Likewise, in OSS_E, we observe two sub-line specific segmental deletions on chromosome arm 3R of 900kb and 100kb, respectively, that lack heterozygous SNPs and TE insertions in the corresponding regions (Figure 2.15; Figure 2.17). We also observe a sub-line specific segmental deletion on chromosome arm 3R of 800kb in OSC_C that exhibits a BAF profile enriched at 0.85 (Figure 2.15; Figure 2.17) rather than the homozygosity expected for complete LOH due to hemizygosity. Similar to patterns of LOH in mammalian tumors that have incomplete purity [226], we interpret the incomplete LOH in this region as being caused by clonal heterogeneity in the OSC_C sub-line, with the majority of cells having the segmental deletion but a small proportion of cells lacking it. If this hypothesis is correct, the median copy number should be slightly over 1 in the segmentally deleted LOH region in OSC_C: as predicted, the median copy number of OSC_C in the putative LOH region is 1.14 (Figure 2.17). Additionally, the OSS_E sub-line also exhibits a terminal deletion on the tip of chromosome X (Figure 2.15; Figure 2.17), which does not lead to LOH in the SNP profile because of the primary whole-arm LOH event proposed to have occurred in the ancestor of the OSS_E/OSC lineage. However, similar to the secondary LOH event proposed to have occurred by somatic recom-

bination on the distal region of chromosome X in OSC_DGRC, recovery of heterozygosity by ongoing TE insertion allows secondary LOH by segmental deletion to be observed in the lineage-specific TAF profile at the tip of chromosome X in OSS_E. Finally, we note that LOH events that can be explained by segmental deletions provide further support for a diploid stem line in the OSS_E/OSC lineage, since hemizygosity in a diploid is more parsimonious than scenarios such as multiple identical deletions or deletions followed by mitotic recombination required to explain LOH by segmental deletion in genomes with higher ploidies.

As LOH has not previously been reported as a mechanism of genome evolution in *Drosophila* cell culture, we sought to find additional evidence for this process by inspecting BAF profiles for other *Drosophila* cell lines in the expanded dataset. This led us to additional evidence for large-scale LOH events defined by SNPs on chromosome arms 2R and 3L of the CME-W2 and CME-W1-Cl.8+ cell lines (Figure 2.6B; Figure 2.18A), both of which are reported to have a diploid baseline autosomal copy number [129]. As with OSS_E, we propose that the more extensive heterozygous BAF profile on these chromosome arms in CME-W2 represents the pre-LOH ancestral-like state, and the homozygous BAF profile of CME-W1-Cl.8+ represents the post-LOH derived state. This scenario is consistent with the reported establishment of CME-W1-Cl.8+ from a single cloned cell of a polyclonal cell line (CME-W1) with the same ancestral genotype as CME-W2 [61, 184]. The lack of difference in the baseline copy number profiles on chromosome arms 2R and 3L of CME-W2 and CME-W1-Cl.8+ suggests these large-scale LOH events were also due to mitotic recombination (Figure 2.18B). As predicted under the LOH model, we observed many TE insertions shared by CME-W2 and CME-W1-Cl.8+ are heterozygous in CME-W2 but are nearly all homozygous in CME-W1-Cl.8+ in LOH regions (Figure 2.18C). Like in OSC sub-lines, we also observed many heterozygous TE insertions that are specific to CME-W1-Cl.8+ in LOH regions (Figure 2.18D), consistent with recovery of TE insertion variation after LOH. Similar to the OSS_E/OSC lineage, we also find evidence in the CME-W2/CME-W1-Cl.8+ lineage for smaller scale LOH events on chromosome arm 3R that can be explained by segmental

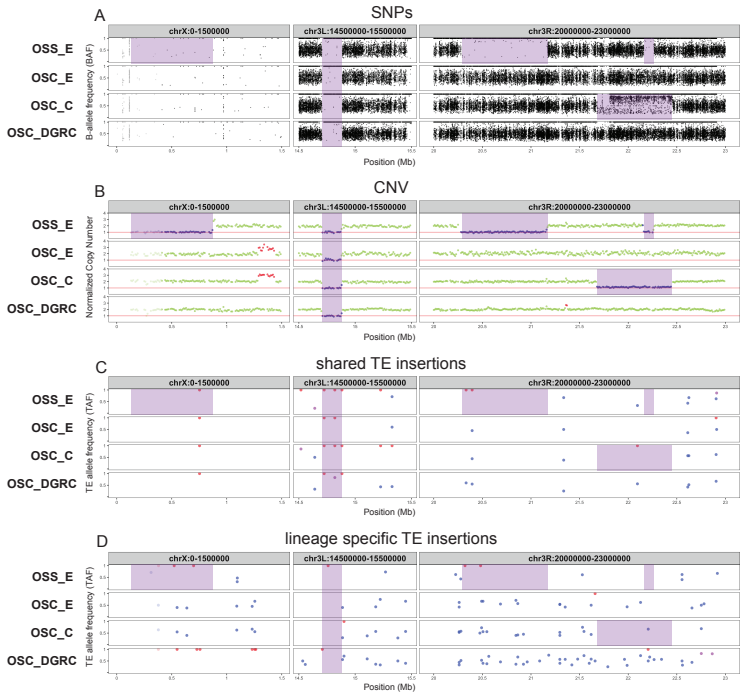

Figure 2.17: **Patterns of genomic variation in regions with loss of heterozygosity putatively caused by segmental deletion in *Drosophila* ovarian somatic cell lines.** Genome-wide profiles for OSS_E and OSC sub-lines of (A) intra-sample allele frequency based on SNP variants, (B) copy number, (C) intra-sample allele frequency based on TE insertions shared by OSS_E and OSC sub-lines, and (D) intra-sample allele frequency based on lineage specific TE insertions restricted to only OSS_E or the OSC sub-lines. For SNP profiles, the B-allele frequency (BAF) was determined as the coverage of reads supporting the non-reference allele divided by total coverage at that variant positions; regions of heterozygosity in a diploid genome are shown in BAF profiles where clusters of SNPs have allele frequencies centered around 0.5. For copy number profiles, each data point represents normalized copy number (ratio*ploidy) for a given 10kb window estimated by Control-FREEC [37]; data points for each window are colorized by CNV status (red: CNV gain; green: no CNV; blue: CNV loss), which are based on the comparison between normalized copy number for that window and baseline ploidy for the chromosome arm. For TE profiles, TE insertions are classified as being homozygous (red), heterozygous (blue), or undefined (purple) based on intra-sample allele frequencies estimated by ngs_te_mapper2. Purple shading indicates LOH regions that are putatively caused by segmental deletion.

deletion (Figure 2.18; Figure 2.19), with clonal heterogeneity explaining incomplete LOH by segmental deletion at the tip of chromosome arm 3R in CME-W2. Finding evidence for

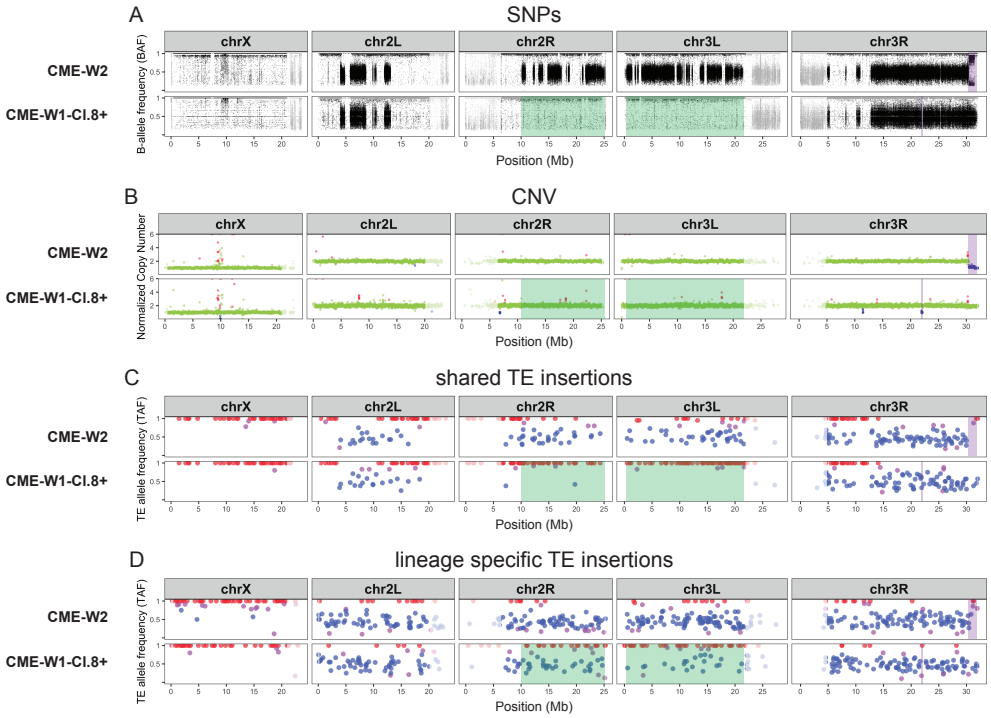

Figure 2.18: **Loss of heterozygosity, copy number evolution and ongoing transposition shape TE profiles in *Drosophila* imaginal disc derived cell lines.** Allele frequency profiles for CME-W2 and CME-W1-Cl.8+ cell lines based on (A) SNP variants, (B) copy number, (C) intra-sample allele frequency based on TE insertions shared by CME-W2 and CME-W1-Cl.8+, and (D) intra-sample allele frequency based on lineage specific TE insertions restricted to only CME-W2 or CME-W1-Cl.8+. SNPs and TE insertions in highly-repetitive low recombination regions are shaded in grey. For SNP profiles, the B-allele frequency (BAF) was determined as the coverage of reads supporting the non-reference allele divided by total coverage at that variant positions; regions of heterozygosity in a diploid genome are shown in BAF profiles where clusters of SNPs have allele frequencies centered around 0.5. For copy number profiles, each data point represents normalized copy number (ratio*ploidy) for a given 10kb window estimated by Control-FREEC [37]; data points for each window are colorized by CNV status (red: CNV gain; green: no CNV; blue: CNV loss), which are based on the comparison between normalized copy number for that window and baseline ploidy for the chromosome arm. For TE profiles, TE insertions are classified as being homozygous (red), heterozygous (blue), or undefined (purple) based on intra-sample allele frequencies estimated by ngs_te_mapper2. Green shading indicates LOH regions that are putatively caused by mitotic recombination defined by the more extensive pattern of SNP heterozygosity in CME-W2 relative to CME-W1-Cl.8+. Purple shading indicates LOH regions that are putatively caused by segmental deletion.

both mechanisms of LOH in distinct cell lineages developed in different labs generalizes the inference that LOH shapes TE profiles in *Drosophila* cell lines, and suggests that LOH in *Drosophila* culture is not dependent on the genetic background of ancestral fly donor.

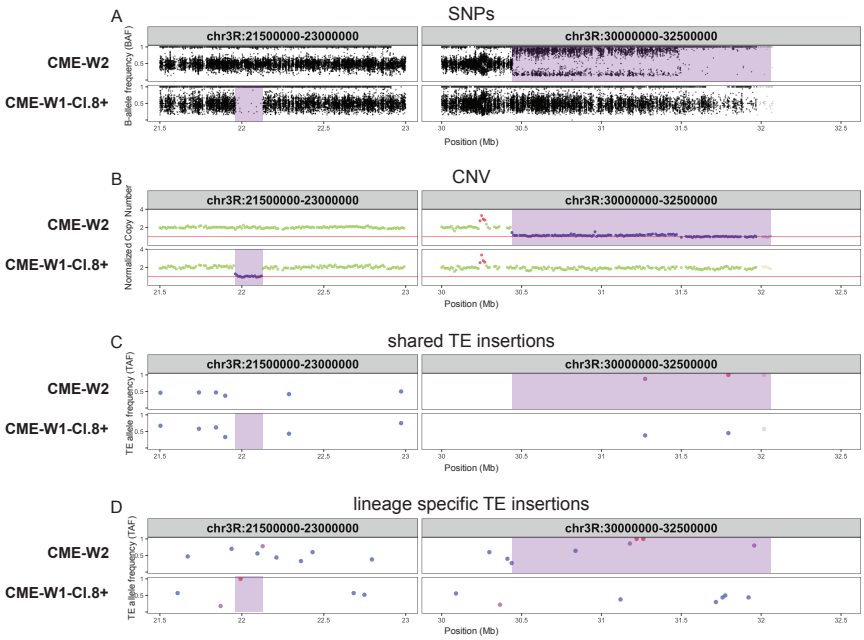

Figure 2.19: **Patterns of genomic variation in regions with loss of heterozygosity putatively caused by segmental deletion in *Drosophila* imaginal disc derived cell lines.** Genome-wide profiles for CME-W2 and CME-W1-Cl.8+ of (A) intra-sample allele frequency based on SNP variants, (B) copy number, (C) intra-sample allele frequency based on TE insertions shared by CME-W2 and CME-W1-Cl.8+, and (D) intra-sample allele frequency based on lineage specific TE insertions restricted to only CME-W2 or CME-W1-Cl.8+. For SNP profiles, the B-allele frequency (BAF) was determined as the coverage of reads supporting the non-reference allele divided by total coverage at that variant positions; regions of heterozygosity in a diploid genome are shown in BAF profiles where clusters of SNPs have allele frequencies centered around 0.5. For copy number profiles, each data point represents normalized copy number (ratio*ploidy) for a given 10kb window estimated by Control-FREEC [37]; data points for each window are colorized by CNV status (red: CNV gain; green: no CNV; blue: CNV loss), which are based on the comparison between normalized copy number for that window and baseline ploidy for the chromosome arm. For TE profiles, TE insertions are classified as being homozygous (red), heterozygous (blue), or undefined (purple) based on intra-sample allele frequencies estimated by ngs_te_mapper2. Purple shading indicates LOH regions that are putatively caused by segmental deletion.

2.5 Conclusions

Here we demonstrate that TE insertion profiles can successfully identify *Drosophila* cell lines and use this finding to clarify several aspects of cell line provenance in *Drosophila*. The success of this approach validates our basic model for how the joint processes of germline transposition in whole flies and somatic transposition in cell culture create TE profiles that uniquely mark *Drosophila* cell lines (Figure 2.2). We also show that TE insertion profiles can shed light on the evolutionary history of *Drosophila* cell lines derived from a common ancestral cell line, and that LOH is an additional mechanism of genome evolution in cell culture that adds complexity to our basic model (Figure 2.20). During cell culture, LOH resulting from mitotic recombination (Figure 2.20, green shading) or segmental deletion (Figure 2.20, purple shading) purges ancestral variation and causes previously heterozygous SNPs and TE insertions to become fixed or lost within a cell line genome. Ongoing transposition in cell culture leads to the relatively rapid recovery of TE but not SNP heterozygosity, allowing secondary LOH events to be identified using TE insertions in regions that have previous lost ancestral variation due to primary LOH events (Figure 2.20, yellow shading). The emerging model of TE evolution in cell culture motivated by results presented here has direct implications for the development of protocols for cell line identification in *Drosophila* and contributes to our general understanding of the mechanisms of genome evolution in cell lines derived from multicellular organisms.

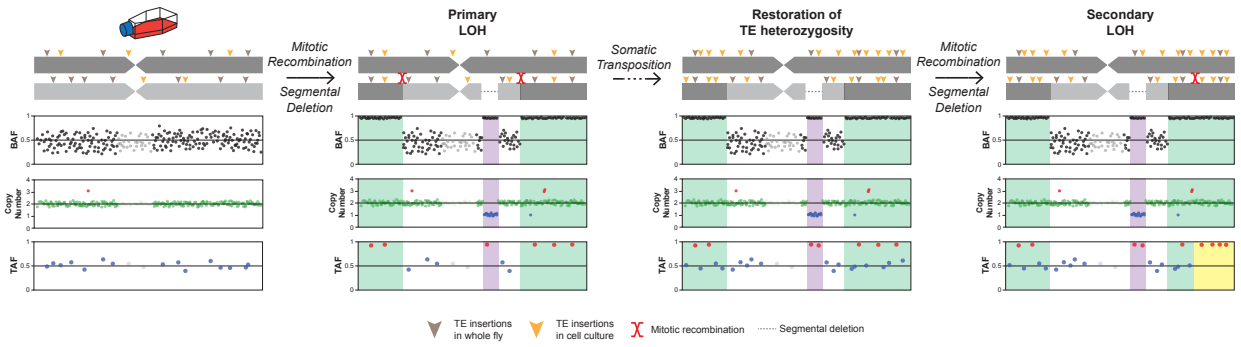

Figure 2.20: **Schematic model of how loss of heterozygosity and somatic transposition interact to shape TE profiles in diploid *Drosophila* cell line genomes.** Mitotic recombination (green shading) or segmental deletion (purple shading) can cause LOH of pre-existing heterozygous SNP and TE variants, as revealed by changes in B-allele frequency (BAF) and TE-allele frequency (TAF) profiles. Ongoing transposition in cell culture leads to accumulation of new haplotype-specific heterozygous TE insertions inside and outside of primary LOH regions. Recovery of TE heterozygosity allows detection of secondary LOH events (yellow shading) in regions of the genome that have previously undergone primary LOH events. Secondary LOH can occur by either mitotic recombination or segmental deletion, but only mitotic recombination is depicted here. We note that this model depicts a simplified case of diploidy which applies to some cell lines such as the OSS_E/OSC and CME-W2/CME-W1-Cl.8+ lineages, however many cell culture genomes can have complex genome structure due to polyploidy and aneuploidy.

Chapter 3

Evidence for ongoing transposition during long-term *Drosophila* cell culture[1]

[1]Shunhua Han, Guilherme B Dias, Preston J Basting, Michael G Nelson, Sanjai Patel, Mar M Marzo, and Casey M Bergman, Evidence for ongoing transposition during long-term *Drosophila* cell culture.

To be submitted to *Genetics*.

3.1 ABSTRACT

Cultured cells are widely used in molecular biology, although we know little about how cell line genomes change over time. Previous work has shown that *Drosophila* cultured cells have a higher transposable element (TE) content than whole flies, but whether this increase in TE content resulted from an initial burst of transposition during cell line establishment or ongoing transposition in cell culture remains unknown. Here we sequence the genomes of 28 sub-lines of *Drosophila* S2 cells and show that TE insertions provide abundant markers for the reconstruction of phylogenetic relationships of sub-lines in a model animal cell culture system. Analysis of DNA copy number evolution across S2 sub-lines revealed dramatically different patterns of genome organization that support the overall evolutionary history reconstructed using TE insertions. Ancestral state reconstruction of TE insertions on the S2 phylogeny support a model of ongoing TE insertion dominated by episodic activity of a small number of retrotransposon families. Our work demonstrates extensive TE insertion and DNA copy number diversity among S2 sub-lines that may impact the reproducibility of cell culture experiments that do not control for sub-line identity. Moreover, our work reveals that ongoing transposition in cell culture leads to useful genomic markers that can be used to verify sub-line identity in S2 cells and possibly other cell culture systems.

3.2 INTRODUCTION

Animal cell lines play vital roles in biology by providing an abundant source of material to study molecular processes or as cellular factories to express important biomolecules. Like all living systems, animal cell lines undergo genomic changes during routine propagation [205], leading to genetic diversity across time and laboratories that can lead to irreproducible research outcomes [98]. Despite the current emphasis on reducing sources of irreproducibility in biological research, relatively little attention has been paid to understand the pattern and process of *in vitro* evolution that leads to genomic diversity among sub-lines of long-

term cell cultures [105, 67, 24, 141], or how to identify and minimize the impact of such diversity [98, 24]. Establishing general rules for cell culture genome evolution and mitigating its influence will likely require analysis of multiple cell lines from many different species since the pattern and process of genome evolution is known to vary across taxa [145].

Early studies in the model insect *Drosophila melanogaster* showed a high abundance of multiple transposable element (TE) families in cell lines relative to the genomes of whole flies [189]. Recent study using next-generation sequencing (NGS) technique showed that between ∼800 to ∼3000 non-reference TE insertions can be detected among *Drosophila* cell lines, with LTR retrotransposons making up the bulk of these new insertions [193]. In comparison, only ∼200 to ∼1400 non-reference TE insertions can be detected among *Drosophila* lab strains [193]. Proliferation of TEs in *Drosophila* cultured cell genomes could be explained by a burst of transposition during initial establishment of cell lines, by ongoing TE insertion during routine cell culture, or a combination of both processes [70]. Di Franco *et al*, 1992 [67] contrasted the stability of TE profiles among sub-lines of one of the oldest *Drosophila* cell lines (Kc) [105] with elevated TE abundance in a newly-established cell line (inb-c) and concluded that the increased TE abundance in *Drosophila* cell lines resulted from an initial burst of transposition during the establishment of a new cell line, with relative stasis thereafter. However, comparison of old and new cultures from different cell lines is not a definitive test of whether ongoing TE proliferation occurs during routine culture because of differences in the founder genotypes and cell type of independently established cell lines. More recently, Sytnikova *et al*, 2014 [231] provided evidence for transposition after initial cell line establishment in *Drosophila* by showing an increase in abundance of the *ZAM* element in a continuously cultured sub-line of the OSS cell line (OSS_C) relative to a putative frozen progenitor sub-line (OSS_E). More recent work [88] suggests that the early version of the OSS reported in Sytnikova *et al*, 2014 (OSS_E) [231] is actually a mislabeled version of a related cell line (OSC) and thus it is unclear if the *ZAM* activation in OSS occurred during or after the establishment of the *bona fide* OSS lineage. Documenting whether ongoing transposition

in cell culture occurs is important since this process can lead to genomic variation among sub-lines that could impact functional studies and, more practically, provide useful markers for cell line identification and reconstruction of cell line evolutionary history.

Here we contribute to the understanding of genome evolution in long-term animal cell culture using a large sample of sub-lines of *Drosophila* Schneider Line 2 (S2) cells, one of the most widely-used non-mammalian cell culture systems. S2 cells were established from embryonic tissue of an unmarked stock of Oregon-R flies in December 1969 [213] and are likely to be derived from macrophage-like hemocytes [213, 70]. Two other cell lines, S1 (August 1969) and S3 (February 1970), were derived from the same ancestral fly stock [213] that can be used as outgroups to analyze evolution in the S2 lineage. Since their establishment, S2 cells have been distributed widely and grown more extensively than S1 or S3 cells [129]. Many different sub-lines of S2 cells have been established by labs in the *Drosophila* community, some of which have been donated back to the *Drosophila* Genomics Resource Center (DGRC) for maintenance and distribution. In general, the provenance and relationships among sub-lines of S2 cells are unknown, as is the extent of their genomic or phenotypic diversity. At least one sub-type of S2 cells, called S2R+, is known to have distinct phenotypes from other S2 cell lines such as expressing the Dfrizzled-1 and Dfrizzled-2 membrane proteins and having the desirable property of being more adherent to surfaces in tissue culture [245]. In addition to their ubiquity and diversity, S2 cells are a good model to study genome evolution in animal cell culture because of the ability to perform cost-effective whole-genome sequencing and the wealth of prior biological knowledge in *D. melanogaster*.

In this study, we report and analyze whole-genome shotgun sequence data for 29 sub-lines of S2 cells as well as the outgroup S1 and S3 cell lines. We show that TE insertions provide abundant markers to reconstruct the evolutionary history of S2 sub-lines, and that major phylogenetic relationships among S2 sub-lines inferred from TE insertions correlate with genome-wide copy number differences. These data show that publicly available S2 sub-lines form one monophyletic group defined by two major clades (A and B), and reveal no

evidence for widespread cross-contamination of available S2 cultures by other *Drosophila* cell lines. We infer that S2 cells underwent tetraploidization early in their evolutionary history and use copy number profiling to support the evolutionary history of S2 sub-lines based on TE profiles. Using ancestral state reconstruction, we infer that TE insertion has occurred on all internal branches of the S2 phylogeny, but that only a small subset of *D. melanogaster* TE families have proliferated during S2 evolution, most of which are retrotransposons that do not encode a retroviral envelope gene. Together, these results support the conclusions that TE insertions provide useful markers of S2 sub-line identity and genome organization and that TE proliferation in *Drosophila* somatic cell culture is an ongoing, cell-autonomous process that does not result from ubiquitous deregulation of global transpositional control mechanisms.

3.3 Materials and Methods

3.3.1 Genome sequencing

We surveyed the genomes of 32 samples of S1, S2, or S3 cells to understand the diversity and evolutionary relationships of publicly available sub-lines of S2 cells. Frozen stocks for each of these 32 samples were ordered from the *Drosophila* Genomics Resource Center (DGRC), American Type Culture Collection (ATCC), Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ), and Thermo Fisher. DNA was prepared directly from thawed samples without further culturing. Stock or catalogue numbers for these publicly available cell lines can be found in Table 3.1. Cells were defrosted and 250µl of the cell suspension was aliquoted and spun down for 5 min at 300g. The supernatant was discarded and the DNA from the cell pellet was extracted using the Qiagen DNeasy Blood & Tissue Kit (Cat. No. 69504). DNA preps were done in three batches, each of which contained an independent sample of S2-DRSC (DGRC-181) to identify any potential sample swaps and to assess the reproducibility of TE detection systems. The triplicate samples of S2-DRSC were from the same freeze of this cell sub-line performed by DGRC (Daniel Mariyappa, personal communication). Illumina

sequencing libraries were generated using the Nextera DNA sample preparation kit (Cat. No. FC-121-1030), AMPure XP beads were then used to purify and remove fragments <100bp, and libraries were normalized and pooled prior being sequenced on an Illumina HiSeq 2500 flow cell using a 101bp paired-end layout. A summary of the sequence data generated for each sample in this study can be found in Table 3.1.

Table 3.1: **Summary of 31 Schneider sub-lineages analyzed in this study.** *Drosophila* Genomics Resource Center (DGRC) cell line names are given for all cell line samples except for samples obtained from other sources, which are denoted within parenthesis. The lab origin represents the lab that originally created the sub-line. Inferred ploidy represents the ploidy estimated by analyzing DNA density of whole genome data using the method in Lee *et al.* 2014 [129]. Inferred sex represents the sex of the cell line inferred by analyzing DNA density of whole genome data and analysis of sex determination gene expression based on *et al.* 2014 [129]. Read pairs represents the number of paired-end reads for a given sample. Coverage represents the average mapped depth of coverage after quality and adaptor trimming. N.A. indicates that this information is not available.

Cell line	DGRC ID	Lab origin	Inferred ploidy	Inferred sex	SRA	Read length	Read pairs	Coverage
mbn2	DGRC-147	Gateff	4	male	SRR13360020	151	55531647	109.62
mbn2 (Gorski)	N.A.	Gateff	4	male	SRR13360019	151	63738692	130.60
mbn2 (Strand)	N.A.	Gateff	4	male	SRR13360018	151	68440069	132.00
S1	DGRC-9	Cherbas	2	male	SRR10981795	101	34904345	35.79
S2	DGRC-6	Cherbas	4	male	SRR10981796	101	28189507	31.67
S2 (ATCC-CRL-1963)	N.A.	Others	4	male	SRR10981814	101	50088154	47.26
S2 (DSMZ-ACC-130-C)	N.A.	Others	4	male	SRR10981794	101	51683568	48.33
S2 (Invitrogen-R69007)	N.A.	Others	4	male	SRR10981793	101	43038240	42.18
S2-act-GFP-alphaTub84B	DGRC-170	Rogers	4	male	SRR10981789	101	37705915	37.57
S2-DRSC-1	DGRC-181	Perrimon	4	male	SRR10981786	101	31515040	34.76
S2-DRSC-2	DGRC-181	Perrimon	4	male	SRR10981812	101	49916928	51.87
S2-DRSC-3	DGRC-181	Perrimon	4	male	SRR10981811	101	50084326	49.18
S2-GFP-SKL	DGRC-197	Rogers	4	male	SRR10981805	101	47302631	48.96
S2-Mt-Dl	DGRC-152	Klueg	4	male	SRR10981802	101	37297961	38.88
S2-Mt-EB1-GFP	DGRC-171	Rogers	4	male	SRR10981788	101	34454428	39.07
S2-Mt-Fog-myc	DGRC-218	Rogers	4	male	SRR10981803	101	38085006	40.59
S2-Mt-GFP	DGRC-194	Rogers	4	male	SRR10981808	101	41882171	43.14
S2-Mt-GFP-Act5C	DGRC-169	Rogers	4	male	SRR10981790	101	44383445	46.85
S2-Mt-mCherry-alphaTub84B	DGRC-195	Rogers	4	male	SRR10981807	101	44960162	44.06
S2-Mt-Msps-GFP	DGRC-206	Rogers	4	male	SRR10981804	101	47561052	49.00
S2-Mt-N	DGRC-154	Klueg	4	male	SRR10981792	101	40733228	43.91
S2-Mt-Slit	DGRC-192	Rogers	4	male	SRR10981810	101	41033908	41.86
S2-SQH-GFP	DGRC-172	Rogers	4	male	SRR10981787	101	26690953	29.73
S2-SQH-GFP+Mt-mCherry-actin	DGRC-193	Rogers	4	male	SRR10981809	101	42525580	41.74
S2-Tub-wg	DGRC-165	Nusse	4	male	SRR10981791	101	41826501	43.40
S2R+	DGRC-150	Wheeler	4	male	SRR10981813	101	20056094	23.23
S2R+ (DRSC)	N.A.	Perrimon	4	male	SRR11000336	151	47640215	82.65
S2R+-NPT005	DGRC-229	Perrimon	4	male	SRR10981801	101	45413880	45.17
S2R+-NPT017	DGRC-230	Perrimon	4	male	SRR10981800	101	34459982	35.27
S2R+-NPT050	DGRC-231	Perrimon	4	male	SRR10981799	101	29162882	28.27
S2R+-NPT101	DGRC-232	Perrimon	4	male	SRR10981798	101	43114190	43.26
S2R+-SQH-GFP	DGRC-196	Rogers	4	male	SRR10981806	101	49206117	47.08
S3	DGRC-5	Cherbas	4	male	SRR10981797	101	23764412	27.52

3.3.2 Prediction of non-reference TE insertions

Non-reference TE insertions were detected in each sample using trimmed paired fastq sequences as input for McClintock (v2.0) [175]. We used the TEMP [259] module in McClintock to predict non-reference TEs since it has been shown that the number of non-reference TE predictions from TEMP is least dependent on data coverage and read length compared to other component methods in McClintock [88]. The major sequences (chr2L, chr2R, chr3L, chr3R, chr4, chrM, chrY, and chrX) from the *D. melanogaster* dm6 assembly were used as a reference genome [95]. The TE library used for McClintock runs was a slightly modified version of the Berkeley *Drosophila* Genome Project canonical TE dataset v9.4.1 described in Sackton *et al*, 2009 [207] (https://github.com/bergmanlab/transposons/blob/master/releases/D_mel_transposon_sequence_set_v10.2.fa).

Genome-wide non-reference TE predictions generated by McClintock were filtered to exclude TEs in low recombination regions using boundaries defined by Cridland *et al*, 2013 [60] lifted over to dm6 coordinates, as in Han *et al*, 2021 [88]. Filtered non-reference TE predictions were then clustered across genomics coordinates and samples. TE predicted in different samples in the same cluster are required to directly overlap and be on the same strand. Clustered non-reference TE predictions were then filtered to exclude low-quality predictions using the same criteria as in Han *et al*, 2021 [88]. Briefly, the regions included in our analyses were defined as chrX:405967–20928973, chr2L:200000–20100000, chr2R:6412495–25112477, chr3L:100000–21906900, chr3R:4774278–31974278. We also excluded *INE-1* family from the subsequent analysis since this family has been reported to be inactive in *Drosophila* for million of years [220, 238].

3.3.3 Phylogenetic analysis of cell sub-line samples using TE insertion profiles

Genome-wide non-reference TE predictions were then converted to a binary presence/absence matrix as input for phylogenetic analysis. Phylogenetic trees of cell sub-lines were built using

Dollo parsimony in PAUP (v4.0a168) [230]. The phylogenetic analysis was performed using heuristic searches with 50 replicates. A hypothetical ancestor carrying the assumed ancestral state (absence) for each locus was included as root in the analysis [23, 88]. "DescribeTrees chgList=yes" option was used to assign character state changes to all branches in the tree. Finally, node bootstrap support for the most parsimonious tree was computed by integrating 100 replicates generated by PAUP using SumTrees (v4.5.1) [229].

3.3.4 Copy number and DNA density analysis

BAM files generated by McClintock were used to generate copy number profiles for non-overlapping 10kb windows of the dm6 genome using Control-FREEC (v11.6) [37]. Windows with less than 85% mappability were excluded from the analysis based on mappability tracks generated by GEM (v1.315 beta) [65]. Baseline ploidy was set to diploid for S1 and tetraploid for all other samples, according to ploidy levels estimated by Lee *et al*, 2014 [129]. The minimum and maximum expected value of the GC content was set to be 0.3 and 0.45, respectively. Copy number profiles generated by Control-FREEC were also used to create DNA density profiles as in Lee *et al*, 2014 [129] to confirm ploidy for all samples used in our study.

3.4 Results

3.4.1 Genome-wide TE profiles reveal the relationship among S2 sub-lines

A previous study has shown that clustering of genome-wide TE profiles can be used to uniquely identify *Drosophila* cell lines and provide insight into the evolutionary history of clonally evolving cell sub-lines derived from the same cell line [88]. In this study, we propose that TE profiles can also be used to infer the currently unknown evolutionary relationship among a large panel of cell sub-lines originating from different labs. We generated paired-end Illumina whole-genome sequencing (WGS) data for a panel of 28 *Drosophila* S2 sub-lines from multiple lab origins (Table 3.1) and predicted non-reference TE insertions using TEMP

[259]. We also included a S2R+ sub-line from *Drosophila* RNAi Screening Center (DRSC) with publicly available WGS data (SRR11000336). Several samples from other *Drosophila* cell lines are included in the analysis, including S1 and S3 cell lines that were independently established from the same fly stock (i.e., Oregon-R) that S2 was derived from [213], and samples from mbn2 cell line, which was previously reported to have a distinct origin [82] but inferred by a recent study to descend from a divergent lineage of S2 cells in Han *et al.* 2021 [88] (Table 3.1). We predicted between 656 and 2924 non-reference TE insertions in the euchromatic regions of these S2 sub-line samples (Table 3.2). Each sample of S2 sub-line had a unique profile of non-reference TE insertions.

We performed phylogenetic analysis using genome-wide TE profiles of all S2 sub-line samples using the Dollo parsimony approach [88]. This approach fits the assumptions of the homoplasy-free nature of TE insertions [217, 209, 244, 187, 123, 122] while also accommodates the false negative TE predictions inherent to short-read-based TE detection methods [175, 198, 235]. The most parsimonious tree revealed several expected patterns that suggest using TE profiles to infer the evolutionary relationship among *Drosophila* S2 sub-lines is reliable (Figure 3.1). First, most internal nodes are highly supported. All weakly supported nodes are close to the terminal taxa, which presumably is due to the lack of informative TE insertion signals that differentiate very closely related S2 sub-lines or sample replicates. Second, using a hypothetical ancestor as root representing the state without any non-reference insertions, S1 and S3 cell lines were independently reconstructed as outgroups for the S2 sub-lines in the phylogeny as expected based on their independent origin from the same ancestral fly stock. Third, all S2 sub-lines form a monophyletic clade with 100% bootstrap support. Forth, replicate samples of S2-DRSC sub-line cluster as nearest taxa and form a monophyletic clade with 100% bootstrap support. Fifth, all samples from S2R+, which are sub-lines of S2 with unique phenotypic characteristics [245], form a monophyletic clade with 100% bootstrap support. Finally, all samples from the mbn2 cell line recently proposed to be misidentified S2 cells [88] form a monophyletic clade with 100% bootstrap support embedded within S2 sub-

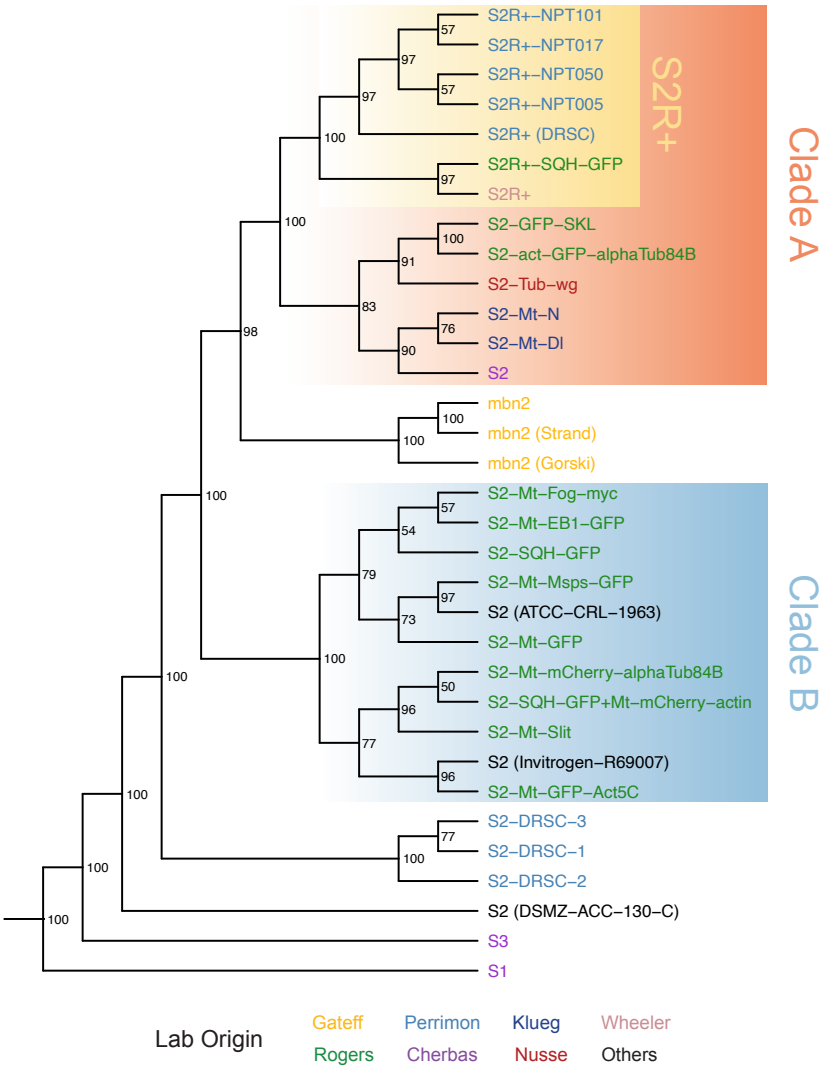

Figure 3.1: **TE profiles reveal the evolutionary relationship among S2 sub-lines.** Dollo parsimony tree including 29 *Drosophila* S2 sub-lines constructed using genome-wide non-reference TE insertions predicted by TEMP [259]. Replicate samples for S2-DRSC were included. Samples from S1 and S3 cell lines were included to serve as outgroups in the phylogenetic analysis. Samples from mbn2 cell line were also include since they are inferred to be misidentified S2 lines [88]. Percentage bootstrap support was annotated beside each node. *Drosophila* Genomics Resource Center (DGRC) cell line names are used as taxa labels. Samples obtained from other sources are labeled in the format of "cell line name (source name)". Taxa labels were colorized based on original labs in which cell sub-lines were developed.

Table 3.2: **Number of non-reference TE predictions made by TEMP for 31 Schneider sub-lineage samples.** Numbers of non-reference TE insertion predictions are based on default McClintock [175] settings. *INE-1* and non-reference TE insertion predictions in low recombination regions were excluded from all methods.

Cell line	DGRC ID	SRA	# of TEs
mbn2	DGRC-147	SRR13360020	1935
mbn2 (Gorski)	N.A.	SRR13360019	2194
mbn2 (Strand)	N.A.	SRR13360018	1980
S1	DGRC-9	SRR10981795	742
S2	DGRC-6	SRR10981796	1268
S2 (ATCC-CRL-1963)	N.A.	SRR10981814	848
S2 (DSMZ-ACC-130-C)	N.A.	SRR10981794	656
S2 (Invitrogen-R69007)	N.A.	SRR10981793	845
S2-act-GFP-alphaTub84B	DGRC-170	SRR10981789	973
S2-DRSC-1	DGRC-181	SRR10981786	1282
S2-DRSC-2	DGRC-181	SRR10981812	1084
S2-DRSC-3	DGRC-181	SRR10981811	1058
S2-GFP-SKL	DGRC-197	SRR10981805	858
S2-Mt-Dl	DGRC-152	SRR10981802	1264
S2-Mt-EB1-GFP	DGRC-171	SRR10981788	1244
S2-Mt-Fog-myc	DGRC-218	SRR10981803	1298
S2-Mt-GFP	DGRC-194	SRR10981808	1318
S2-Mt-GFP-Act5C	DGRC-169	SRR10981790	804
S2-Mt-mCherry-alphaTub84B	DGRC-195	SRR10981807	1041
S2-Mt-Msps-GFP	DGRC-206	SRR10981804	1063
S2-Mt-N	DGRC-154	SRR10981792	1133
S2-Mt-Slit	DGRC-192	SRR10981810	1038
S2-SQH-GFP	DGRC-172	SRR10981787	1534
S2-SQH-GFP+Mt-mCherry-actin	DGRC-193	SRR10981809	992
S2-Tub-wg	DGRC-165	SRR10981791	1210
S2R+	DGRC-150	SRR10981813	1820
S2R+ (DRSC)	N.A.	SRR11000336	2924
S2R+-NPT005	DGRC-229	SRR10981801	1444
S2R+-NPT017	DGRC-230	SRR10981800	1604
S2R+-NPT050	DGRC-231	SRR10981799	1275
S2R+-NPT101	DGRC-232	SRR10981798	1470
S2R+-SQH-GFP	DGRC-196	SRR10981806	1305
S3	DGRC-5	SRR10981797	1204

line diversity. These results suggest that TE profiles can be used to infer the evolutionary relationship among sub-lines of the S2 cell line, and that there is no evidence for cross-contamination between S2 sub-lines and other *Drosophila* cell lines in our dataset.

The S2 phylogeny built using TE profiles revealed a major split in the history of S2 cell line evolution resulting in two sister lineages, which we annotated as "Clade A" and "Clade B" (Figure 3.1). Clade A includes all seven S2R+ sub-lines and six S2 sub-lines while Clade B includes 11 S2 sub-lines. These results imply that S2 cells are paraphyletic (i.e., some S2 sub-lines are more closely related to S2R+ than other S2 sub-lines). In some cases, S2 sub-lines from the same lab cluster together (S2R+ sub-lines from the Perrimon lab, S2 sub-lines from the Klueg lab). However, S2 sub-lines from the Rogers lab were placed in different major clades of the S2 phylogeny (three S2-sub-lines in Clade A, nine S2-sub-lines in Clade B, Figure 3.1), demonstrating that the same lab can use sub-lines of S2 from divergent clades which have potentially different genome organization.

Majority of S2 sub-lines we surveyed in this study were placed within Clade A and Clade B based on their TE profiles. However, two S2 sub-lines, S2-DRSC and S2 (DSMZ-ACC-130-C), were independently placed as outgroups for the two major clades of S2, suggesting that they are divergent S2 lineages. S2-DRSC is routinely used for RNAi screens at the *Drosophila* RNAi Screening Center (DRSC) and was recently donated to DGRC. Its relationship to the canonical S2 sub-line from DGRC (i.e., DGRC-6) was previously not known. Our results suggest that S2-DRSC and S2 (DGRC-6) are not closely related sub-lines, which could explain the phenotypic and functional differences between these two sub-lines reported in previous studies [56, 241, 129, 130].

Samples of the mbn2 line cluster in a monophyletic subclade that is sister to Clade A (98% bootstrap support) but is clearly contained within the S2 lineage. This observation is consistent with previous study by Han *et al.* 2021 proposing that mbn2 is a misidentified S2 lineage [88]. Han *et al.* 2021 also showed that mbn2 clusters with S2-DRSC before clustering with S2R+ [88]. However, our results showed that the mbn2 clade clusters with

S2R+ before clustering with S2-DRSC. We interpret this discrepancy as caused by the low coverage sequencing data of S2 and S2R+ samples included in the previous study [88], which led to insufficient TE insertion signals to infer the evolutionary relationship among mbn2 and diverse S2 sub-lines.

3.4.2 Genome-wide copy number profiles correlate with history of S2 sub-lines

To further understand the genomic heterogeneity among S2 sub-lines, we performed DNA density analysis on all S2 sub-lines surveyed in this study (Figure 3.3). Briefly, any CNVs would lead to deviations relative to the mean peak of DNA-seq read count density at "1", the level of the deviation can then be used to infer the minimal ploidy of the sample genome [129]. This analysis revealed that all S2 sub-lines have tetraploid genomes, which could be explained by an ancestral tetraplodization event that occurred in the early stage of S2 following cell line establishment [213]. We also generated copy number variant (CNV) profiles for all S2 sub-line samples (Figure 3.2) using Control-FREEC [37], in which the baseline ploidy for each sample was determined from DNA density analysis (Figure 3.3). Several patterns were observed in the CNV profiles that suggest the robustness of using this profiling approach to characterize cell sub-lines: 1) CNV profiles of S2R+, S2-DRSC, S1, and S3 are consistent with profiles for their replicates from modENCODE reported in previous studies [253, 129], and 2) we observed high concordance among CNV profiles of replicate samples for S2-DRSC.

CNV profiles revealed a substantial amount of segmental copy number changes among S2 sub-lines (Figure 3.2). Importantly, patterns of segmental copy number changes are consistent with the evolutionary relationship among S2 sub-lines inferred from their TE profiles (Figure 3.2). The majority of CNVs in S2 were observed in the autosomal regions in sub-lines within Clade A (Figure 3.2B), including the ∼15Mbp copy number gains and losses on chromosome arm 3L that are exclusively shared by sub-lines in Clade A (Figure 3.2B, red shading). In addition, several copy number changes on chromosome X, arm 2L, and arm 2R are exclusively

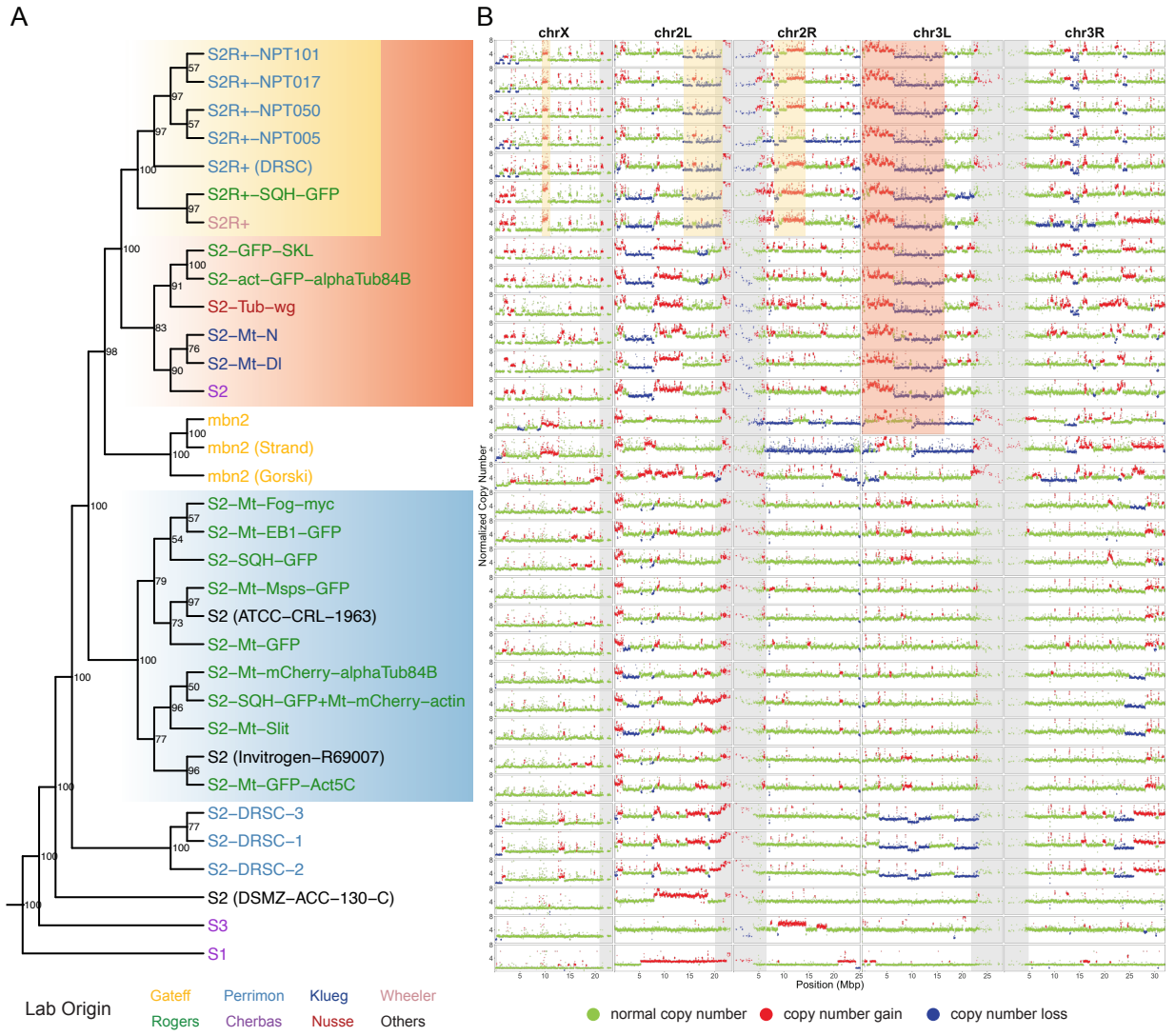

Figure 3.2: **DNA copy number profiles reveal diverse and extensive segmental aneuploidy among *Drosophila* S2 sub-lines.** (A) Dollo parsimony tree of 29 *Drosophila* S2 sub-lines based on non-reference TE predictions made by TEMP [259]. Samples from the S1, S3 and mbn2 cell lines were also included in the analysis. Taxa labels were colorized in the same way as Figure 3.1. (B) Copy number profiles for samples included in panel (A) separated by chromosome arms. Each data point represents normalized copy number (ratio*ploidy) for a given 10kb window estimated by Control-FREEC [37]. Data points for each window are colorized by CNV status (red: CNV gain; green: no CNV; blue: CNV loss), which are based on the comparison between normalized copy number and baseline ploidy. Low recombination regions are shaded in grey.

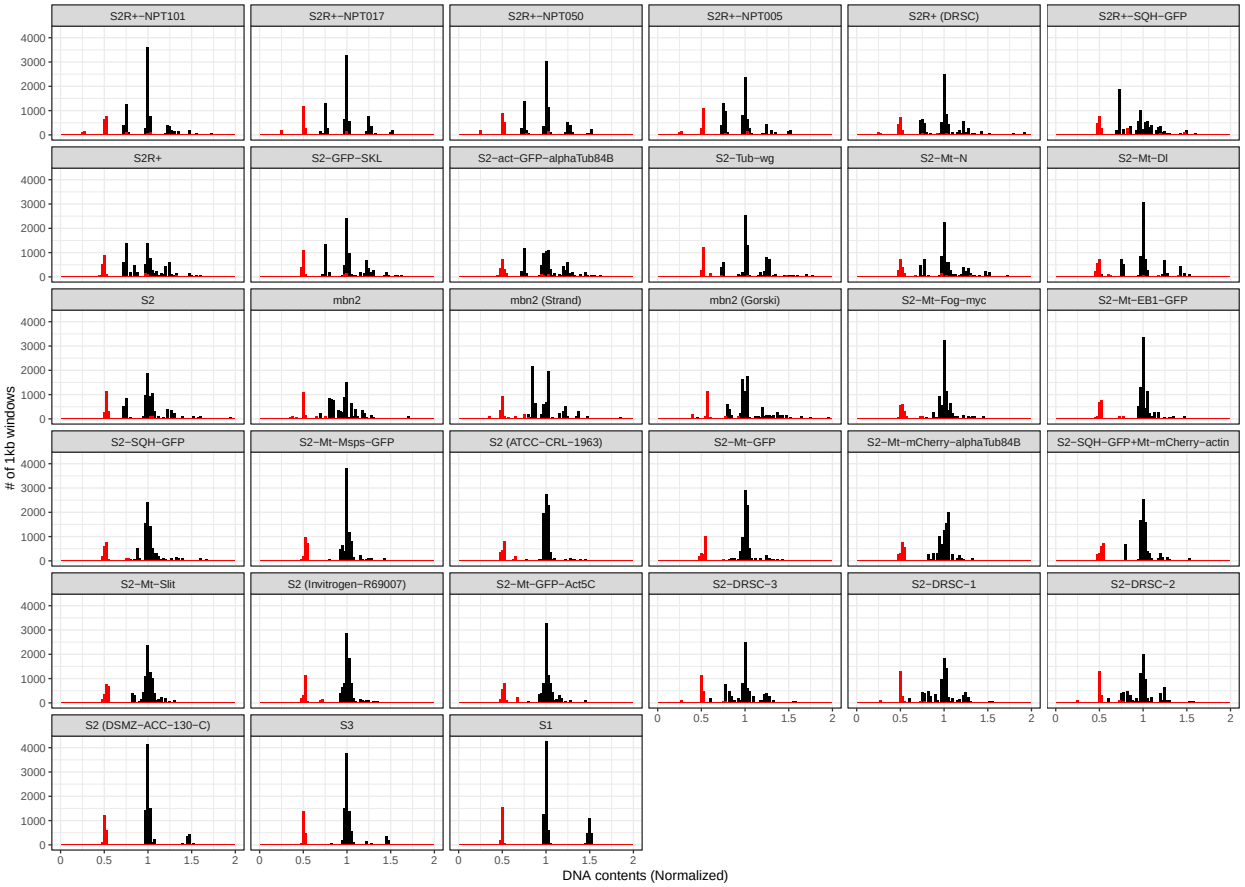

Figure 3.3: **Normalized DNA content for *Drosophila* S1, S2, S3 and mbn2 lines.** Histograms of normalized DNA read density of 1kb windows using the method described in Lee *et al.* 2014 [129]. Reads mapping to chromosome X and autosomes are show in red and black, respectively. Minimal ploidy can be inferred based on peaks that deviate from mean peak of normalized DNA read density at 1. For example, minimal tetraploidy can be marked by peaks at 0.75 and 1.25.

shared by S2R+ sub-lines (Figure 3.2B, yellow shading). In comparison, sub-lines in Clade B have very few CNVs throughout the genome (Figure 3.2B), which explains the lack of 0.75 and 1.25 peaks in the DNA density profiles for these samples. Both S2-DRSC and S2 (DSMZ-ACC-130-C) sub-lines have distinct CNV patterns that are different from other S2 sub-lines in Clade A and Clade B (Figure 3.2B), which is consistent with these two S2 sub-lines being reconstructed as divergent lineages of S2 using TE profiles. Finally, CNV profiles

of samples from mbn2 have distinct CNVs that are different from S2 sub-lines, which is consistent with the interpretation that mbn2 is a divergent lineage of S2. In addition, it is worth noting that the extensive segmental aneuploidy in mbn2 resembles the characteristic of S2 sub-lines in Clade A (Figure 3.2B).

S2R+, S2R+-SQH-GFP, and most S2 sub-lines in Clade A (except S2-Tub-wg) share a ~5Mbp copy number loss in chromosome arm 2L (Figure 3.2B), which could be explained by a segmental deletion event occurred in the common ancestor of sub-lines in Clade A, followed by reversals of the deletion in S2-Tub-wg and in the common ancestor of S2R+ sub-lines from Perrimon lab through somatic recombination (Figure 3.2B). In addition, a copy number loss on the entire chromosome arm 2L can be observed for S2R+-NPT005 but not for other S2R+ sub-lines. These results suggest that copy number changes frequently occur in the S2 cultured cells, which further contributes to diversity in genome organization among sub-lines of the S2 cell line.

3.4.3 A subset of LTR retrotransposons have ongoing transposition activities in S2 cultured cells

S2 sub-lines evolves clonally, which implies the non-reference TEs in S2 sub-lines from the ancestral fly strain and early passages should be shared by all sub-lines. In the absence of any other factors, these ancestral insertions can not provide phylogenetic signal to infer the evolutionary history of S2. Since TE profiles clearly provide signals that can be used to infer the evolutionary history of S2 sub-lines, we infer that phylogenetically informative TE insertions mainly came from ongoing somatic TE transpositions during prolonged S2 cell culture (i.e., the "ongoing transposition" hypothesis). However, the corrleation between CNV profiles and S2 phylogeny raises an alternative hypothesis, which is that the majority of TE insertions in the S2 sub-lines happened soon after the establishment of the original S2 cell line (i.e., the "initial TE burst" hypothesis), and that the S2 phylogeny built from TE profiles of S2 sub-lines is driven by subsequent clonally inherited copy number loss events. If

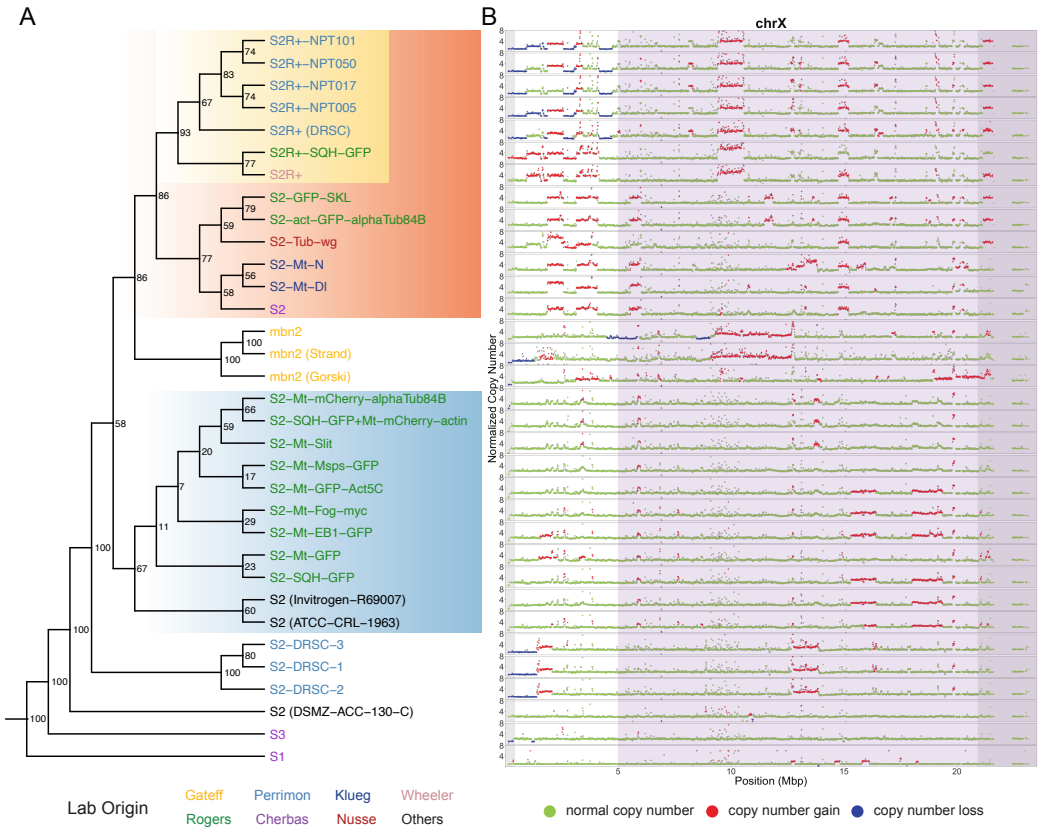

Figure 3.4: **Evolutionary relationship among S2 sub-lines inferred using TE profiles in regions without copy number loss.** (A) Dollo parsimony tree of 29 *Drosophila* S2 sub-lines constructed using non-reference TE insertions predicted by TEMP [259] in regions of chromosome X without copy number losses. Samples from S1, S3 and mbn2 cell lines were also included. Percentage bootstrap support was annotated beside each node. *Drosophila* Genomics Resource Center (DGRC) cell line names are used as taxa labels. Samples obtained from other sources are labeled in the format of "cell line name (source name)". Taxa labels were colorized based on original labs in which cell sub-lines were developed. (B) Copy number profiles of chromosome X for samples included in panel (A). Each data point represents normalized copy number (ratio*ploidy) for a given 10kb window estimated by Control-FREEC [37]. Data point for each window is colorized by CNV status (red: CNV gain; green: no CNV; blue: CNV loss), which are based on the comparison between normalized copy number and baseline ploidy. Regions without copy number loss across all S2 sub-lines are shaded in purple. Low recombination regions are shaded in grey.

this alternative hypothesis is true, we should expect all S2 sub-lines to have highly similar TE profiles in regions without significant copy number loss events. Furthermore, we should

expect that TE profiles in regions without copy number loss do not have sufficient signal to infer the evolutionary history of S2 sub-lines. To test the alternative hypothesis, we analyzed TE profiles in a ∼15Mbp region in chromosome X that does not include significant copy number loss across all S2 sub-lines we surveyed (Figure 3.2B; Figure 3.4B, purple shading). Our analysis revealed that the majority of TE insertions in regions of the X chromosome without copy number loss are exclusive to one or a subset of S2 sub-line samples (Figure 3.5). We then built the Dollo parsimony tree of *Drosophila* S2 sub-lines using non-reference TE insertions in the same region of chromosome X. The most parsimonious tree has the same major topological features as the one built from genome-wide TE profiles (Figure 3.4). Together, these results provide evidence against the initial burst and CNV loss model and suggest that the genome-wide TE profiles used to infer evolutionary relationship of S2 sub-lines are contributed mainly by ongoing S2 sub-line- or lineage-specific somatic transposition.

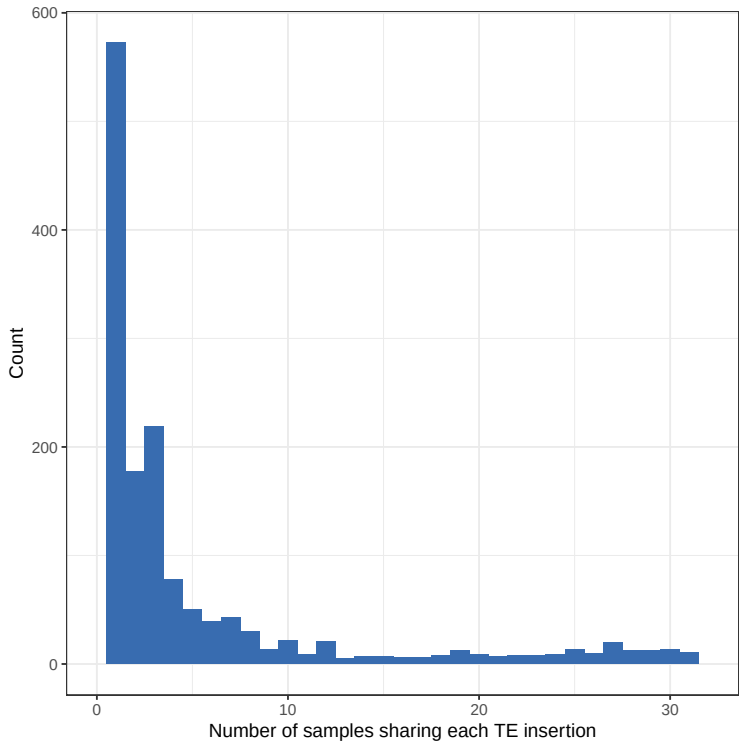

Figure 3.5: **Majority of non-reference TE insertions are exclusive to one or a subset of sub-line samples in S2.** Histogram shows the number of *Drosophila* S2 sub-line samples that share each given TE insertion in regions of chromosome X without copy number losses.

To gain additional insights into the dynamics of TE transposition activity during the history of S2 cell line evolution, we mapped TE insertions on the phylogeny of *Drosophila* S2 sub-lines using ancestral state reconstruction based on the most parsimonious scenario of TE gain and loss under the Dollo model [23, 194, 88]. If insertions happened in the early stage of S2 and were later purged in S2 sub-lineages through copy number loss, the Dollo assumption would favor early insertion over parallel gains of TEs in different sub-lineages [75]. This unique characteristic of the Dollo assumption suggests that the phylogenetic analysis from TE profiles is not biased towards a tree with more recent TE insertions, thus is conservative regarding the "ongoing transposition" hypothesis. Finally, if the "ongoing transposition" hypothesis is true, we expect to observe TE insertions on virtually all branches in the tree. The most parsimonious scenario of TE insertions mapped on the phylogeny including 29 *Drosophila* S2 sub-lines is shown in Figure 3.6. Although the Dollo model favors insertions that are shared by all samples over more recent parallel insertions [75], a substantial amount of TE insertions were observed on both ancestral and recent branches in the phylogeny, thereby supporting the "ongoing transpositions" hypothesis.

We then aggregated the insertions mapped on each branch of the most parsimonious tree by TE families to visualize branch- and family-specific TE insertion profiles. This analysis revealed that only a subset of 125 curated TE families in *D. melanogaster* exhibit high transposition activity in the S2 cell culture (Figure 3.7). The top 10 most active TE families are all LTR retrotransposons except for *jockey*, which is long interspersed nuclear element (LINE). In addition, ancestral state reconstruction analysis also revealed branch- and TE family-specific activities. Branch 32 that goes from the hypothetical ancestor (i.e., ISO1) to the ancestor of all Schneider cell samples (i.e., Oregon-R) is enriched with *roo* insertions, which can be expected since *roo* is one of the most active TE families in the natural populations of *D. melanogaster* [193]. Branch 34 that leads to the ancestor of all S2 sub-lines is enriched with *17.6*, *297* and *1731* insertions. Branches 49, 48, and 37 that leads to the ancestor of S2R+, Clade A, Clade B, respectively, are enriched with a variety of TE families

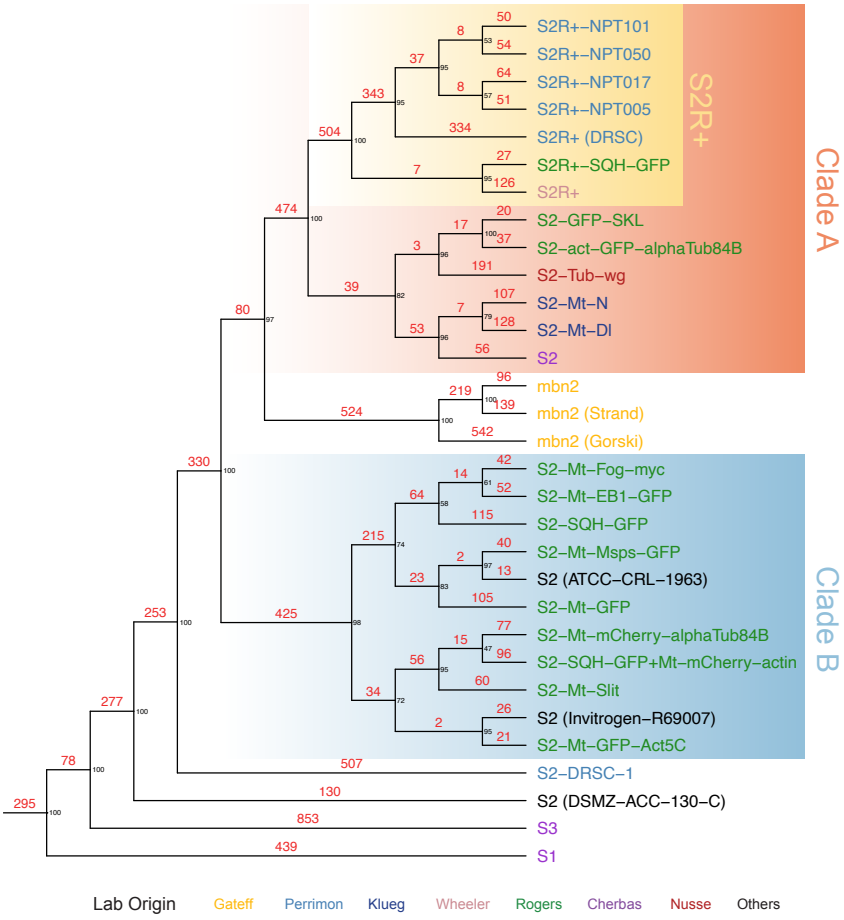

Figure 3.6: **Ancestral state reconstruction supports ongoing TE transpositions in S2 cell culture.** Dollo parsimony tree of 29 *Drosophila* S2 sub-lines constructed using non-reference TE insertions predicted by TEMP [259]. Samples from S1, S3 and mbn2 cell lines were also included in the analysis. The number of TE insertions estimated using ancestral state reconstruction were annotated above each branch. Percentage bootstrap support was annotated beside each node. *Drosophila* Genomics Resource Center (DGRC) cell line names are used as taxa labels. Samples obtained from other sources are labeled in the format of "cell line name (source name)". Taxa labels were colorized based on original labs in which cell sub-lines were developed.

including *jockey*, *blood*, *copia*, *297*, *3S18*, *1731*, *diver*, *mdg1* and *17.6*. Insertions from TE families on ancestral branches that were active during the S2 cell culture evolution should play important roles in shaping unique TE profiles of different S2 lineages. Compared to

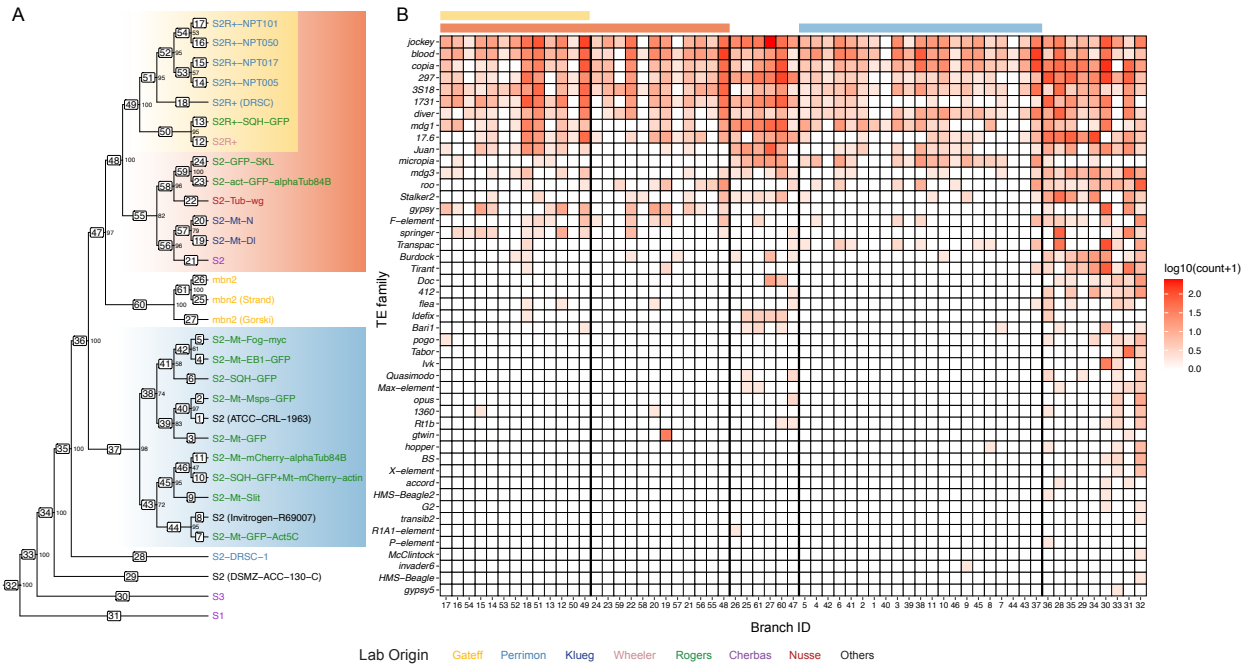

Figure 3.7: **Ongoing TE insertions are contributed by a small subset of LTR retrotransposon families in *Drosophila* S2 culture.** (A) Dollo parsimony tree of 29 *Drosophila* S2 sub-lines based on non-reference TE predictions made by TEMP [259]. Samples from S1, S3, and mbn2 cell lines were also included. Taxa labels were colorized in the same way as Figure 3.1. (B) Heatmap showing the number of family-specific TE insertions on each branch of the tree in (A) based on ancestral state reconstruction. The heatmap is colorized by log-transformed (log10(count+1)) number of gains per family per branch, sorted top to bottom by overall non-reference TE insertion gains per family across all branches, and sorted left to right into clades representing major clades of S2 phylogeny with major clade color codes indicated at the top of the heatmap.

ancestral branches, more recent branches within the two major clades of S2 were mapped with fewer TE insertion activities except for several branches within the S2R+ clade. We observed a significant number of insertions mapped to branches 49, 52, and 18, which lead to the ancestor of S2R+ clade, S2R+ (DRSC), and the remaining S2R+ sub-lines from the Perrimon lab, respectively. The high TE insertion activities on these branches correlate with the popularity and extensive passage of the S2R+ cell line. Together, these results provide

evidence that different TE families were active at different times during the the evolutionary history of S2, and support an episodic model of ongoing transposition in cell culture.

3.5 Discussion

In this study, we used TE profiles to reveal the evolutionary relationship among 29 *Drosophila* S2 sub-lines. All S2 sub-lines form a single monophyletic clade in the phylogeny, suggesting that no cross-contamination between S2 and other *Drosophila* cell lines in our dataset. Our results also revealed two major subclades of S2 that are supported by copy number profiles. One major clade we labeled as "Clade A" includes all S2R+ sub-lines and several S2 sub-lines. This clade can be characterized by substantial copy number changes in the autosomes. The other major clade we labeled as "Clade B" includes only S2 sub-lines with mostly intact genomes. These results imply that "S2" is paraphyletic and revealed substantial heterogeneity in sub-lines labeled as S2. We also found that some S2 sub-lines originating from the same lab were reconstructed in different major clades of S2, providing evidence that heterogeneity in S2 genome content has the potential to influence results obtained in one lab. Future experiments on spontaneous TE accumulation in S2 would be needed to estimate the TE transposition rate in the *Drosophila* cell culture and date the divergence time between major clades of S2.

We also performed ancestral state reconstruction of TE insertions on the S2 phylogeny, which provided evidence for the "ongoing transposition" hypothesis in the S2 cell culture. One potential issue with this analysis is the presence of false-positive (FP) and false-negative (FN) non-reference TE predictions. In principle, a random FP prediction is unlikely to be shared by multiple cell samples thus should lead to a falsely reconstructed insertion on the terminal branch under the Dollo model. This suggests that the number of TE insertions reconstructed on the terminal branches of our trees may be overestimated. Conversly, a random FN would most likely lead to falsely reconstructed deletion on the terminal branch

under the Dollo model. Thus, random FP and FN TE predictions should have a limited impact on the phylogenetic and ancestral state reconstruction analyses.

We infer from the ancestral state reconstruction analysis that transposition has occurred throughout the S2 phylogeny but that only a subset of TE families have high transposition activity in the S2 culture. Most of the active TEs in S2 are retrotransposons that do not encode a functional retroviral envelope (*env*) gene (except for *297* and *17.6*), suggesting that TE proliferation in cell culture is an endogenous process. The fact that we do not observe global activation of all TE families suggests transposition in S2 is under some form of TE-family-specific regulation. Arkhipova *et al*, 1995 [9] provided two non-mutually exclusive hypotheses for proliferation of TEs in cell lines: 1) the ongoing transposition is more easily tolerated in cultured cells and is no longer under strong negative selection as in the whole flies, and 2) there exist specific factors that control TE transposition, and their actions are altered significantly in cell culture. More work is needed to understand the mechanism by which TE copy number regulation is relaxed in a family-specific fashion in the *Drosophila* cell culture.

Overall, this study revealed ongoing somatic TE insertions and copy number changes as mechanisms for genome evolution in the *Drosophila* S2 cell culture, which only has 50 years of history since establishment [213]. The genomic and phenotypic heterogeneities within cell culture have also been reported for the human HeLa cell line [141] and the MCF-7 breast cancer cell line [24], suggesting that rapid genome evolution and within culture heterogeneity are common features of animal cell culture. Future work is needed to further understand the genome evolution of animal cell culture and how the genome content changes affect cell phenotypes and functional studies.

Chapter 4

Local assembly of long reads enables phylogenomics of transposable elements in a polyploid cell line[1]

[1]Guilherme B Dias (co-first author), Shunhua Han (co-first author), Preston J Basting, Raghuvir Viswanatha, Norbert Perrimon, and Casey M Bergman, Local assembly of long reads enables phylogenomics of transposable elements in a polyploid cell line.

To be submitted to *Genetics*.

4.1 ABSTRACT

Animal cell lines cultured for extended periods often undergo extreme genome restructuring events, including polyploidy and segmental aneuploidy that impede *de novo* whole-genome assembly (WGA). In *Drosophila*, many established cell lines also exhibit massive proliferation of specific transposable element (TE) families relative to wild-type flies. To better understand the process of TE evolution in *Drosophila* somatic cell culture, we sequenced the genome of a tetraploid *Drosophila* cell line called S2R+ using long-read and linked-read technologies. WGAs of S2R+ provided variable estimates of TE content and were highly fragmented relative to assembly of genomes of whole-flies from inbred strains. In order to study TE sequence evolution without depending on WGAs, we developed a novel bioinformatics tool called "TELR" that identifies, locally assembles, and estimates allele frequency of transposable elements from long-read data (https://github.com/bergmanlab/telr). Application of TELR to a ∼130x PacBio dataset for S2R+ revealed many haplotype-specific TE insertions that occurred after tetraploidization of the S2 genome and therefore must have arisen by somatic transposition in cell culture after initial establishment of the cell line. Local assemblies from TELR also allowed phylogenetic analysis of paralogous TE copies within the S2R+ genome, which revealed that proliferation of different TE families in *Drosophila* cell lines could be driven by single or multiple source lineages. Our work provides a model for the analysis of TEs in complex heterozygous or polyploid genomes that are not amenable to WGA and yields new insights into the mechanisms of TE sequence evolution in animal cell culture.

4.2 INTRODUCTION

Cell lines are commonly used in biological and biomedical research, however little is known about how cell line genomes evolve during routine cell culture. Early studies using cells extracted from abnormally grown tissue such as mammalian tumors and plant galls revealed

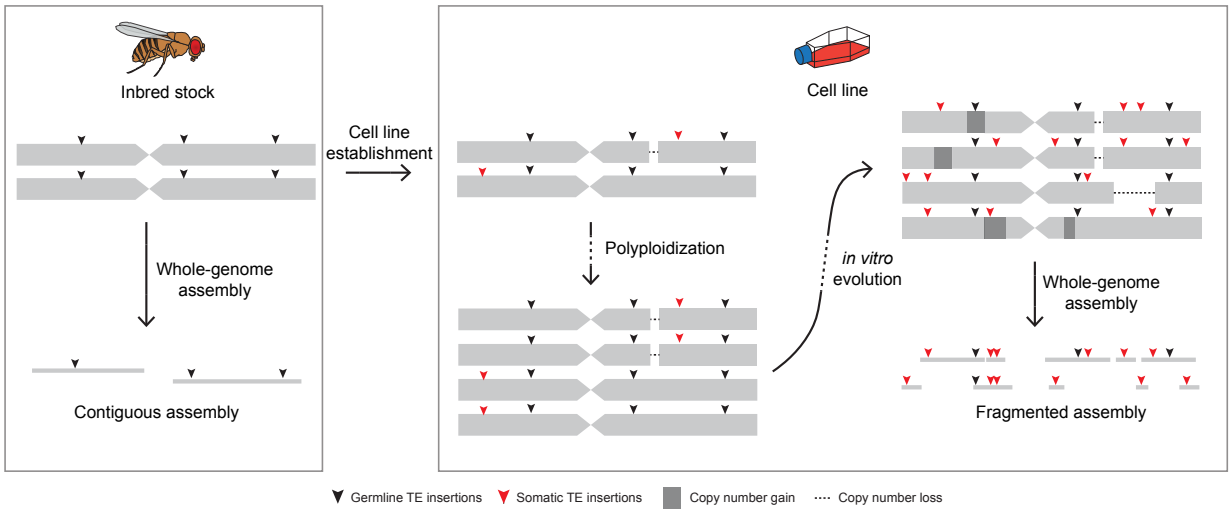

Figure 4.1: **Genome architecture complexity hinders whole-genome assembly in long-term cultured cell lines.** The inbred fly stock has diploid genome that includes homozygous variations, which allows contiguous whole-genome assembly (WGA). In comparison, cell lines established from inbred fly stock undergo polyploidization and accumulates heterozygous variations including segmental aneuploidy and haplotype-specific TE insertions during long-term culture. The complexity of polyploid genome with heterozygous variants may lead to highly fragmented WGA and as a result limit the utility of using WGA to study TE sequence evolution.

large changes in genome structure including polyploidy and aneuploidy [131]. Likewise, cell lines derived from normal plants or animal euploid tissues often develop polyploidy and aneuploidy in cell culture [79, 93, 180, 13]. More recently, the use of DNA sequencing has further uncovered that segmental aneuploidy and other types of submicroscopic structural variation are widespread in cell lines [166, 2, 129, 172, 24, 257, 256, 141]. These observations indicate that cells in culture often evolve a complex genome architecture that deviates substantially from their original source material. This complexity can impose limitations on efforts to perform *de novo* whole-genome assembly (WGA) [162, 163, 172] and thus limit the ability to study cell line genome structure and evolution using WGA-based approaches. Resolving the evolutionary processes that govern the transition from wild-type to complex cell line

genome architectures is important for understanding the stability of cell line genotypes and the reproducibility of cell-line-based research.

Like many cell lines, Schneider-2 (S2) cells from the model insect *Drosophila* have undergone polyploidization [213], and display substantial small- and large-scale segmental aneuploidy [253, 129]. In addition, S2 and other *Drosophila* cell lines exhibit a higher abundance of transposable element (TE) sequences compared to whole flies [189, 100, 193], with TE families that are abundant in S2 being different from those amplified in other *Drosophila* cell lines [70, 193, 88, 155]. However, little is known about the pattern of TE sequence evolution in S2 or other *Drosophila* cell lines. For example, it is typically unknown whether the proliferation of TEs in *Drosophila* cell lines [152] is contributed by one or more source lineages. The lack of understanding on TE sequence evolution in *Drosophila* cell lines in previous studies is mainly due to the limitation of using legacy short read data, which typically does not allow contiguous WGA or complete assembly of TE insertions or other structural variants [5, 232, 119, 254].

Recent advances in long-read DNA sequencing technologies have substantially improved the quality of WGAs, including a better representation of repetitive sequences such as TEs [30]. In *Drosophila*, long-read WGAs of homozygous diploid genomes such as those from inbred fly stocks can achieve high contiguity and permit detailed analysis of structural variation including TE insertions [30, 51, 39, 167]. However, WGA using long reads still remains limited by complex genome features including polyploidy, heterozygosity, and high repeat content, all of which are present in cell lines such as *Drosophila* S2 cells [213, 189, 100, 253, 129, 193, 88, 155]. In fact, the state-of-the-art long-read assemblies of wild-type diploid genomes still suffer from the presence of repeats and heterozygosity, which may result in assembly gaps and haplotype duplication artifacts [197, 185]. Therefore, it is expected that the assembly of highly complex *Drosophila* cell line genomes like S2 should result in substantially more fragmented WGAs than those generated from homozygous diploid fly stocks, and is likely to impact the subsequent annotation and analysis of TE sequences (Figure 4.1).

To gain insight into the mechanisms of genome evolution in *Drosophila* cell culture, we sequenced the genome of a commonly used variant of S2 cells, the S2R+ cell line [245], using PacBio long-read and 10x Genomics linked-read technologies. As predicted, WGAs of S2R+ from long-read sequencing data were highly fragmented and yielded highly variable estimates of TE content using different assembly methods (Figure 4.7). To circumvent the limitations of WGA and comprehensively characterize TE content in *Drosophila* cell lines, we developed a TE detection tool called TELR (Transposable Elements from Long Reads, pronounced "Teller") that can predict non-reference TE insertions based on a long-read dataset, a reference genome, and a TE library. Importantly, TELR can detect haplotype-specific TE insertions, reconstruct TE sequences, and estimate intra-sample TE allele frequencies (TAFs) from polyploid genomes that are not amenable to WGA. We applied TELR to PacBio long-read datasets of S2R+ and a geographically-diverse panel of *D. melanogaster* inbred fly strains from the *Drosophila* Synthetic Population Resource (DSPR) [50]. We discovered a large number of simplex TE insertions from a subset of LTR retrotransposons in the tetraploid S2R+ cell line. We inferred that these simplex insertions came from somatic transpositions after S2 cells went tetraploid following initial cell line establishment. We also performed phylogenomics analysis on the full-length TE sequences that were assembled by TELR. The analysis revealed that amplification of TEs in *Drosophila* cell lines could originate from single or multiple source lineages. Together, our work provides a novel computational framework to study polymorphic TEs in complex polyploid genomes. It also improves our understanding of TE dynamics in the long-term *Drosophila* cell culture.

4.3 Materials and Methods

4.3.1 Cell culture

An initial sample of S2R+ cells, which we define as passage 0, was obtained from a routine freeze of cells made by the *Drosophila* RNAi Screening Center (DRSC). Cells from passage 0 were defrosted and recovered in Schneider's *Drosophila* medium (Thermo) containing 10%

FBS (Thermo) and 1X Penicillin-Streptomycin (Thermo), then expanded continually for two additional passages in T75 flasks. Aliquots of cells from passage 3 flasks were frozen, and the remaining cells were expanded to 10 T75 flasks (passage 4A). Passage 4A cells were pooled and harvested to make DNA for PacBio libraries. A frozen stock was defrosted and expanded for two additional passages (passages 4B-5B). Passage 5B cells were harvested to make DNA for 10X Genomics libraries. The provenance of the cell line samples used in this study is depicted in Figure 4.2.

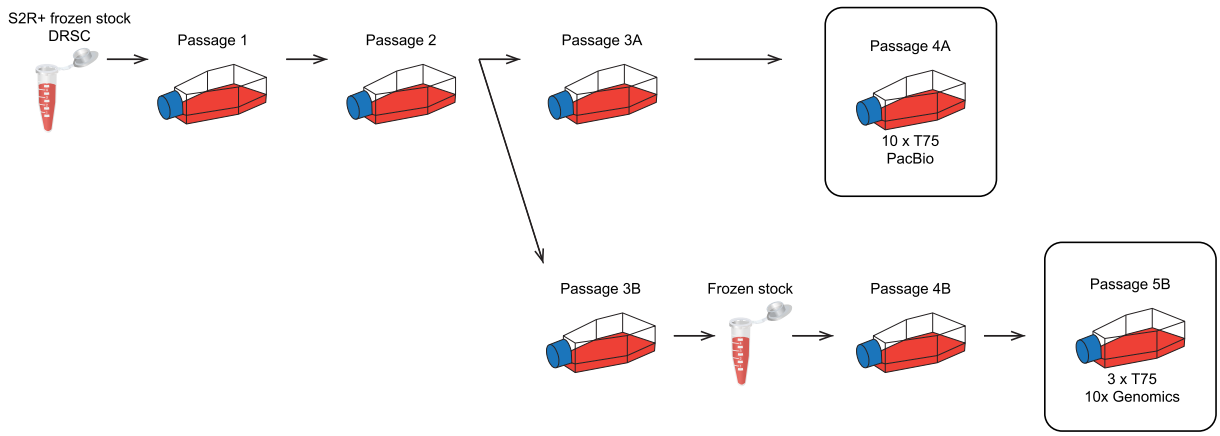

Figure 4.2: **Provenance of the S2R+ cells used in this study.** Cells were harvested from passages 4A and 5B for PacBio and 10x Genomics linked-read DNA sequencing, respectively.

4.3.2 Fly stocks

A stock of *D. melanogaster* strain A4 from the Drosophila Synthetic Population Resource (DSPR) [114] was obtained from Stuart Macdonald (University of Kansas) and reared on Instant Drosophila Medium (Carolina Biological, Cary NC) until DNA extraction.

4.3.3 DNA purification, library preparation, and sequencing

Cells were harvested from passages 4A and 5B (Figure 4.2) for PacBio and 10x Genomics linked-read DNA sequencing, respectively. Detailed DNA purification and library prep methods are described in the Supplementary Methods. Briefly, DNA for the PacBio library

was obtained from 10 confluent T75 flasks using a modified phenol:chloroform:isoamyl alcohol protocol. DNA for the 10x Genomics linked-read library was obtained from ∼1x10^7 cells using the 10x Genomics recommended "salting out" protocol (https://support.10xgenomics.com/permalink/5H0Dz33gmQOea02iwQU0iK).

DNA for *D. melanogaster* strain A4 linked-read library was obtained from a single female fly following the 10x Genomics recommended protocol for DNA purification from single insects (https://support.10xgenomics.com/permalink/7HBJeZucc80CwkMAmA4oQ2).

PacBio SMRTbell libraries were prepared for S2R+ using the Procedure & Checklist 20 kb Template Preparation with BluePippin Size Selection. The resulting library was sequenced on a RS II instrument using 31 SMRT cells with a movie time of 240 minutes per cell, generating a total of 3,510,012 reads (∼28.5 Gbp).

Linked-read libraries were prepared for both S2R+ and A4 after DNA size selection with BluePippin to remove fragments shorter than 15 kb. Libraries were prepared following the 10x Genomics Chromium Genome Reagent Kit Protocol v2 (RevB) using a total DNA input mass of 0.6 ng for each sample. The linked-read libraries were sequenced on an Illumina NextSeq 500 instrument mid-output flow cell with 150 bp paired-end layout, generating 95,280,430 reads for S2R+ (∼13.3 Gbp) and 127,009,398 reads for A4 (∼17.7 Gbp).

4.3.4 Whole-genome assembly and QC

Raw PacBio reads from S2R+ (generated here; SRX7661404) and A4 from Chakraborty *et al.* 2018 [51] (SRX4713156) were independently used as input for whole-genome assembly with Canu (v2.1.1; genomeSize=180m corOutCoverage=200 "batOptions=-dg 3 -db 3 -dr 1 -ca 500 -cp 50" -pacbio-raw), FALCON-Unzip (pb-falcon v0.2.6; seed coverage = 30, genome_size = 180000000), wtdbg2 v2.5 (-x rs -g 180m), and Flye (v2.8.2) [57, 118, 117, 203]. The reads were re-aligned to the resulting assemblies with pbmm2 (v1.3.0; --preset SUBREAD --sort) and the assemblies were polished with the 'Arrow' algorithm from GenomicConsensus (v2.3.3)

using default parameters. FALCON-Unzip performs read re-alignment and Arrow polishing automatically as part of its phasing pipeline.

10x Genomics linked-reads generated here were used as input for whole-genome assembly with Supernova (v2.1.1) for S2R+ (--maxreads=61508497) and A4 (--maxreads=77907944) [240]. The optimal '--maxreads' parameter was calculated by Supernova in a previous run to avoid excessive coverage. Supernova assemblies were exported in 'pseudohap2' format and pseudo-haplotype1 was analyzed.

10x Genomics reads from S2R+ and A4 were also barcode-trimmed with LongRanger (v2.2.2; basic pipeline) [255] and used as standard paired-end reads in a WGA with SPAdes (v3.15.0) using default parameters [14].

All assemblies were filtered to remove redundancy using the 'sequniq' program from GenomeTools (v1.6.1) [86]. General assembly statistics were calculated with the 'stats.sh' utility from BBMap (v38.83) [43]. Assembly completeness was assessed with BUSCO (v4.0.6) [219, 239] and the Diptera ortholog set from OrthoDB (v10) [121].

4.3.5 Assessment of overall TE content

Transposable elements were annotated in all WGAs with RepeatMasker (v4.0.7; -s -no_is -nolow -x -e ncbi) (https://www.repeatmasker.org/RepeatMasker/) using v10.2 of the curated library of *D. melanogaster* canonical TE sequences (https://github.com/bergmanlab/transposons). TE abundance was calculated from RepeatMasker's ".out.gff" files as the percentage of bases masked in each assembly.

Barcode-trimmed linked-reads were also used as an assembly-free estimate of TE content in S2R+ and A4. Reads were filtered for adapters and low quality bases, and trimmed to 100 bp using fastp (v0.20.0; --max_len1 100 --max_len2 100 --length_required 100) [53]. A random sample of 5 million read pairs (10 million reads) was extracted for each dataset using seqtk (v1.3; -s2) (https://github.com/lh3/seqtk) and masked using RepeatMasker (v4.0.7; -s -no_is -nolow -x -e ncbi) and the *D. melanogaster* canonical TE set (v10.2; https:

//github.com/bergmanlab/transposons). Abundance for each TE family was calculated as the percentage of read bases that were RepeatMasked.

4.3.6 Detection of non-reference TE insertions using long reads

The TELR pipeline consists of four main stages: (1) general SV detection and filter for TE insertion candidate, (2) local reassembly and polishing of the TE insertion, (3) identification of TE insertion coordinates, and (4) estimation of intra-sample TE insertion allele frequency.

In stage 1, long reads are aligned to the reference genome using NGMLR (v0.2.7) [216]. The alignment output in BAM format is provided as input for Sniffles (v1.0.12) to detect structural variations (SVs) [216]. TELR then filters for TE insertion candidates from SVs reported by Sniffles using following criteria: 1) The type of SV is an insertion, 2) The insertion sequence is available, and 3) The insertion sequences include hits from user provided TE consensus library using RepeatMasker (v4.0.7; http://www.repeatmasker.org/).

In stage 2, reads that support the TE insertion candidate locus based on Sniffles output are used as input for wtdbg2 (v2.5) to assemble local contig that covers the TE insertion for each TE insertion candidate locus [203]. The local assemblies are then polished using minimap2 (v2.20) [136] and wtdbg2 (v2.5) [203].

In stage 3, TE consensus library is aligned to the assembled TE insertion contigs using minimap2 and used to define TE-flank boundaries. TE region in each contig is annotated with family info using RepeatMasker (v4.0.7). Sequences flanking the TE insertion are then re-aligned to the reference genome using minimap2 to determine the precise TE insertion coordinates and target site duplication (TSD).

In stage 4, raw reads aligned to the reference genome are extracted within a 1kb interval on either side of the insertion breakpoints initially defined by Sniffles. The reads are then aligned to the assembled polished contig to identify reads that support the non-reference TE insertion and reference alleles, respectively, in following steps: 1) Reads are aligned to the forward strand of the contig, 5' flanking sequence depth (5p_flank_cov) and

5' TE depth (5p_te_cov) are calculated. 2) Reads are aligned to the reverse complement strand of the contig, 5' flanking sequence depth (3p_flank_cov) and 5' TE depth (3p_te_cov) are calculated. 3) The TE allele frequency is estimated as (5p_te_cov/5p_flank_cov + 3p_te_cov/3p_flank_cov)/2.

TELR (v0.2; revision bb90a5) was applied to the S2R+ PacBio dataset and to a panel of 13 *D. melanogaster* strains from the *Drosophila* Synthetic Population Resource (DSPR) (Bioproject ID PRJNA418342) [50]. The mapping reference used was release 6 of the *D. melanogaster* reference genome (chr2L, chr2R, chr3L, chr3R, chr4, chrX, chrY, chrM) [95] and the TE library was v10.2 of the *D. melanogaster* canonical TE sequence library (https://github.com/bergmanlab/transposons/blob/master/releases/D_mel_transposon_sequence_set_v10.2.fa).

The marked enrichment of TE insertions with TAF ∼0.25 and ∼0.75 in strain A7 (Figure 4.14) indicates sample contamination/mislabeling. We used BEDTools (v2.29.0) [192] to investigate the possibility of contamination of sample A7 with another strain by intersecting TE predictions between A7 and all other DSPR strains.

4.3.7 Evaluate TELR on predicting non-reference TE insertion coordinate and family

We generated synthetic datasets using reads simulated from the ISO1 (dm6) and A4 (GCA_003401745.1) [51] genome assemblies to evaluate TELR on predicting non-reference TE insertion under different ploidy, zygosity and coverage settings. In principle, a good predictor should be able to accurately predict "non-reference" insertions that are present in genome 1 (e.g., ISO1) but absent from genome 2 (e.g., A4) using reads simulated from genome 1 and 2 mapped to genome 2. Synthetic datasets under different settings were created as follows: 1) We simulated Pacbio reads from ISO1 to model diploid homozygous insertions. 2) We simulated and combined Pacbio reads from both ISO1 and A4 with equal coverages to model diploid heterozygous insertions. 3) We simulated and combined Pacbio

reads ISO1 and A4 with 1:3, 2:2, 3:1, and 4:0 ratio to model tetraploid simplex, duplex, triplex and quadruplex insertions, respectively (Table 4.1). The simulations were conducted using pbsim2 (v2.0.1; P6C4 HMM model) [181] under 50X, 100X, 150X, and 200X coverages for all ploidy, zygosity, and coverage settings. The synthetic datasets were used as input to TELR to detect non-reference TE insertions (v0.2; revision bb90a5a; options: –assembler wtdbg2 –polisher flye -p 1). The A4 assembly was used as the reference genome and the Berkeley *Drosophila* Genome Project canonical TE dataset v10.2 was used for these analyses.

As ground truth for evaluating TELR performance, curated TE annotations from the release 6.38 version of *D. melanogaster* ISO1 genome (http://ftp.flybase.net/releases/FB2021_01/dmel_r6.38/gff/dmel-all-r6.38.gff.gz) were lifted over to A4 genome assembly. After excluding *INE-1* insertions and TE insertions in low recombination regions, 1163 curated TEs in ISO1 could be lifted over to A4 on the basis of their flanking regions. TELR predictions were considered true positives if the predicted TE insertion coordinates were within a 5bp window of a lifted over ISO1 TE annotation and if the predicted TE family was the same as the lifted over annotation. The final benchmark results for TELR applied to synthetic datasets are summarized in Table 4.1. The results suggested that TELR has high precision ($\geq$95%) under all ploidy, zygosity and coverage settings. In contrast, TELR's recall was much lower, especially at low effective coverage levels. These results indicate that the non-reference TE insertion predictions made by TELR are highly accurate, however, the method has an appreciable false negative rate especially when the effective coverage is lower than 50X.

4.3.8 Evaluation of a classifier for predicting homozygous or heterozygous TE insertion in diploid and tetraploid genome using long-read data

To fill a gap in tools available to analyze intra-sample TE allele frequencies (TAF) in long-read data, we built a classifier to determine zygosity of TE insertion predicted by TELR.

Table 4.1: **TELR performance benchmark using genome-wide synthetic data from ISO1 and A4 genome assemblies.** Non-reference TE insertion predictions made by TELR using the A4 genome assembly as reference were evaluated against curated TE annotations in ISO1 lifted over to A4 coordinates (see Section 4.3.7 for details). Zygosity represents the ratio of simulated reads generated from ISO1 and A4 (see details in Section 4.3.7). "True Positives" and "False Positives" represent the number of predictions that match and doesn't match with lifted over insertion annotations, respectively. "False Negatives" represent the number of lifted over non-reference TE insertion annotations that are not predicted by TELR. "Precision" represents the number of true positives divided by total number of predictions. "Recall" represents the number of true positives divided by total number of lifted over non-reference TE insertion annotations.

Ploidy	Zygosity	Coverage	Total	True Positives	False Positives	False Negatives	Precision	Recall
diploid	homozygous	50	483	466	17	145	96.5%	75.4%
diploid	homozygous	100	521	503	18	107	96.5%	81.4%
diploid	homozygous	150	545	526	19	83	96.5%	85.1%
diploid	homozygous	200	535	516	19	94	96.4%	83.5%
diploid	heterozygous	50	423	414	9	200	97.9%	67.0%
diploid	heterozygous	100	504	483	21	126	95.8%	78.2%
diploid	heterozygous	150	520	502	18	109	96.5%	81.2%
diploid	heterozygous	200	516	495	21	113	95.9%	80.1%
tetraploid	simplex	50	131	129	2	488	98.5%	20.9%
tetraploid	simplex	100	421	408	13	205	96.9%	66.0%
tetraploid	simplex	150	491	472	19	138	96.1%	76.4%
tetraploid	simplex	200	503	487	16	122	96.8%	78.8%
tetraploid	duplex	50	422	406	16	204	96.2%	65.7%
tetraploid	duplex	100	494	473	21	135	95.7%	76.5%
tetraploid	duplex	150	523	504	19	105	96.4%	81.6%
tetraploid	duplex	200	523	504	19	104	96.4%	81.6%
tetraploid	triplex	50	488	470	18	140	96.3%	76.1%
tetraploid	triplex	100	530	508	22	99	95.8%	82.2%
tetraploid	triplex	150	521	501	20	110	96.2%	81.1%
tetraploid	triplex	200	517	496	21	111	95.9%	80.3%
tetraploid	quadruplex	50	502	487	15	122	97.0%	78.8%
tetraploid	quadruplex	100	519	499	20	110	96.1%	80.7%
tetraploid	quadruplex	150	538	519	19	90	96.5%	84.0%
tetraploid	quadruplex	200	531	509	22	100	95.9%	82.4%

If the dataset is diploid, our model classifies a TE insertion as homozygous if TAF>0.875, as heterozygous if 0.375<TAF≤0.625, and as unclassified if neither of these conditions are met. If the expected ploidy is tetraploid, our model classifies a TE insertion as quadruplex if TAF>0.875, as triplex if 0.625<TAF≤0.875, as duplex if 0.375<TAF≤0.625, and as simplex if TAF≤0.375. To evaluate this classifier, we used synthetic datasets generated from ISO1 and A4 genome assemblies described in Section 4.3.7. The benchmark results were summarized in Table 4.2 and Table 4.3 for diploid and tetraploid genome, respectively. Our classifier had over 91% precision for diploid genomes and over 88% precision for tetraploid genomes under all coverage levels.

Table 4.2: **Performance benchmark for intra-sample TE insertion zygosity classifier on diploid genome.** TELR predictions on synthetic data from ISO1 and A4 genome assemblies were used as input for the classifier. Zygosity represents whether the simulated reads were generated from both ISO1 and A4 (heterozygous) or ISO1 only (homozygous). Precision represents the proportion of predictions being correctly classified as heterozygous or homozygous by the classifier.

Ploidy	Zygosity	Coverage	Total_TAF_Pred	Num_Homozygous_Pred	Num_Heterozygous_Pred	Num_Unclassified_Pred	Precision
diploid	homozygous	50	481	479	0	2	99.6%
diploid	homozygous	100	510	509	1	0	99.8%
diploid	homozygous	150	534	533	1	0	99.8%
diploid	homozygous	200	521	520	1	0	99.8%
diploid	heterozygous	50	399	4	366	29	91.7%
diploid	heterozygous	100	481	2	469	10	97.5%
diploid	heterozygous	150	496	4	484	8	97.6%
diploid	heterozygous	200	495	4	479	12	96.8%

Table 4.3: **Performance benchmark for intra-sample TE insertion zygosity classifier on tetraploid genome.** TELR predictions on synthetic data from ISO1 and A4 genome assemblies were used as input for the classifier. Zygosity represents whether the simulated reads were generated from both ISO1 and A4 (heterozygous) or ISO1 only (homozygous). Precision represents the proportion of predictions being correctly classified as heterozygous or homozygous by the classifier.

Ploidy	Zygosity	Coverage	Total_TAF_Pred	Num_Simplex_Pred	Num_Duplex_Pred	Num_Triplex_Pred	Num_quadruplex_Pred	Precision
tetraploid	simplex	50	127	112	15	0	0	88.2%
tetraploid	simplex	100	401	391	9	0	1	97.5%
tetraploid	simplex	150	467	458	8	0	1	98.1%
tetraploid	simplex	200	474	466	6	0	2	98.3%
tetraploid	duplex	50	400	11	360	22	7	90.0%
tetraploid	duplex	100	468	7	445	11	5	95.1%
tetraploid	duplex	150	498	5	484	4	5	97.2%
tetraploid	duplex	200	497	4	483	8	2	97.2%
tetraploid	triplex	50	476	0	19	437	20	91.8%
tetraploid	triplex	100	523	0	10	493	20	94.3%
tetraploid	triplex	150	513	0	3	485	25	94.5%
tetraploid	triplex	200	514	2	3	483	26	94.0%
tetraploid	quadruplex	50	496	1	1	0	494	99.6%
tetraploid	quadruplex	100	510	0	1	0	509	99.8%
tetraploid	quadruplex	150	526	0	1	0	525	99.8%
tetraploid	quadruplex	200	516	1	1	1	513	99.4%

4.3.9 Evaluate TELR on TE sequence quality

We used the same synthetic datasets used in Section 4.3.7 to evaluate quality of TE sequences assembled by TELR. In theory, a good predictor should produce local contig assemblies that can be perfectly aligned to the corresponding TE loci in the ISO1 genome assembly. For a

given TELR run, each reported TE sequence plus 500bp flanks upstream and downstream of TE region in the local contig assembly (later referred to as "TELR TE locus") was aligned to the ISO1 genome assembly. All TELR TE loci can be uniquely aligned to ISO1 (Figure 4.3). Next, we checked whether each region in ISO1 that was covered by a unique TELR TE locus intersects a curated annotation of the same TE family. The results are summarized in 4.4. The majority of TELR TE sequences have at least 95% of their region aligned to ISO1. If a match is found, we then compared the TELR TE sequence with corresponding TE sequence in ISO1 based on curated TE annotation. The SNVs and INDELs between TELR TE sequences and corresponding ISO1 TE sequences are summarized in Figure 4.16 and 4.17, respectively.

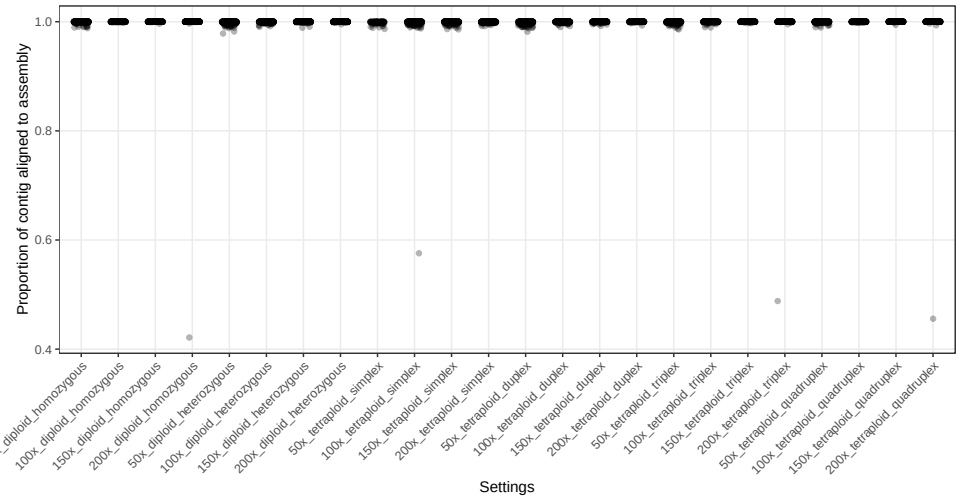

Figure 4.3: **Distribution on proportions of TELR TE loci aligned to the ISO1 genome assembly using synthetic long read sequencing data.** For a given TELR run using simulated data with a specific coverage, ploidy and zygosity setting, each predicted TE sequence plus 500bp flanking sequences on 5' and 3' side of the TE locus in the local contig assembly (referred to as "TELR TE locus") was aligned to the ISO1 genome assembly. The proportion of each TELR TE locus aligned to ISO1 was computed as "Proportion of contig aligned to assembly" in the y axis. See details in Section 4.3.9.

4.3.10 Cross-validation of TELR results using short-read methods

To cross-validate results obtained by TELR, we employed seven short-read TE detection methods implemented in McClintock (v2.0; revision 93369ef) [175] that output TAF values,

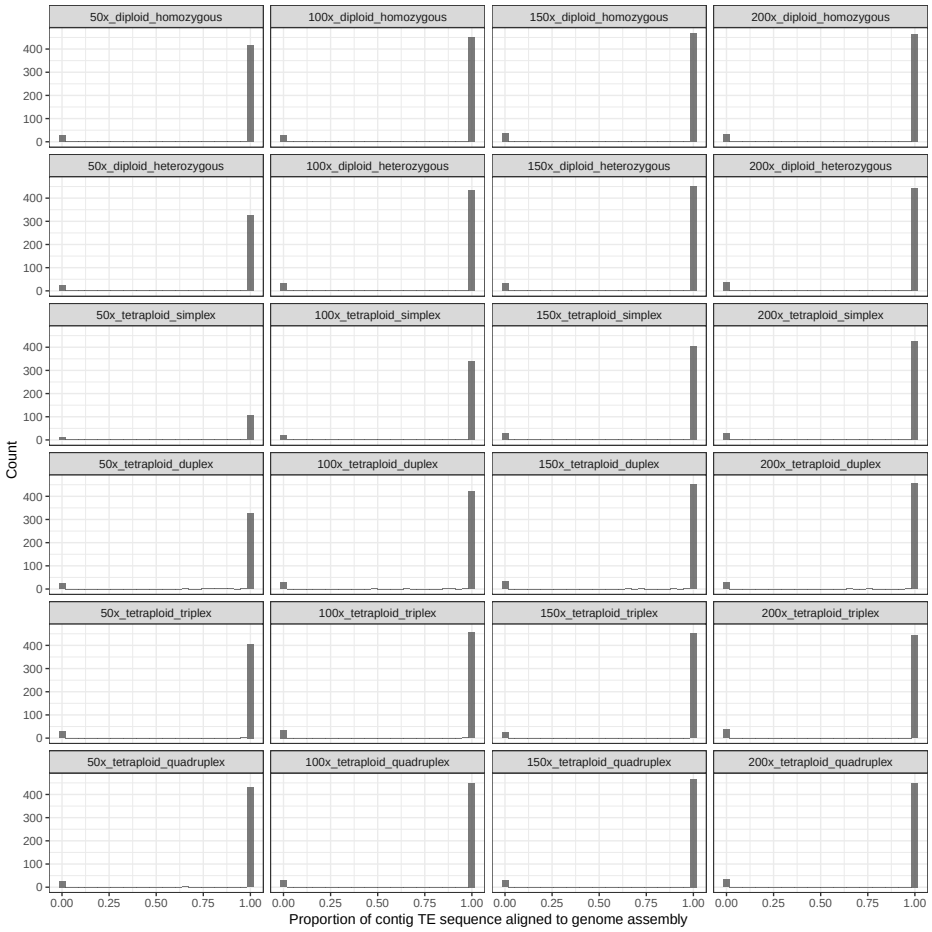

Figure 4.4: **Distribution on proportions of TELR TE sequences aligned to the corresponding TE sequences in ISO1 using synthetic long read sequencing data.** For a given TELR run using simulated data with a specific coverage, ploidy and zygosity setting, each predicted TE sequence plus 500bp flanking sequences on 5' and 3' side of the TE locus in the local contig assembly (referred to as "TELR TE locus") was aligned to the ISO1 genome assembly. The proportion of TE sequence in the TELR TE locus aligned to the corresponding TE sequences in ISO1 was computed as "Proportion of contig TE sequence aligned to genome assembly" in the y axis. If the TE family predicted by TELR doesn't match the annotated TE family in ISO1, the proportion will then be labelled as 0. See details in section 4.3.9.

which include ngs_te_mapper2 [88], PoPoolationTE [115], PoPoolationTE2 [116], RetroSeq [112], TEFLoN [3], TEMP [259], and TEMP2 [249]. Linked-read data obtained for S2R+ and A4 was barcode-trimmed with LongRanger (v2.2.2; basic pipeline) [255], de-interleaved, and

trimmed to 100bp using fastp (v0.20.0; –max_len1 100 --max_len2 100 --length_required 100) [53]. This data was downsampled to ∼50X mean mapped read depth for S2R+ (74,648,362 reads) and A4 (76,045,544 reads) before being used as input in McClintock to generate non-redundant non-reference TE insertion predictions. The genome-wide TAF profiles for S2R+ and A4 using TELR and seven short-read TE detection methods are shown in Figure 4.5 and 4.6, respectively.

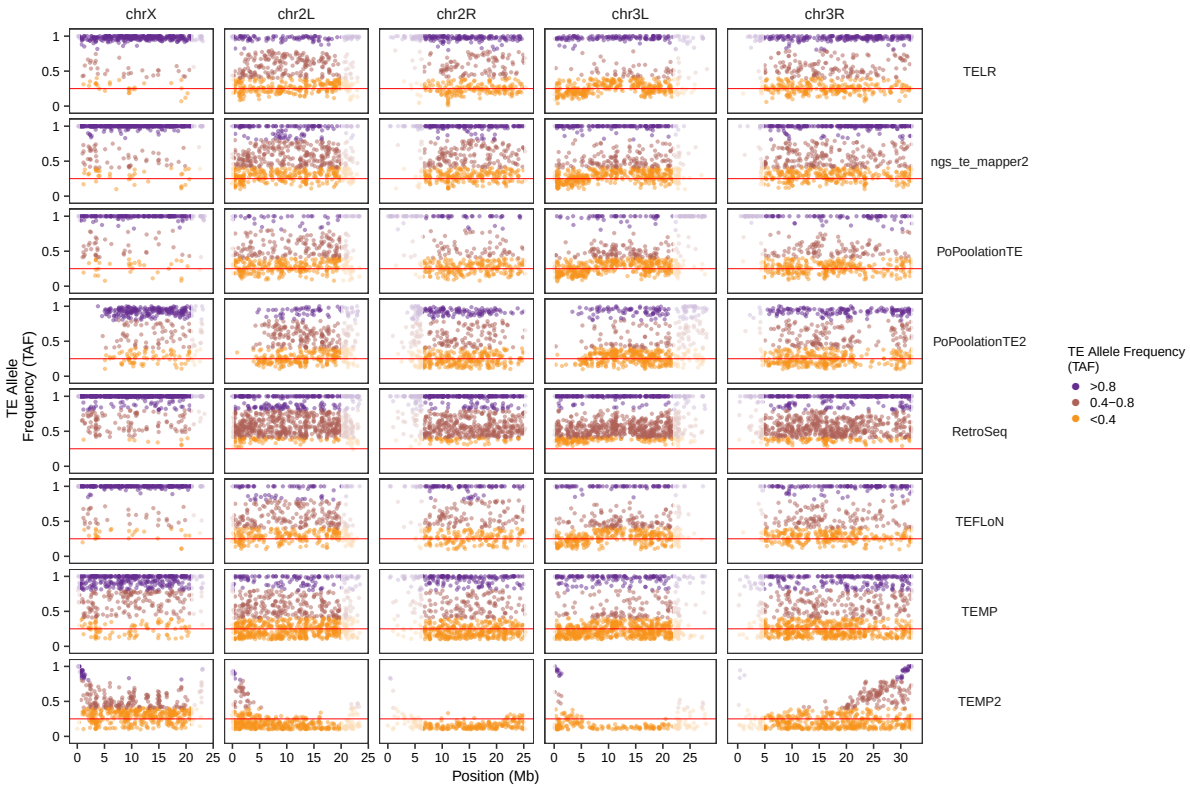

Figure 4.5: **Comparison of genome-wide distribution of TAF profiles for S2R+ between TELR and short-read methods.** Only insertions with flank alignment support for both sides, and for which the TAF could be calculated were included.

4.3.11 Construction of phylogenetic trees using TE sequences from TELR

TE sequences predicted, assembled, and polished by TELR on S2R+ and DSPR dataset were filtered using the following criteria: 1) Sequences from A2 were excluded due to potential inversion-induced gain of heterozygosity (see Section 4.5 for details). 2) Sequences from

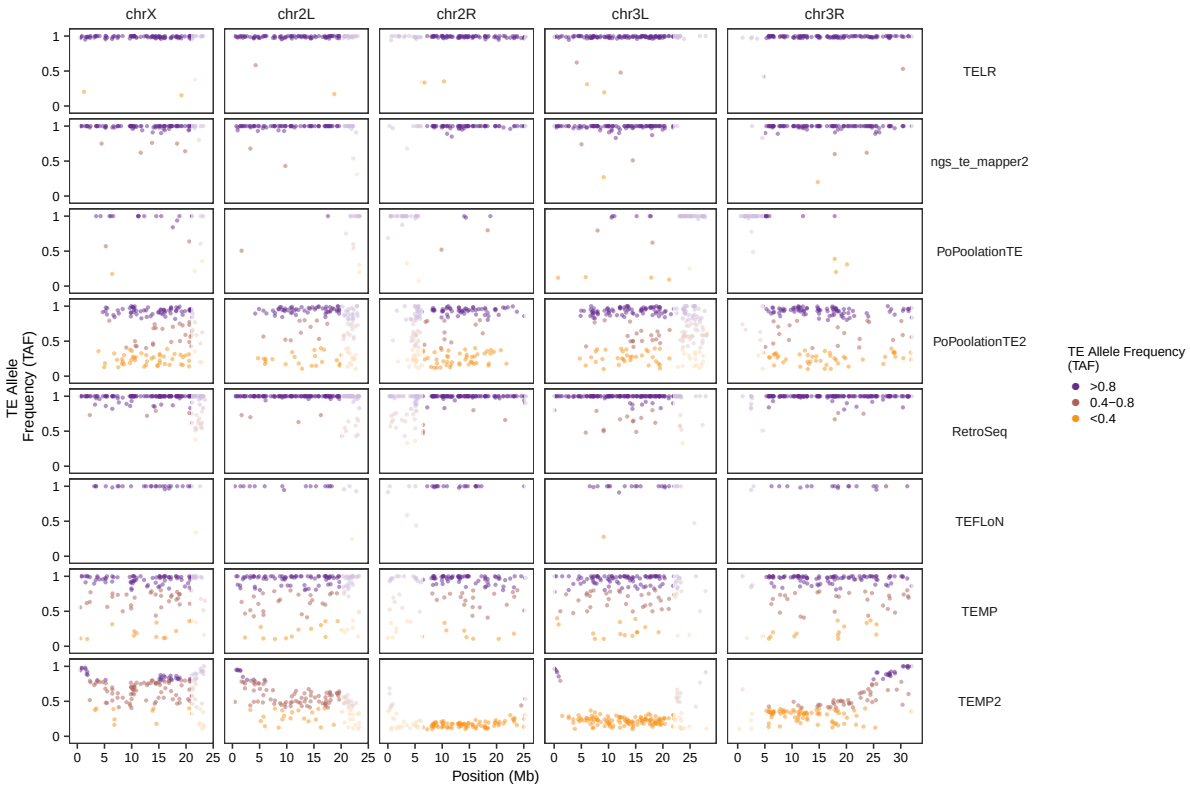

Figure 4.6: **Comparison of genome-wide distribution of TAF profiles for A4 between TELR and short-read methods.** Only insertions with flank alignment support for both sides, and for which the TAF could be calculated were included.

A7 were excluded due to potential sample contamination (see Section 4.5 for details). 3) Sequences from chromosome X were excluded due to lower coverage compared to autosomes and loss of heterozygosity (LOH) events. 4) Exclude sequences from low recombination regions using boundaries defined by Cridland *et al.* 2013 [60] lifted over to dm6 coordinates. Normal recombination regions included in our analyses were defined as chrX:405967–20928973, chr2L:200000–20100000, chr2R:6412495–25112477, chr3L:100000–21906900, chr3R:4774278–31974278. We restricted our analysis to normal recombination regions since low recombination regions have high reference TE content which reduces the ability to predict non-reference TE insertions [29, 154]. 5) Only full-length TE elements

based on canonical sequences were included. We first calculated the ratio between each TELR sequence length and the corresponding canonical sequence length. Next, we filtered TELR sequences for full-length copies using a 0.95-1.05 ratio cutoff. 6) Only sequences with both 5' and 3' flanks mapped to reference genome were included.

TELR sequences from each family were aligned with MAFFT (v7.487) [108]. The multiple sequence alignments (MSAs) were filtered by trimAI (v1.4.rev15; parameters: -resoverlap 0.75 -seqoverlap 80) to remove spurious sequences. The filtered MSAs were used as input to IQ-TREE (v2.1.4-beta; parameters: -m GTR+G -B 1000) [165] to generate maximum likelihood trees.

4.4 Results

4.4.1 Fragmented assemblies yield variable estimates of TE content in the S2R+ genome

To better understand the structure and evolution of TEs in the S2R+ cell line genome, we initially used a *de novo* assembly-based approach by generating PacBio long-read (132X average depth) and 10x Genomics linked-read (89X average depth) sequencing data and assembled these data using a variety of state-of-the-art WGA software [14, 57, 118, 240, 203, 117]. All S2R+ whole-genome assemblies (WGAs) using long reads (Canu, FALCON-Unzip, wtdbg2, and Flye) or linked reads (Supernova) had better contiguities compared to a SPAdes assembly using Illumina paired-end short reads (Figure 4.7A; Table 4.4). However, S2R+ WGAs from different sequencing technologies and assemblers varied substantially in their contiguities and levels of duplicated BUSCOs (Figure 4.7A,B; Table 4.4). The S2R+ assemblies using Canu and FALCON-Unzip are among the longest and display the highest level of BUSCO duplication (Figure 4.7A,B; Table 4.4). Unlike Canu, FALCON-Unzip explicitly separates phased heterozygous regions into a set of contigs called haplotigs [57]. Combining the haplotigs with the primary FALCON-Unzip assembly resulted in highest level of BUSCO duplication

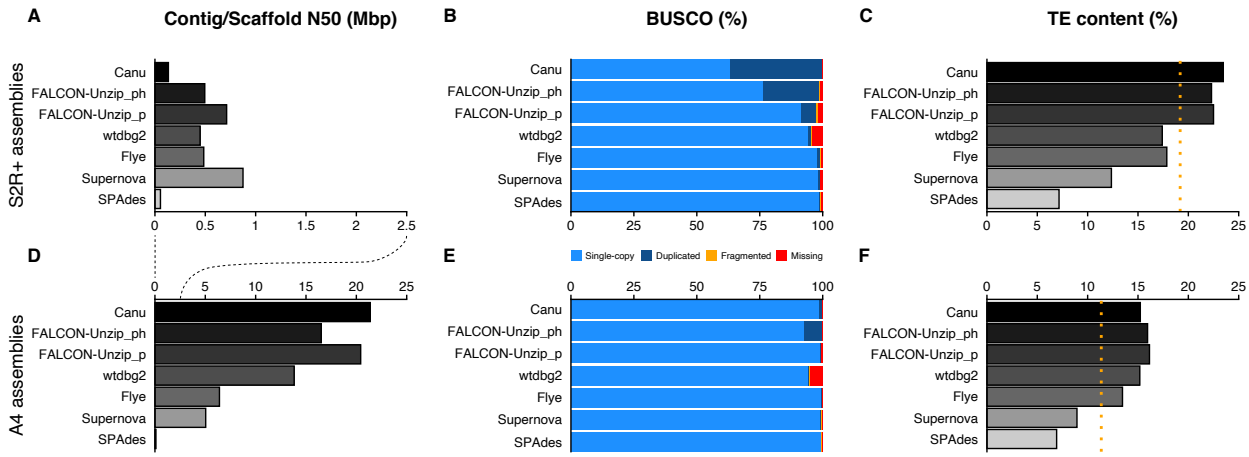

Figure 4.7: **Lower contiguity, and higher BUSCO duplication and TE content in whole-genome assemblies of S2R+ compared to those from an inbred fly strain.** Contig (Canu, FALCON-Unzip, and wtdbg2), and Scaffold (Flye, Supernova, and SPAdes) N50 values for S2R+ (A) and A4 (D) whole-genome assemblies. BUSCO (Benchmarking Universal Single-Copy Orthologs) analysis with the Diptera gene set from OrthoDBv10 on S2R+ (B) and A4 (E) assemblies. RepeatMasker estimates of TE content in WGAs of S2R+ (C) and A4 (F). Orange dotted lines in C and F represent RepeatMasker estimates of TE content from raw Illumina reads. FALCON-Unzip_p = primary contigs, FALCON-Unzip_ph = primary + haplotigs. Note that the scale bar is different in A and D.

("FALCON-Unzip_ph" in Figure 4.7B,E). This result suggested that many regions of the S2R+ genome are heterozygous, which adds more complexity to its tetraploid nature.

Table 4.4: **Statistics for S2R+ genome assemblies.**

	Canu	Falcon-Unzip_p	Falcon-Unzip_h	Falcon-Unzip_ph	wtdbg2	Flye	Supernova	SPAdes
Assembly size (bp)	288,352,585	166,182,448	43,221,666	209,404,114	146,747,034	144,559,338	136,942,585	138,619,786
Contig count	4,388	599	1,132	1,731	1,516	1,258	6,083	144,828
Contig N50 (bp)	133,624	711,969	41,606	494,779	448,074	478,824	78,941	42,786
Scaffold count	4,388	599	1,132	1,731	1,516	1,243	3,687	143,874
Scaffold N50 (bp)	133,624	711,969	41,606	494,779	448,074	484,692	874,516	54,714
GC content (%)	41.49	41.75	41.75	41.75	41.62	41.88	42.2	42.72
TEs (bp)	67,687,512	37,394,886	9,297,197	46,692,083	25,542,314	25,822,786	16,928,725	9,923,944
TEs (%)	23.47	22.50	21.51	22.30	17.41	17.86	12.36	7.16
BUSCO (%)								
Complete	99.6	97.1	25.6	98.5	95.1	99.0	98.7	98.7
Single-copy	63.1	91.4	25.1	76.3	93.8	97.6	98.0	98.5
Duplicated	36.5	5.7	0.5	22.2	1.3	1.4	0.7	0.2
Fragmented	0.1	0.8	2.4	0.2	0.4	0.3	0.3	0.6
Missing	0.3	2.1	72.0	1.3	4.5	0.7	1.0	0.7

FALCON-Unzip_p = primary contigs; FALCON-Unzip_h = haplotigs; FALCON-Unzip_ph = primary + haplotigs.

The N50s for all S2R+ WGAs are less than 1 Mbp, which is significantly smaller than the chromosome size in the *Drosophila* reference genome [95]. To assess how the WGAs for the S2R+ cell line compared to WGAs for whole flies of inbred stocks, we also assembled genome for a highly inbred *D. melanogaster* strain called A4 using available PacBio long-read data (110X average depth) [50] and a new 10x Genomics linked-read dataset generated in this study (118X average depth). Using identical assembly software and parameters, we found that A4 WGAs have reference-grade contiguities and exhibit lower variation in levels of BUSCO duplication than the WGAs for the S2R+ cell line (Figure 4.7D,E; Table 4.4 and 4.5). Given that the A4 strain is diploid homozygous [50], these results suggest that the highly fragmented WGAs for the S2R+ cell line are caused by the polyploid and heterozygous nature of the cell line genome.

Table 4.5: **Statistics for A4 genome assemblies.**

	Canu	Falcon-Unzip_p	Falcon-Unzip_h	Falcon-Unzip_ph	wtdbg2	Flye	Supernova	SPAdes
Assembly size (bp)	141,737,450	141,292,095	13,036,468	154,328,563	137,203,985	135,706,987	126,833,864	135,820,998
Contig count	181	107	289	396	431	234	3115	119370
Contig N50 (bp)	21,369,333	20,430,351	47,610	16,510,272	13,821,893	5,389,879	182,420	74,783
Scaffold count	181	107	289	396	431	229	1,818	118,640
Scaffold N50 (bp)	21,369,333	20,430,351	47,610	16,510,272	13,821,893	6,405,908	5,040,789	98,019
GC content (%)	42.07	42.14	41.89	42.12	41.83	42.11	42.27	42.33
TEs (bp)	21580457	22,804,030	1,803,901	24,607,939	20,832,191	18,259,520	11,339,728	9,393,794
TEs (%)	15.23	16.14	13.84	15.95	15.18	13.46	8.94	6.92
BUSCO (%)								
Complete	99.4	99.2	8.7	99.4	94.6	99.5	99.2	99.2
Single-copy	98.5	98.7	8.6	92.5	94.1	99.1	98.9	99.1
Duplicated	0.9	0.5	0.1	6.9	0.5	0.4	0.3	0.1
Fragmented	0.2	0.2	0.8	0.2	0.2	0.2	0.3	0.3
Missing	0.4	0.6	90.5	0.4	5.2	0.3	0.5	0.5

FALCON-Unzip_p = primary contigs; FALCON-Unzip_h = haplotigs; FALCON-Unzip_ph = primary + haplotigs.

Estimates of TE content from all WGAs and unassembled short reads varied substantially for both S2R+ and in A4 (Figure 4.7C,F; Table 4.4 and 4.5). Compared to estimates based on unassembled short reads (dotted line in Figure 4.7C,F), long-read WGAs of both the S2R+ and A4 genomes always gave higher estimates of TE content, while WGAs from short reads always gave lower estimates. Regardless of read technologies and assemblers used to produce WGAs, higher estimates of TE content can be observed in S2R+ relative to A4 (Figure 4.7C,F; Table 4.4 and 4.5). Given the RepeatMasker analysis of the raw Illumina reads is

not affected by the assembly quality and therefore give the least biased overall estimates, we estimated the TE content in S2R+ is 19.2%, which is substantially higher compared to 11.4% in A4 (Figure 4.7C,F). These results provide further support for the conclusion that *Drosophila* cell lines have increased TE content relative to whole flies [189, 100, 193].

In addition to differences in overall TE content, we observed much higher variation in the abundance of different TE families in S2R+ compared to A4 (Figure 4.8). Moreover, the rank order abundance of TE families differed among WGAs of the same long-read data in S2R+, but less so in A4 (Figure 4.8). Due to the poor contiguity and high variations of TE content among WGAs built from S2R+ long-read and linked-read data using multiple assemblers, we concluded that an alternative approach was necessary to reliably study TE content in the S2R+ genome that is well adapted to the polyploid and heterozygous nature of cell line genomes.

4.4.2 A novel long-read approach reconstructs TE sequences and allele frequencies in the S2R+ genome

To circumvent the issue of fragmented WGAs that limit the analysis of TE content in polyploid and heterozygous cell line genomes, we developed a new reference-based method called "TELR" (Transposable Elements from Long Reads; `https://github.com/bergmanlab/telr`) that allows the identification, assembly, and allele frequency estimation of non-reference TE insertions using long-read data (Figure 4.9). Briefly, TELR first aligns long reads to a reference genome to identify new insertions using Sniffles [216]. The insertions identified by Sniffles are then filtered by aligning putative insertion sequences to a TE library to generate TE insertion candidate loci. For each candidate locus, TELR performs a local assembly using all reads that support the putative TE insertion event. Finally, TELR extracts TE sequence from each assembled contig, identifies the precise insertion coordinates and estimates TAF (See details in Section 4.3.6).

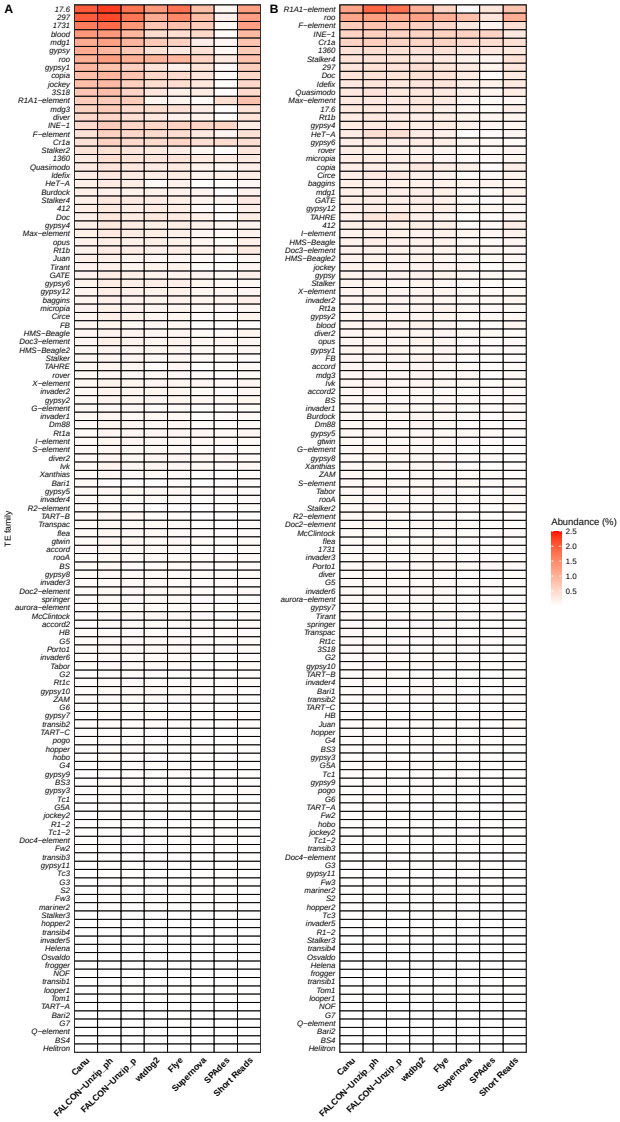

Figure 4.8: **TE abundance varies substantially between assembly methods.** TE abundance was estimated for different assemblies and directly from raw Illumina reads in S2R+ (A) and A4 (B) using RepeatMasker and the curated canonical *D. melanogaster* TE library.

Using TELR we identified 3070 non-reference TE insertions in S2R+, which is a ∼5-fold increase relative to the number of insertions identified in A4 (624; Figure 4.10). To control for differences in read coverage and read length distributions between S2R+ and A4 datasets, we also ran TELR on normalized datasets for S2R+ and A4 with a mean mapped read

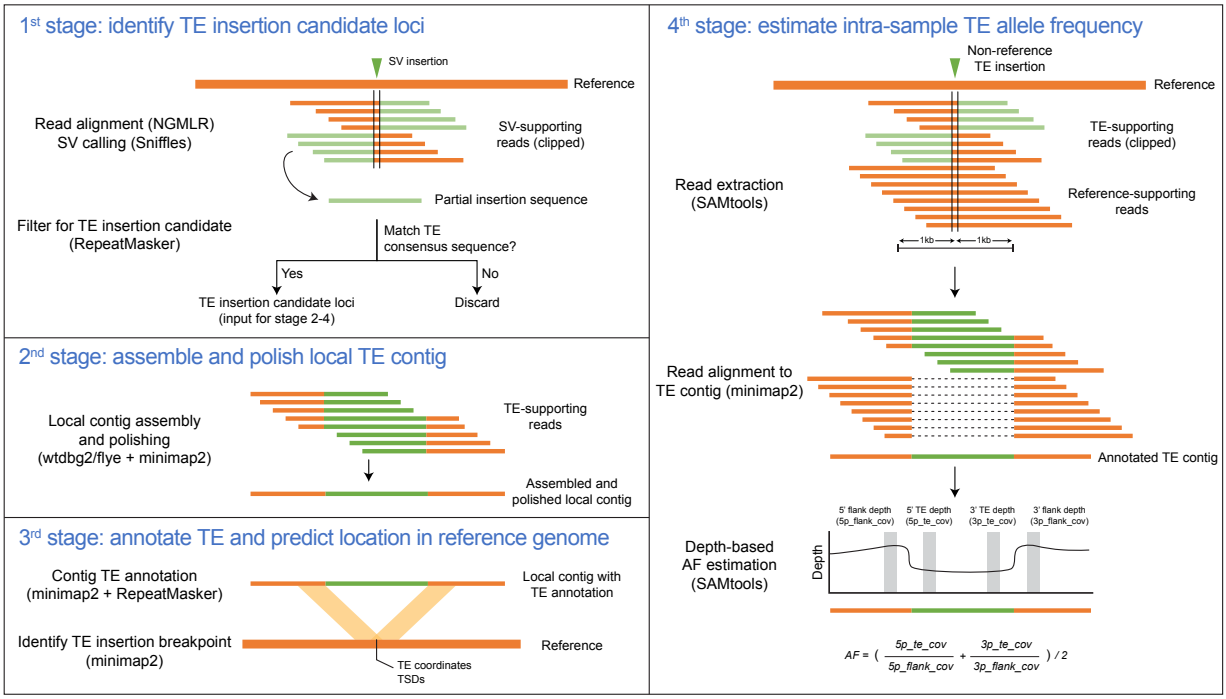

Figure 4.9: **TELR workflow to predict non-reference TE and estimate intra-sample allele frequency.** TELR is a non-reference transposable element (TE) detector from long read sequencing data. The TELR pipeline consists of four main stages. In the first stage, TELR aligns long reads to a reference and identify insertions using Sniffles [216]. TELR then screens for non-reference TE insertion candidate locus by computing nucleotide similarity between partial insertion sequence provided by Sniffles and TE consensus sequences. In the second stage, TELR use SV-supporting reads from Sniffles to assemble and polish local contig using wtdbg2 [203] and minimap2 [136]. In the third stage, The TE boundaries and family are annotated in the local contig using minimap2 and RepeatMasker, and the TE flanking sequences are used to determine the TE coordinates and target-site duplications by mapping to the reference genome with minimap2. In the fourth stage, TELR determines the intra-sample allele frequency of each TE insertion by extracting all reads in a 2 kb span around the insertion locus and aligning them to the TE contig. The mapped read depth over TE and flanking sequences are then used to calculate the insertion intra-sample allele frequency.

depth of ∼25, 50, and 75X. Despite a drop in the number of predictions in the normalized S2R+ data relative to the full dataset, TELR still predicted substantially more TEs in S2R+ compared to A4 at all coverage levels (Figure 4.11). This analysis also revealed that, unlike A4, which displays a plateauing of the number of non-reference insertions at a read depth of

50X, detection of non-reference TEs in S2R+ is likely not saturated even at 75X. Therefore, in order to maximize TE prediction sensitivity, we used the complete non-normalized Pacbio data for S2R+ and all whole-fly strains in subsequent analyses. Comparison on the number of family-specific non-reference TE insertions predicted by TELR in S2R+ and A4 revealed a number of TE families that are enriched in S2R+ compared to A4 (Figure 4.13A). These TE families specifically enriched in S2R+ consist mostly of long terminal repeat (LTR) retrotransposons except for *Jockey* and *Juan*, which are long interspersed nuclear elements (LINEs) (Figure 4.13A; Figure 4.12).

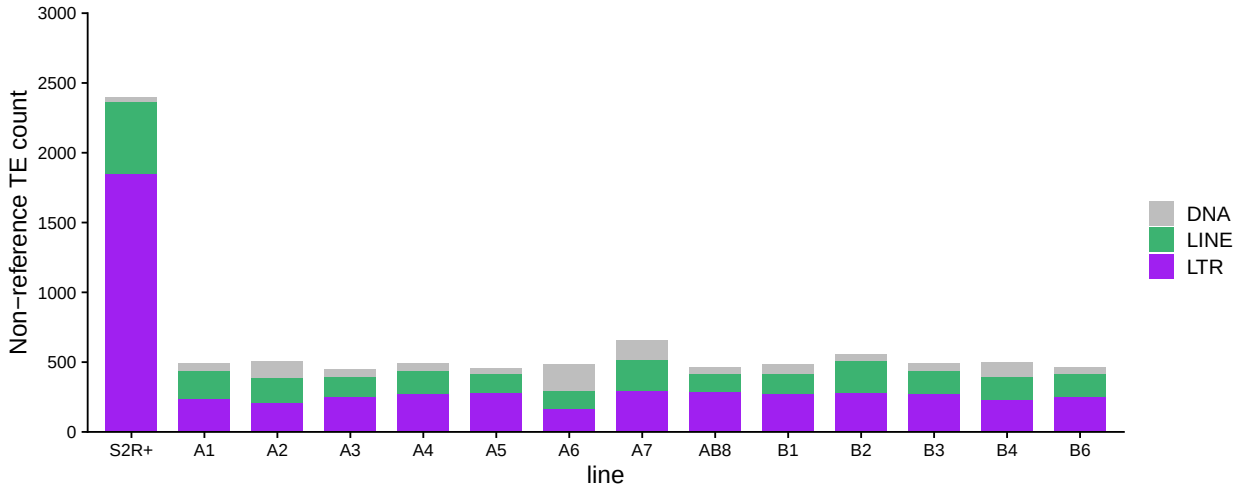

Figure 4.10: **Increased number of non-reference TE insertions in S2R+ compared to inbred fly stocks derived from natural populations.** Number of non-reference TE insertions predicted by TELR for S2R+ and inbred stocks from geographically diverse natural fly populations [50]. Only insertions from normal recombination regions (see details in Section 4.3.11), with flank alignment support for both sides, and for which the TAF could be calculated were included in this analysis.

We also used TELR to predict non-reference TEs on a panel of 13 geographically diverse *D. melanogaster* inbred strains from the *Drosophila* Synthetic Population Resource (DSPR) [50]. This analysis revealed that S2R+ includes a higher number of non-reference TE insertions than any of the fly strains surveyed (Figure 4.10). Partitioning TE predictions by family reveals that only a small fraction of the 125 curated TE families in *D. melanogaster* make up the bulk of non-reference insertions in S2R+ (Figures 4.13; Figure 4.12). In S2R+, roughly

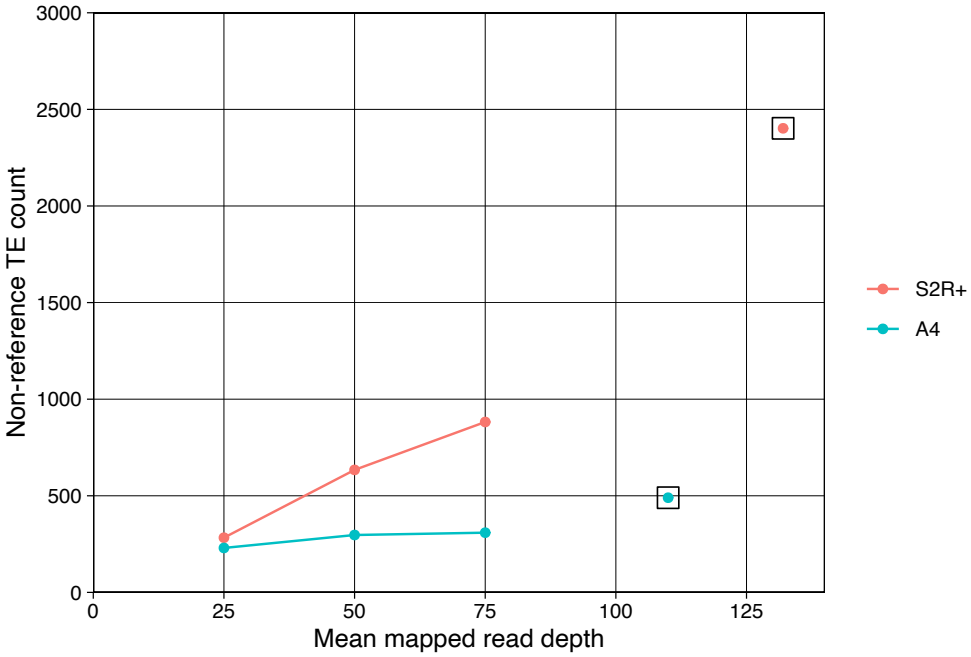

Figure 4.11: **Effect of mean mapped read depth in the number of non-reference TE predictions.** Comparison of the number of non-reference TEs predicted by TELR using length- and depth-normalized long-read datasets for S2R+ and A4. Reads from both datasets were split into 8 kb segments and sub-sampled to achieve the desired mean mapped read depth. The boxed points indicate the number of TE predictions made by TELR when the full non-normalized data was supplied. Only insertions from normal recombination regions (see details in Section 4.3.11), with flank alignment support for both sides, and for which the TAF could be calculated were included in this analysis.

6% of TE families account for 75% of all non-reference TE insertions, and they are composed mostly of LTR elements (Figure 4.10). In comparison, roughly 9-13% of TE families contribute $\geqslant$ 75% of all non-reference TE insertions in the DSPR strains, and they represent a greater variety of TE types including LTRs, LINEs, and DNA transposons. We also observed strain-specific TE expansions in DSPR strains with some strains displaying over 3-fold increase in the number of non-reference TE insertions for specific families compared to the mean values among all samples, i.e. A2 (*1360*), A6 (*hopper*) and B2 (*Doc*, *Quasimodo*) (Figure 4.12).

Figure 4.12: **Increased TE abundance and unique profile of TE activation in S2R+ compared to multiple fly strains.** Heatmap showing the copy number of non-reference insertions from multiple TE families in S2R+ and inbred fly strains from the Drosophila Synthetic Population Resource [114]. TE families are ordered firstly according to their abundance in S2R+, and secondly alphabetically. Only insertions from normal recombination regions (see details in Section 4.3.11), with flank alignment support for both sides, and for which the TAF could be calculated were included in this analysis.

An important feature of TELR is the ability to estimate the intra-sample allele frequency of non-reference TE insertions, which allowed us to observe drastic differences between S2R+ and A4 in the intra-sample TE allele frequency (TAF) across chromosomes. As expected for a highly inbred fly stock [114], non-reference TEs in DSPR strains are mostly homozygous,

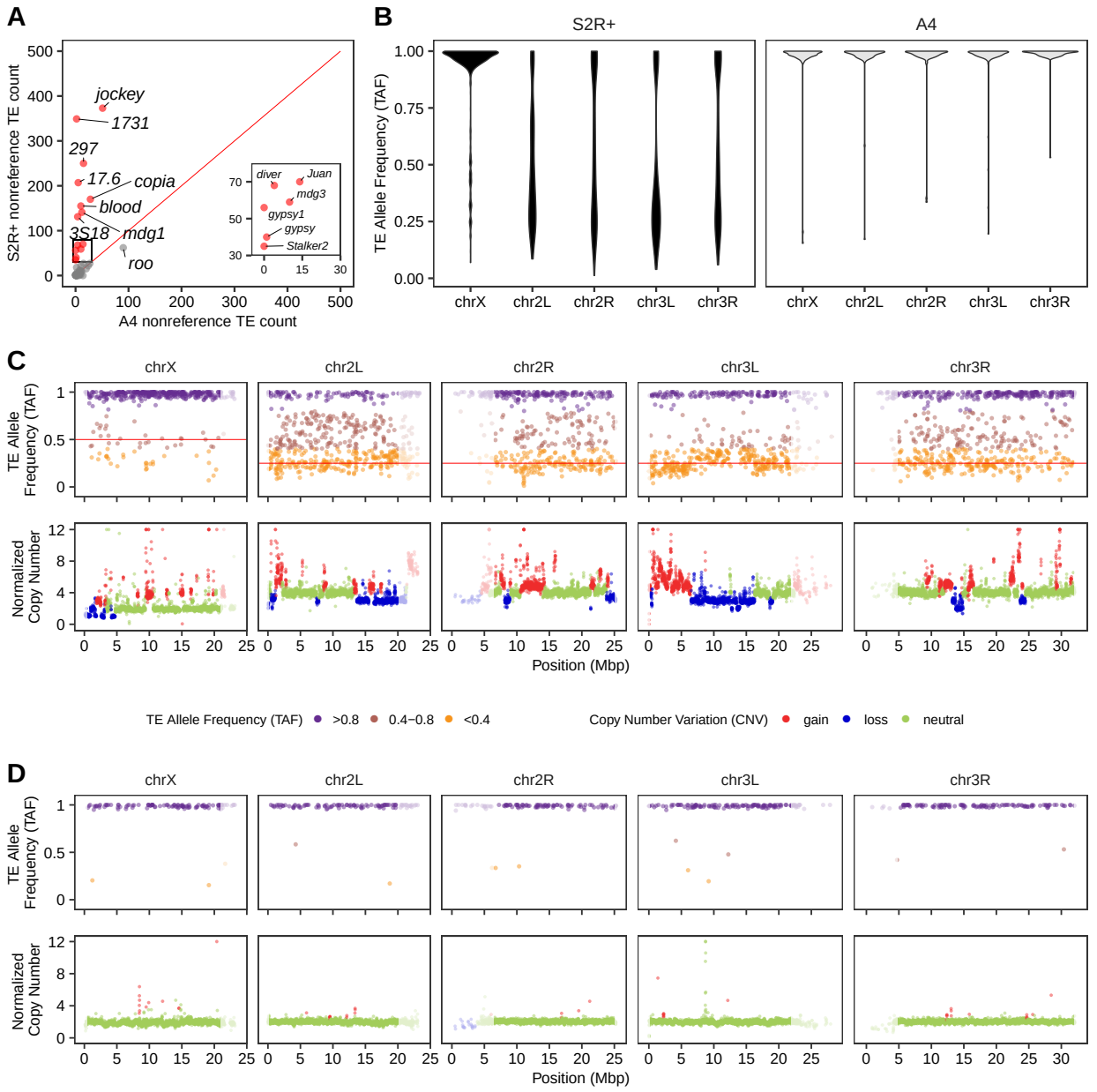

Figure 4.13: **Long-read non-reference TE prediction with TELR reveals multiple families amplified during cell culture. A** Number of non-reference TE predictions made by TELR for S2R+ and A4 with the 8 most abundant families in S2R+ highlighted in red. **B** TE allele frequency (TAF) distribution by chromosome arm. **C-D** Genome-wide TAF and copy number variation (CNV) profiles for S2R+, and A4, respectively.

with TAF clustered at ∼1 (Figure 4.13B; Figure 4.14). However, we found two curious exceptions on the homozygous TAF patterns in DSPR strains. First, strain A2 displays mostly heterozygous TE insertions across chromosome arm 3R, which coincides with the presence of a known heterozygous chromosomal inversion in this strain, *In(3R)P*, that prevents full inbreeding [114]. Second, TELR reveals a genome-wide profile of putative heterozygous TE insertions in *D. melanogaster* strain A7, with TAF enriched at ∼0.25 and ∼0.75 (Figure 4.14). This TAF pattern is unusual since strain A7 is thought to be fully inbred and devoid of large chromosomal inversions. We hypothesized that the bi-modal TAF profile in strain A7 could be indicative of contamination with data from a different fly strain. Indeed, intersecting TELR predictions between strain A7 and other DSPR strains revealed a large overlap between strain A7 and strain B3. Moreover, the shared TE insertions between strains A7 and B3 have TAF enriched at ∼0.25, which could be explained by contaminated data from B3 taking up ∼25% of the A7 data we analyzed. Finally, the WGA for DSPR strains reported in Chakraborty *et al.* 2019 [50] suggested that strain A7 has the highest level of BUSCO duplication, highest assembly length, and highest number of scaffolds among all DSPR strains. Together, these results suggest that the long reads for strain A7 reported in Chakraborty *et al.* 2019 [50] was likely contaminated with reads from strain B3. The results also highlighted the utility of using TAF estimated by TELR to clarify the identity of high-throughput long-read dataset and identify sample contamination.

In contrast to DSPR strains exhibiting genome-wide TAFs that are enriched at 1, TE insertions in S2R+ display a wide range of allele frequencies (Figure 4.1). Despite the fact that the the X chromosome of S2R+ is diploid, TAF in the X chromosome of S2R+ clustered at 1 (Figure 4.13B). This could be explained by a recent loss of heterozygosity (LOH) event in the X chromosome of S2R+ through mitotic recombination. This explanation is plausible since a previous study has shown that LOH could occur and shape TAF profiles in the *Drosophila* cell lines [88]. Importantly, we observed a clear enrichment of TAFs in S2R+ at ∼0.25 in autosomes (Figure 4.1), which can be most parsimoniously explained as

Figure 4.14: **Genome-wide distribution of TE Allele Frequency (TAF) profiles for S2R+ and *D. melanogaster* inbred fly strains.** Strain A2 has a known heterozygous inversion, *In(3R)P*, that prevents the complete inbreeding in chr3R [114]. Only insertions with flank alignment support for both sides, and for which the TE Allele Frequency (TAF) could be calculated were included.

somatic insertions after the S2 genome became tetraploid following initial cell line establishment [213, 129]. An alternative hypothesis (Figure 4.15) is that these TEs were either ancestrally heterozygous in the diploid Oregon-R lab strain that S2 was established from, or they inserted in the pre-tetraploid stage of S2, and that the ∼0.25 TAF was driven by mitotic

recombination events in the post-tetraploid state of S2 that changed one haplotype from TE-present to TE-absent. Under the alternative hypothesis, assuming that the heterozygous TEs are randomly distributed in two different haplotypes of the Oregon-R/pre-tetraploid state of S2, and that mitotic recombination have the same probability of increasing and decreasing TE allele frequency, we should observe simplex and triplex TE loci in tetraploid regions of the S2R+ cells with ∼0.25 and ∼0.75 TAF, respectively, with equal probability. However, the number of TEs with ∼0.25 TAF is significantly higher than the number of TEs with ∼0.75 TAF in autosomal regions of S2R+ (Figure 4.13), providing evidence against the alternative hypothesis. Thus, we interpreted TEs with ∼0.25 TAF as being enriched in insertions that occurred after tetraploidization and therefore must have occurred during cell culture.

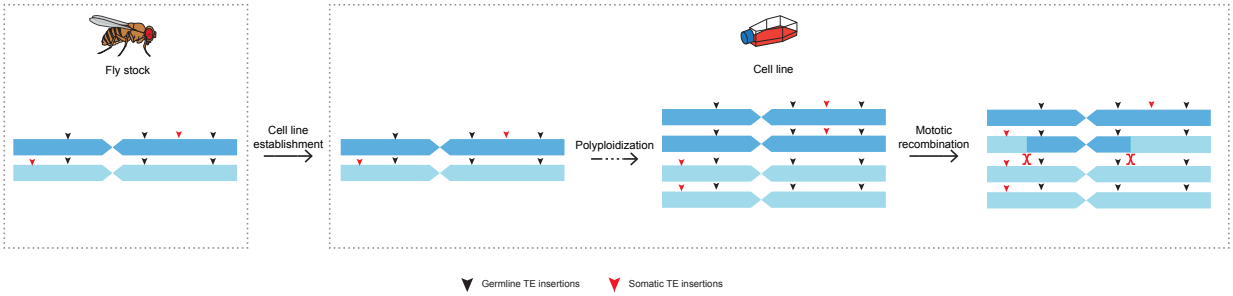

Figure 4.15: **Pre-tetraploid TE insertion followed by post-tetraploid somatic recombinations to explain non-reference TEs in cell line with TAF around 0.25.** The inbred fly stock has diploid genome that includes homozygous variations. Cell lines established from inbred fly stock accumulates heterozygous variations with TAF to be equal to 0.5, then cell lines underwent polyploidization, which is followed by somatic recombinations that change TAF from 0.5 to to 0.25 or 0.75.

All things being equal, post-tetraploid TE insertions should have TAF at ∼0.25. However, the extensive CNVs in S2R+ result in changes in ploidy in the affected genome segments, which could subsequently affect TAFs (Figure 4.13). To facilitate the interpretation of TAF values under varying CNV status, we first computed normalized TE copy number by multiplying TAF by estimated local copy number. This "normalized TE copy number" should approximate the number of TE-bearing haplotypes in the cell line sample. However, they are variable non-integer values since TAFs estimated by TELR are non-integer. We then built

a classifier that can estimate integer TE copy number. Finally, we labelled TE insertions as post-tetraploid if the integer TE copy number is equal to one in regions that are minimally triploid. Our analysis revealed that a significant proportion of non-reference TE insertions identified in S2R+ are due to post-tetraploid somatic transposition, which provide a rich set of resources to study the sequence evolution of TEs that are activated and contribute to overall expansions of TEs in the cell line genome.

4.4.3 TE expansions in *Drosophila* cell lines can be caused by one or more source lineages

The TE expansion events in the S2R+ cell line dominated by post-tetraploid somatic transpositions may originate from a single or multiple source lineages. A previous study used PCR to amplify TEs in S2 cultured cells and provided evidence that all *1731* neocopies in S2 cultured cells were derived from a single source copy that is strongly activated in cultured cells [152]. However, the number of *1731* new insertions are likely underestimated due to the limitations of the PCR technique. In addition, only a single TE family was surveyed in the previous study. To comprehensively test these two hypotheses on all TE families that likely expanded in S2R+ (Figure 4.13A), we took advantage of TELR's ability to assemble TE sequences and used a phylogenetic approach to cluster non-reference TE sequences from S2R+ and 13 whole-fly strains from the DSPR panel (Figure 4.18). Evaluation on the quality of TE sequences reconstructed by TELR on simulated datasets suggested that the TELR produced high-quality TE sequences (Figure 4.16; Figure 4.17), and thus can be reliably used to infer the sequence evolution of TEs amplified in the polyploid cell line genomes like S2R+.

We designed a set of criteria to identify TE expansion events in S2R+ that start from a single source lineage. First, the TE expansion event should be marked by a monophyletic clade in which $\geq$30% of TEs are enriched with post-tetraploid insertions in S2R+. Second, the candidate TE expansion clade should have at least 70% bootstrap support. Using these

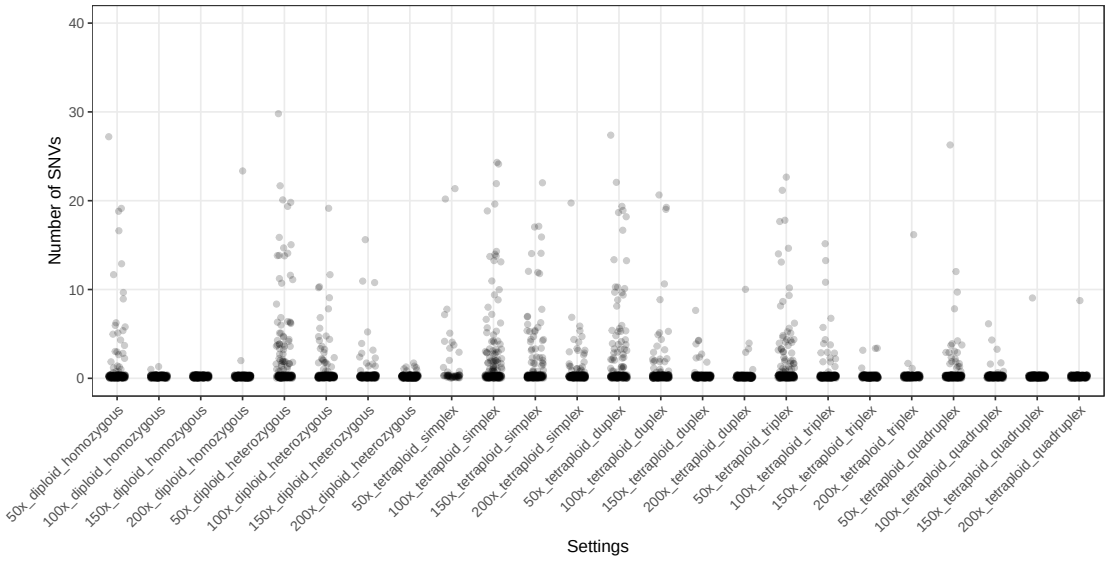

Figure 4.16: **Distribution on number of SNVs between TELR TE sequences and corresponding TE sequences in ISO1 using synthetic long read sequencing data.** For a given TELR run using simulated data with a specific coverage, ploidy and zygosity setting, each predicted TE sequence plus 500bp flanking sequences on 5' and 3' side of the TE locus in the local contig assembly (referred to as "TELR TE locus") was aligned to the ISO1 genome assembly. We then compared TE sequences from TELR with annotated TE sequence from the corresponding region in ISO1 using paftools from minimap2 [136] and calculated the number SNVs. See details in Section 4.3.9.

criteria, we annotated TE expansion events in the sequence phylogeny for each of the 14 TE families that are enriched in S2R+ relative to A4 (Figure 4.13A, red dots). We identified a single TE expansion clade for TE families such as *1731*, *gypsy1*, *diver*, *gypsy*, *mdg3*, and *Stalker2* (Figure 4.18; Figure 4.19), suggesting that the TE expansion events in the S2R+ cell line for these families came from a single source lineage. We also identified multiple TE expansion clades for TE families such as *Jockey*, *Juan*, *copia*, *3S18*, and *mdg1* (Figure 4.18; Figure 4.19), suggesting multiple source lineages for these families. Together, our results revealed that TE expansions in S2R+ can be caused by single or multiple source lineages, and that the pattern of source lineage activation in somatic cell culture is TE family-dependent (Figure 4.18; Figure 4.19).

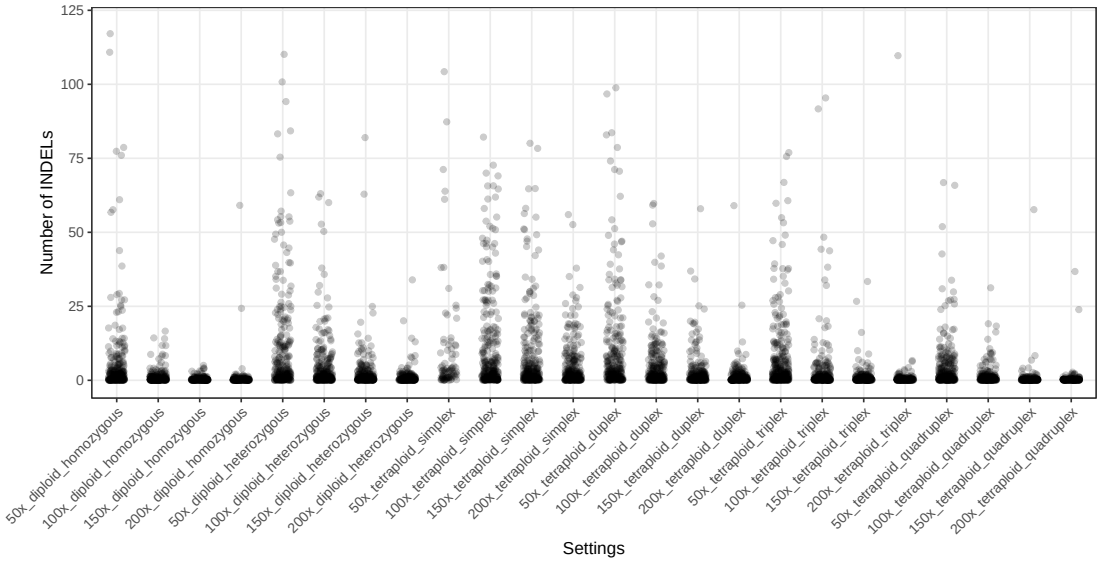

Figure 4.17: **Distribution on number of INDELs between TELR TE sequences and corresponding TE sequences in ISO1 using synthetic long read sequencing data.** For a given TELR run using simulated data with a specific coverage, ploidy and zygosity setting, each predicted TE sequence plus 500bp flanking sequences on 5' and 3' side of the TE locus in the local contig assembly (referred to as "TELR TE locus") was aligned to the ISO1 genome assembly. We then compared TE sequences from TELR with annotated TE sequence from the corresponding region in ISO1 using paftools from minimap2 [136] and calculated the number INDELs. See details in Section 4.3.9.

4.5 Discussion

As an alternative to WGA, several methods have been developed to detect non-reference TEs from long reads using a reference-based approach (Table 4.6). These methods use different strategies for TE detection and report different information about non-reference TEs. Importantly, none of these methods reviewed here that use a WGA-free approach for TE detection can estimate the intra-sample TE allele frequency (TAF). This feature is available in multiple short-read-based TE identification tools and has enabled the detection of copy-neutral loss of heterozygosity (LOH) in *Drosophila* cell line genomes where the allele frequency of single-nucleotide variants (SNVs) was not informative [88]. In this study, the intra-sample TAF estimation included in TELR enabled the identification of haplotype-specific TE inser-

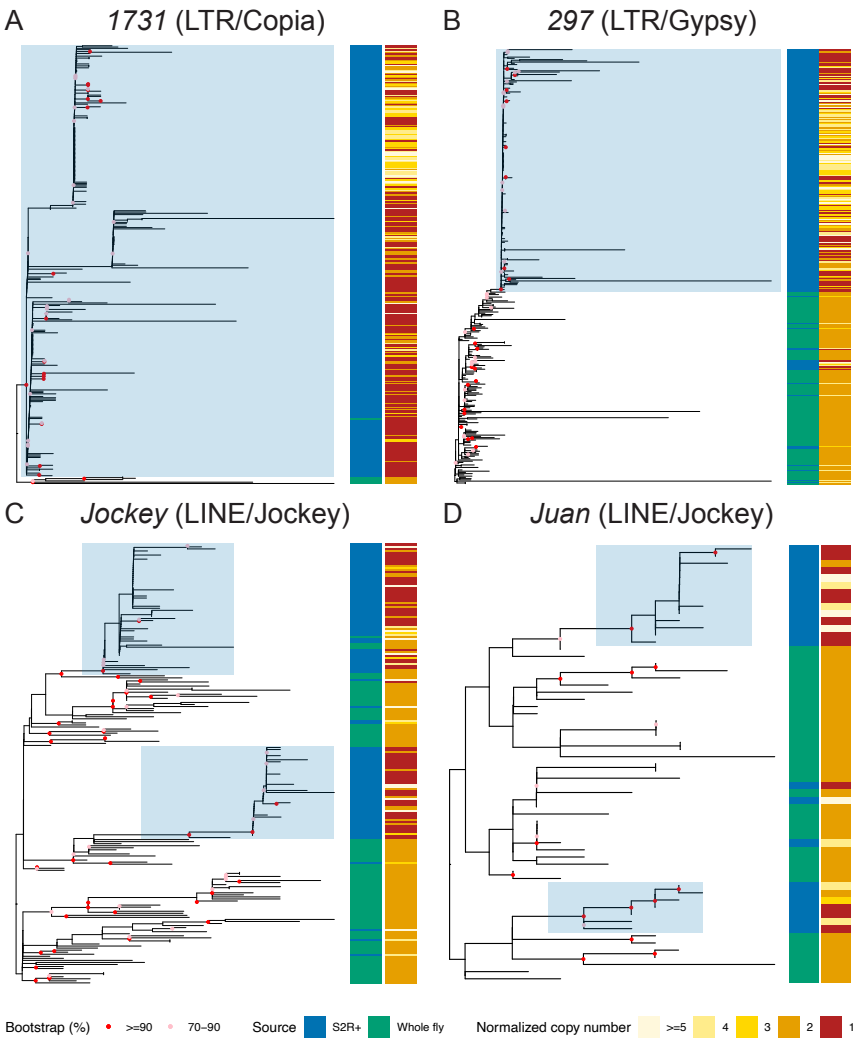

Figure 4.18: **Single and multiple TE source lineage activation in S2R+ cell line.** Non-reference TE insertion sequences from S2R+ and 11 inbred *Drosophila* fly strains were predicted and assembled by TELR. Full length TE sequences for each family were aligned using MAFFT (v7.487) [108]. The multiple sequence alignments were used as input in IQ-TREE (v2.1.4-beta) [165] to build phylogeny for *1731* (A), *297* (B), *Jockey* (C) and *Juan* (D) elements using maximum likelihood approach. All trees were midpoint rooted. The sample source and intra-sample TE allele frequency were annotated in the side bar. Clades with blue shading indicate post-tetraploid TE expansion event in S2R+ based on following criteria: 1) The bootstrap support for the clade should be $\geq$70% and 2) The proportion of sequences from cell line with less than 0.4 TAF should be $\geq$30%.

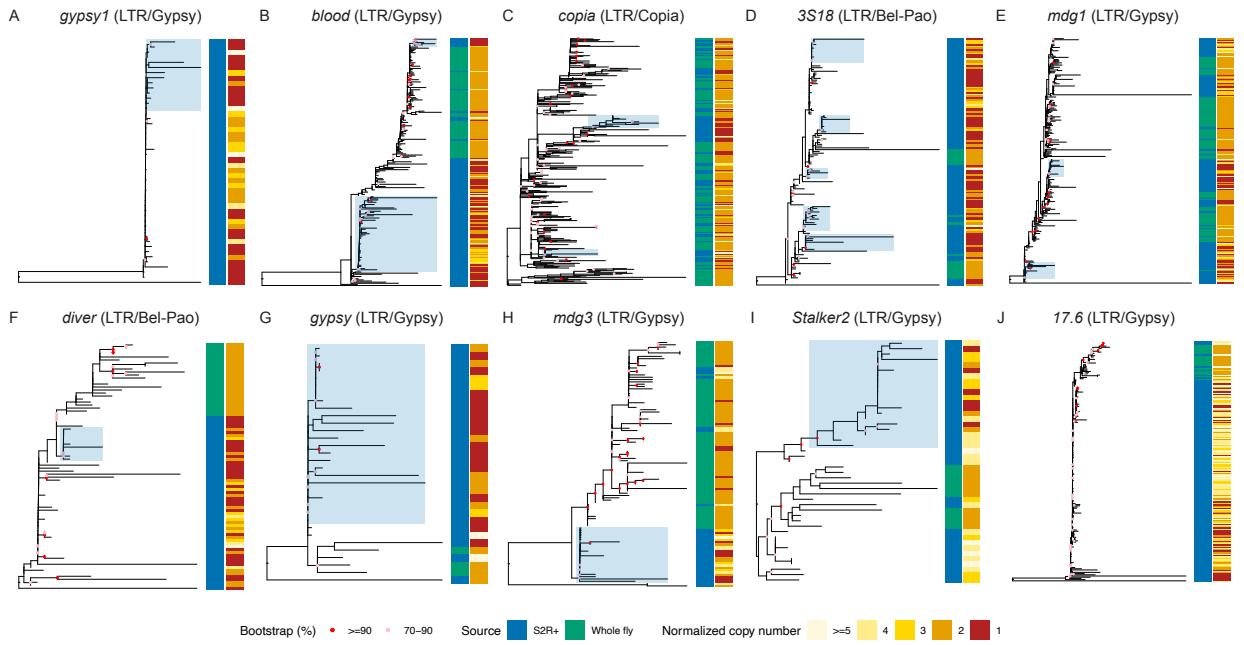

Figure 4.19: **Single and multiple TE source lineage activation in S2R+ cell line.** Non-reference TE insertion sequences from S2R+ and 11 inbred *Drosophila* fly strains were predicted and assembled by TELR. Full length TE sequences for each family were aligned using MAFFT (v7.487) [108]. The multiple sequence alignments were used as input in IQ-TREE (v2.1.4-beta) [165] to build phylogeny for *gypsy1* (A), *blood* (B), *copia* (C), *3S18* (D), *mdg1* (E), *diver* (F), *gypsy* (G), *mdg3* (H), *Stalker2* (I) and *17.6* (J) (A) elements using maximum likelihood approach. All trees were midpoint rooted. The sample source and intra-sample TE allele frequency were annotated in the side bar. Clades with blue shading indicate post-tetraploid TE expansion event in S2R+ based on following criteria: 1) The bootstrap support for the clade should be $\geq$70% and 2) The proportion of sequences from cell line with less than 0.4 TAF should be $\geq$30%.

tions in *Drosophila* cell culture and revealed somatic transposition activities in a polyploid cell line genome.

Previously, sequence level resolution of TE insertions could only be obtained from computationally expensive WGA of long reads or labor-intensive targeted sequencing experiments. Recent development of reference-based TE prediction approaches using long-read data enables reconstruction of inserted TE sequences. We compared currently available long

Table 4.6: **Feature comparison between long-read non-reference TE detection methods.**

	TELR	LoRTE	PALMER	TLDR	xTea	rMETL
PubMed ID	-	28405230	31853540	33186547	34158502	30759188
Species-agnostic	Yes	Yes	Yes	Yes	No	Yes
Predicts TSD	Yes	No	Yes	Yes	Yes	No
TE sequence	Polished local assembly	Representative raw read	Representative raw read	Consensus sequence	Unpolished local assembly	Yes
Estimates TAF	Yes	No	No	No	No	No
Predicts genotype	Yes	Polymorphic TE *	No	No	No	Yes
Available in Bioconda	-	No	No	No	Yes	Yes

* LoRTE does not predict genotypes but may flag TEs as "Possible polymorphism" if there is conflicting evidence regarding the presence/absence of a given insertion. This indicates the insertion is heterozygous or potentially a polymorphism if multiple individuals were pooled together for sequencing [68].

read TE detection methods and found that each tool may output one of three types of sequences: (1) a representative raw read supporting the TE insertion (LoRTE, PALMER), (2) a consensus sequence of TE-supporting reads (TLDR), or (3) a TE sequence resulting from local-assembly of TE-supporting reads (xTea). To our knowledge, TELR is the only tool that outputs a polished assembly of the TE locus, and provides both the TE sequence as well as a large contig that includes TE as well as its flanking sequences. The polishing step in TELR is especially important to improve sequence quality when using long-read assemblers such as wtdbg2 [203] that do not perform read error correction prior to the assembly step. High-quality sequences of predicted TE insertions are crucial for understanding the genetic variations of TEs within a sample or among populations.

Using the TELR system, we found a significantly higher number of non-reference TEs in S2R+, a sub-line of *Drosophila* S2 cell line, compared to the whole fly of a highly inbred strain A4. The increased TE copy number in S2R+ relative to A4 is mainly contributed by a subset of LTR retrotransposons (except for *Jockey*, which is LINE). Notably, most TE families enriched in S2R+ are also suggested by Chapter 2 as families with high ongoing transposition activities in S2 cell culture, providing further support that a small subset of LTR retrotransposon families contribute to ongoing insertion activity and TE amplification

in S2 cultured cells. In addition, TELR predicted that a significant proportion of the non-reference TE insertions identified in S2R+ have TAF enriched at ∼0.25, which we interpreted as somatic insertions occurred after S2 cells became tetraploid following the initial establishment of the S2 cell line. This interpretation is in line with the conclusion from Chapter 2 that TE insertions are ongoing in *Drosophila* cell culture and that a tetraploidization event occurred in the early stage of S2 following cell line establishment. If the TE amplification in cell culture is due to an initial burst with relative stasis thereafter, we expect very few TEs to be post-tetraploid. The conclusion from this work that many non-reference TEs are caused by post-tetraploid somatic transposition provides further support that TE insertion activity is ongoing in S2 cell culture. Finally, the phylogenomic analysis using TELR-assembled sequences for TE families enriched in S2R+ suggested that the TE expansion in cell culture could come from a single or multiple source lineages, providing insight into the sequence evolution of TEs in *Drosophila* cell culture.

Chapter 5

Conclusions

5.1 Summary

Cultured cells are widely used in molecular biology, so are the occurrence of cell line misidentification. However, there is a lack of protocol to authenticate cell lines of non-mammalian species. In Chapter 2, I demonstrated that transposable element (TE) insertion profiles can be used to identify long-term *Drosophila* cell lines. I developed a computational framework of *Drosophila* cell line authentication using non-reference TE insertion profiles and thereby clarified the origin of three *Drosophila* cell lines: Sg4, mbn2, and OSS_E. I also provided evidence that a subset of LTR retrotransposon families is sufficient to identify common *Drosophila* cell lines, which lays the foundation for a novel experimental protocol to authenticate *Drosophila* cell lines (see Appendix). The cell line authentication framework requires establishing TE insertion profiles from WGS data. I thereby developed a novel non-reference TE detection approach using WGS data called "ngs_te_mapper2" (https://github.com/bergmanlab/ngs_te_mapper2), which can detect precise TE insertion coordinate and reliably estimate intra-sample TE allele frequency. Furthermore, using TE allele frequency data from ngs_te_mapper2, I was able to identify the occurrence of loss of heterozygosity in several *Drosophila* cell line genomes, which serves as a mechanism that shapes unique TE profiles that differentiate *Drosophila* cell lines and sub-lines of the same cell line. Overall, this work serves as the foundation for high-throughput protocols that use TE insertion to authenticate cell lines in *Drosophila* and other non-mammalian species. Also, it improves our understanding of metazoan genome evolution using long-term cell culture as a model system.

Previous studies have shown that *Drosophila* cell lines underwent TE amplification. However, it is currently unknown whether this amplification is due to an initial burst of TEs after initial cell line establishment or ongoing TE insertions in the prolonged cell culture. In Chapter 3, I surveyed TE insertions in multiple *Drosophila* S2 cell sub-lines and used a phylogenetic approach to investigate TE activity in the history of S2. This analysis revealed that TE insertions are ongoing in the *Drosophila* S2 cell culture. Also, I showed extensive copy number variations among different *Drosophila* sub-lines of the same cell line. Overall, this work suggests that the ongoing TE insertion and copy number changes in the long-term *Drosophila* cell culture could together lead to extensive genomic variations among sub-lines of the same cell line, which could impact functional studies.

Animal cell culture is susceptible to rapid genome evolution including polyploidization, segmental aneuploidy, and TE amplification. However, little is known about the dynamics and sequence evolution of TE transposition in somatic *Drosophila* cell culture. One strategy to study TE amplification and sequence evolution in *Drosophila* cell culture involves generating *de novo* whole-genome assembly (WGA) from long-read sequencing data, which did not work on the complex and heterozygous polyploid genomes like *Drosophila* cell lines. In Chapter 4, I developed a novel bioinformatics approach called "TELR" (https://github.com/bergmanlab/TELR), which can identify, assemble, and estimate intra-sample allele frequencies of non-reference TEs from long-read sequencing data. The TELR system allowed us to discover haplotype-specific TE insertions from a polyploid *Drosophila* cell line called S2R+, which can not be achieved using the current WGA or short-read-based methods. Using TE allele frequency data estimated by TELR, I infer that most non-reference TE insertions detected in S2R+ came from somatic transposition after S2 cells became tetraploid following initial cell line establishment. I also performed phylogenomic analysis using assembled TE sequences, which revealed that amplification of TEs in *Drosophila* cell lines could arise from single or multiple source lineages. Overall, this work provides a computational

framework to study polymorphic TEs in any complex polyploid genomes and improves our understanding of TE dynamics in long-term *Drosophila* cell culture.

5.2 Future work

Chapters 3 and 4 provide strong evidence that TEs are amplified in *Drosophila* cell culture due to ongoing somatic transposition of a subset of LTR retrotransposons. However, the mechanism of TE derepression in *Drosophila* cell culture is still unknown. Thus, future work is needed to understand the mechanism for deregulating different TE families in *Drosophila* cell culture after initial establishment. In addition, experiments on the TE accumulation lines could be performed to analyze the rate of transposition of different TE families in cell culture. These future works would be crucial to understand the rate of cell line genome evolution and date the divergence time between sub-lineages of the same cell line. Finally, further investigation on cell lines from other species is needed to determine if results on *Drosophila* cell culture in this work can be generalized to other animal cell culture systems.

In Chapter 4, I introduced a new computational approach called "TELR" that provides resources for studying TEs from long-read sequencing data. The TELR approach was designed and optimized for detecting non-reference TEs with unique flanking regions. Future work is needed to identify and analyze complex TE loci. For example, TEs can be nested in *Drosophila* genome especially in heterochromatic regions [106, 27], resulting in complex loci with multiple nested and fragmented TE sequences. The current software performs well on detecting non-reference TEs in unique regions of the genome. However, TELR's sensitivity and specificity could drop significantly if the TE insertions are nested, adjacent to each other, or in highly repetitive regions. We did a preliminary study by applying TELR on the yeast genome using both real and simulated datasets. The results on yeast data are underwhelming with low sensitivity, which could be explained by the large portion of nested or adjacent insertions in the yeast genome due to the localized target site preferences of TEs to integrate near tRNA genes [102, 66]. Complex TE loci detection issues could potentially

be resolved by aligning locally assembled contig to the corresponding region in the reference genome using tools such as mummer [156]. The pairwise alignment between two local contigs can then be used for precise variant detection using specialized tools such as Assemblytics [173] and SVMU [51]. In principle, this strategy may not rely on TE annotation of local contigs and should provide precise insertion information.

The ngs_te_mapper2 approach provided in Chapter 2 allows detection of non-reference TEs and estimate TE allele frequency from legacy short read data. In principle, the split-read-based algorithm used by ngs_te_mapper2 could be extended to work on long-read data. Unlike SV detection software such as Sniffles that TELR uses for detecting insertion candidates, ngs_te_mapper2 does not rely on abnormal coverage or multiple reads to find TE insertion signal. Thus using ngs_te_mapper2-like approach could be more sensitive compared with Sniffles-like approaches in terms of detecting somatic transpositions with low allele frequency [1].

The current TELR workflow requires a high-quality TE library that includes consensus sequences of known TEs to represent diverse TE sequences in the target species. Applying TELR on species without an established high-quality TE library would very likely result in poor predictions. Future development on the TELR system is needed to enable *de novo* predictions of non-reference TEs that do not require a high-quality TE library. One possible strategy for the *de novo* TE prediction workflow is: 1) build profile hidden Markov models (HMMs) from TE-encoded protein sequences using databases such as PFAM [22]; 2) predict TE ORFs using assembled TE contigs from the current TELR workflow; 3) run profile HMMs against the predicted ORFs from the assembled contigs; and 4) analyze TELR-assembled contigs that hit profile HMMs, align assembled contig sequences to the reference genome to identify precise insertion coordinate and candidate TE sequences. Another possible strategy is to develop a machine-learning- or deep-learning-based model that could predict TE sequences directly from local contigs assembled by TELR. Both strategies could be

evaluated or benchmarked by generating a gold standard dataset from *Drosophila* genome, which has a high-quality curated TE library.

The current TELR workflow could also be extended to detect reference TEs. The ability to predict presence or absence of reference TE, assemble TE sequence, and estimate TAF are crucial to study the genetic diversity of TEs among *Drosophila* populations [195]. The reference TE detection workflow could involve 1) aligning reads to the reference genome and investigate read coverage around each reference TE locus; 2) assembling reference-TE-supporting reads around each TE candidate locus; and 3) estimating TAF based on coverages between TE region and regions that flanks both side of reference TEs.

Compared to short-read technologies like Illumina NGS, long-read technologies such as Pacbio and Nanopore significantly improve the quality and completeness of genome assemblies for many organisms, giving researchers access to regions in the genome such as TEs that are traditionally difficult to resolve using short-read data. The long-read technologies unlock a new era of TE genomics research by providing new data and strategies to answer key questions in TE biology. First, the ability to reconstruct full-length TE copy would enable researchers to analyze not only overall TE abundance or presence/absence of TEs but also TE sequences, which can be used for TE comparative genomics study to understand genetic variation and evolution of TEs within a single genome, in genomes across populations, or genomes across species. These sequences can also be used to date TE transpositions based on LTR sequence divergence. Second, it is well known that one TE family can be present with different structures. Also, TEs such as LINEs are often present in partial forms. Reconstructing the TE locus allows researchers to study TE subfamilies in a single family with potentially different ORF structures and SVs. Third, long-read data provide new strategies to discover novel TEs. Finally, complete TE sequence reconstruction allows new insight on the function of TEs in the host genome and how they are being regulated.

This dissertation provides examples of using long-read to study TE dynamics and sequence evolution in cell culture genome, offering new insights into how TEs can shape

cell line genome evolution. We are still in the early days of utilizing long-read technologies to study TE biology. With these technologies becoming more affordable, together with improvement in base-calling accuracy and development of computational tools for analyzing long-read data, I expect more studies that improve our understandings of TE biology coming in the near future.

Appendix

A novel transposable element based authentication protocol for Drosophila cell lines

Following the proof of principle analysis on *Drosophila* cell line authentication in Chapter 2, we collaborated with Drosophila Genomics Resource Center (DGRC) to develop an experimental protocol to identify commonly used *Drosophila* cell lines. The details of the authentication protocol and results were described in Mariyappa *et al.* 2021 [155]. I contributed to this work by providing primer design and a bioinformatics workflow called "enrichmenTE" (https://github.com/bergmanlab/enrichmenTE) that analyzes LTR retrotransposon insertions in PCR-enriched NGS samples.

Bibliography

[1] Unraveling the features of somatic transposition in the Drosophila intestine. *The EMBO Journal*, 40(9):e106388, May 2021.

[2] Andrew Adey, Joshua N. Burton, Jacob O. Kitzman, Joseph B. Hiatt, Alexandra P. Lewis, Beth K. Martin, Ruolan Qiu, Choli Lee, and Jay Shendure. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*, 500(7461):207–211, August 2013.

[3] Jeffrey R. Adrion, Michael J. Song, Daniel R. Schrider, Matthew W. Hahn, and Sarah Schaack. Genome-wide estimates of transposable element insertion and deletion rates in Drosophila melanogaster. *Genome Biol Evol*, 9(5):1329–1340, May 2017.

[4] Can Alkan, Bradley P. Coe, and Evan E. Eichler. Genome structural variation discovery and genotyping. *Nat Rev Genet*, 12(5):363–376, May 2011.

[5] Can Alkan, Saba Sajjadian, and Evan E. Eichler. Limitations of next-generation genome sequence assembly. *Nat Methods*, 8(1):61–65, January 2011.

[6] Jamie L. Almeida, Kenneth D. Cole, and Anne L. Plant. Standards for cell line authentication and beyond. *PLOS Biology*, 14(6):e1002476, June 2016.

[7] American Type Culture Collection Standards Development Organization Workgroup ASN-0002. Cell line misidentification: the beginning of the end. *Nat Rev Cancer*, 10(6):441–448, June 2010.

[8] D Anxolabehere, L Charles-Palabost, A Fleuriet, and G Periquet. Temporal surveys of French populations of Drosophila melanogaster: P-M system, enzymatic polymorphism and infection by the sigma virus. *Heredity*, 61(1):121–131, August 1988.

[9] I.R. Arkhipova, N.V. Lyubomirskaya, and Y.V. Ilyin. *Drosophila Retrotransposons*. R.G. Landes Co., Austin, TX, 1995.

[10] Michael Ashburner and Casey M. Bergman. Drosophila melanogaster: A case study of a model genomic sequence and its consequences. *Genome Res*, 15(12):1661–1667, December 2005.

[11] Philip M. Ashton, Satheesh Nair, Tim Dallman, Salvatore Rubino, Wolfgang Rabsch, Solomon Mwaigwisya, John Wain, and Justin O'Grady. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat Biotechnol*, 33(3):296–300, March 2015.

[12] Zeljana Babic, Amanda Capes-Davis, Maryann E Martone, Amos Bairoch, I Burak Ozyurt, Thomas H Gillespie, and Anita E Bandrowski. Incidences of problematic cell lines are lower in papers that use RRIDs to identify cell lines. *eLife*, 8:e41676, January 2019.

[13] Michael W. Bairu, Adeyemi O. Aremu, and Johannes Van Staden. Somaclonal variation in plants: causes and detection methods. *Plant Growth Regul*, 63(2):147–173, March 2011.

[14] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5):455–477, May 2012.

[15] Weidong Bao, Kenji K. Kojima, and Oleksiy Kohany. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*, 6(1):11, June 2015.

[16] Zhirong Bao and S R Eddy. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res*, 12(8):1269–76, August 2002.

[17] Rita Barallon, Steven R. Bauer, John Butler, Amanda Capes-Davis, Wilhelm G. Dirks, Eugene Elmore, Manohar Furtado, Margaret C. Kline, Arihiro Kohara, Georgyi V. Los, Roderick A. F. MacLeod, John R. W. Masters, Mark Nardone, Roland M. Nardone, Raymond W. Nims, Paul J. Price, Yvonne A. Reid, Jaiprakash Shewale, Gregory Sykes, Anton F. Steuer, Douglas R. Storts, Jim Thomson, Zenobia Taraporewala, Christine Alston-Roberts, and Liz Kerrigan. Recommendation of short tandem repeat profiling for authenticating human cell lines, stem cells, and tissues. *In Vitro Cell Dev Biol Anim*, 46(9):727–732, October 2010.

[18] Bridlin Barckmann, Marianne El-Barouk, Alain Pélisson, Bruno Mugat, Blaise Li, Céline Franckhauser, Anna-Sophie Fiston Lavier, Marie Mirouze, Marie Fablet, and Séverine Chambeyron. The somatic piRNA pathway controls germline transposition over generations. *Nucleic Acids Res*, 46(18):9524–9536, October 2018.

[19] Maite G. Barron, Anna-Sophie Fiston-Lavier, Dmitri A. Petrov, and Josefa Gonzalez. Population genomics of transposable elements in Drosophila. *Annu Rev Genet*, 48:561–581, 2014.

[20] Carolina Bartolome, Xulio Maside, and Brian Charlesworth. On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster. *Mol Biol Evol*, 19(6):926–937, June 2002.

[21] Andrew R. Bassett, Charlotte Tibbit, Chris P. Ponting, and Ji-Long Liu. Mutagenesis and homologous recombination in Drosophila cell lines using CRISPR/Cas9. *Biology Open*, 3(1):42–49, January 2014.

[22] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer. The Pfam protein families database. *Nucleic Acids Res*, 30(1):276–80, 2002. 1362-4962 (Electronic) Journal Article.

[23] Mark A. Batzer and Prescott L. Deininger. Alu repeats and human genomic diversity. *Nature Reviews Genetics*, 3(5):370–379, May 2002.

[24] Uri Ben-David, Benjamin Siranosian, Gavin Ha, Helen Tang, Yaara Oren, Kunihiko Hinohara, Craig A. Strathdee, Joshua Dempster, Nicholas J. Lyons, Robert Burns, Anwesha Nag, Guillaume Kugener, Beth Cimini, Peter Tsvetkov, Yosef E. Maruvka, Ryan O'Rourke, Anthony Garrity, Andrew A. Tubelli, Pratiti Bandopadhayay, Aviad Tsherniak, Francisca Vazquez, Bang Wong, Chet Birger, Mahmoud Ghandi, Aaron R. Thorner, Joshua A. Bittker, Matthew Meyerson, Gad Getz, Rameen Beroukhim, and Todd R. Golub. Genetic and transcriptional evolution alters cancer cell line drug response. *Nature*, 560(7718):325–330, August 2018.

[25] Farid Benachenhou, Goran O Sperber, Erik Bongcam-Rudloff, Goran Andersson, Jef D Boeke, and Jonas Blomberg. Conserved structure and inferred evolutionary history of long terminal repeats (LTRs). *Mob DNA*, 4(1):5, 2013.

[26] Casey M. Bergman. A proposal for the reference-based annotation of de novo transposable element insertions. *Mob Genet Elements*, 2(1):51–54, January 2012.

[27] Casey M. Bergman and Douda Bensasson. Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in Drosophila melanogaster. *Proc Natl Acad Sci USA*, 104(27):11340–11345, July 2007.

[28] Casey M. Bergman, Shunhua Han, Michael G. Nelson, Vladyslav Bondarenko, and Iryna Kozeretska. Genomic analysis of P elements in natural populations of Drosophila melanogaster. *PeerJ*, 5:e3824, September 2017.

[29] Casey M. Bergman, Hadi Quesneville, Dominique Anxolabehere, and Michael Ashburner. Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. *Genome Biol*, 7(11):R112, January 2006.

[30] Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James P. Drake, Jane M. Landolin, and Adam M. Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotech*, 33(6):623–630, June 2015.

[31] C. Biemont, L. Monti-Dedieu, and F. Lemeunier. Detection of transposable elements in Drosophila salivary gland polytene chromosomes by in situ hybridization. *Methods Mol Biol*, 260:21–8, 2004. 1064-3745 Journal Article.

[32] C. Biemont, S. Ronsseray, D. Anxolabehere, H. Izaabel, and C. Gautier. Localization of P elements, copy number regulation, and cytotype determination in Drosophila melanogaster. *Genetics Research*, 56(1):3–14, August 1990.

[33] C. Biémont and G. Cizeron. Distribution of transposable elements in Drosophila species. *Genetica*, 105(1):43–62, January 1999.

[34] D M Black, M S Jackson, M G Kidwell, and G A Dover. KP elements repress P-induced hybrid dysgenesis in Drosophila melanogaster. *EMBO J*, 6(13):4125–4135, December 1987.

[35] Justin P. Blumenstiel, Xi Chen, Miaomiao He, and Casey M. Bergman. An age-of-allele test of neutrality for transposable element insertions. *Genetics*, 196(2):523–538, February 2014.

[36] J. D. Boeke, D. J. Garfinkel, C. A. Styles, and G. R. Fink. Ty elements transpose through an RNA intermediate. *Cell*, 40(3):491–500, March 1985.

[37] Valentina Boeva, Tatiana Popova, Kevin Bleakley, Pierre Chiche, Julie Cappo, Gudrun Schleiermacher, Isabelle Janoueix-Lerosey, Olivier Delattre, and Emmanuel Barillot.

Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, 28(3):423–425, February 2012.

[38] N. J. Bowen and J. F. McDonald. Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside. *Genome Res*, 11(9):1527–40, September 2001.

[39] Ryan Bracewell, Kamalakar Chatla, Matthew J Nalley, and Doris Bachtrog. Dynamic turnover of centromeres drives karyotype evolution in Drosophila. *eLife*, 8:e49002, September 2019.

[40] Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, May 2016.

[41] Julius Brennecke, Alexei A. Aravin, Alexander Stark, Monica Dus, Manolis Kellis, Ravi Sachidanandam, and Gregory J. Hannon. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell*, 128(6):1089–103, March 2007.

[42] James B. Brown, Nathan Boley, Robert Eisman, Gemma E. May, Marcus H. Stoiber, Michael O. Duff, Ben W. Booth, Jiayu Wen, Soo Park, Ana Maria Suzuki, Kenneth H. Wan, Charles Yu, Dayu Zhang, Joseph W. Carlson, Lucy Cherbas, Brian D. Eads, David Miller, Keithanne Mockaitis, Johnny Roberts, Carrie A. Davis, Erwin Frise, Ann S. Hammonds, Sara Olson, Sol Shenker, David Sturgill, Anastasia A. Samsonova, Richard Weiszmann, Garret Robinson, Juan Hernandez, Justen Andrews, Peter J. Bickel, Piero Carninci, Peter Cherbas, Thomas R. Gingeras, Roger A. Hoskins, Thomas C. Kaufman, Eric C. Lai, Brian Oliver, Norbert Perrimon, Brenton R. Graveley, and Susan E. Celniker. Diversity and dynamics of the Drosophila transcriptome. *Nature*, 512(7515):393–399, August 2014.

[43] Brian Bushnell. BBMap: a fast, accurate, splice-aware aligner. Technical Report LBNL-7065E, Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), March 2014.

[44] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421, 2009.

[45] Judith Campisi and Fabrizio d'Adda di Fagagna. Cellular senescence: when bad things happen to good cells. *Nat Rev Mol Cell Biol*, 8(9):729–740, September 2007.

[46] Amanda Capes-Davis, George Theodosopoulos, Isobel Atkin, Hans G. Drexler, Arihiro Kohara, Roderick A. F. MacLeod, John R. Masters, Yukio Nakamura, Yvonne A. Reid, Roger R. Reddel, and R. Ian Freshney. Check your cultures! A list of cross-contaminated or misidentified cell lines. *Int J Cancer*, 127(1):1–8, July 2010.

[47] Felipe Castro, Wilhelm G. Dirks, Silke Fähnrich, Agnes Hotz-Wagenblatt, Michael Pawlita, and Markus Schmitt. High-throughput SNP-based authentication of human cell lines. *Int J Cancer*, 132(2):308–314, January 2013.

[48] Susan E Celniker, Laura A L Dillon, Mark B Gerstein, Kristin C Gunsalus, Steven Henikoff, Gary H Karpen, Manolis Kellis, Eric C Lai, Jason D Lieb, David M MacAlpine, Gos Micklem, Fabio Piano, Michael P Snyder, Lincoln Stein, Kevin P White, and Robert H Waterston. Unlocking the secrets of the genome. *Nature*, 459(7249):927–30, June 2009.

[49] Mark J. Chaisson and Glenn Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, 13(1):238, September 2012.

[50] Mahul Chakraborty, J. J. Emerson, Stuart J. Macdonald, and Anthony D. Long. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat Commun*, 10(1):4872, October 2019.

[51] Mahul Chakraborty, Nicholas W. VanKuren, Roy Zhao, Xinwen Zhang, Shannon Kalsow, and J. J. Emerson. Hidden genetic variation shapes the structure of functional elements in Drosophila. *Nat Genet*, 50(1):20–25, January 2018.

[52] B. Charlesworth and C. H. Langley. The population genetics of Drosophila transposable elements. *Annu Rev Genet*, 23:251–87, January 1989.

[53] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890, September 2018.

[54] Lucy Cherbas and Lei Gong. Cell Lines. *Methods*, 68(1):74–81, June 2014.

[55] Lucy Cherbas, Robert Moss, and Peter Cherbas. Chapter 9 Transformation Techniques for Drosophila Cell Lines. In Lawrence S. B. Goldstein and Eric A. Fyrberg, editors, *Methods in Cell Biology*, volume 44, pages 161–179. Academic Press, January 1994.

[56] Lucy Cherbas, Aarron Willingham, Dayu Zhang, Li Yang, Yi Zou, Brian D. Eads, Joseph W. Carlson, Jane M. Landolin, Philipp Kapranov, Jacqueline Dumais, Anastasia Samsonova, Jeong-Hyeon H. Choi, Johnny Roberts, Carrie A. Davis, Haixu Tang, Marijke J. van Baren, Srinka Ghosh, Alexander Dobin, Kim Bell, Wei Lin, Laura Langton, Michael O. Duff, Aaron E. Tenney, Chris Zaleski, Michael R. Brent, Roger A. Hoskins, Thomas C. Kaufman, Justen Andrews, Brenton R. Graveley, Norbert Perrimon, Susan E. Celniker, Thomas R. Gingeras, and Peter Cherbas. The transcriptional diversity of 25 Drosophila cell lines. *Genome Res*, 21(2):301–314, February 2011.

[57] Chen-Shan Chin, Paul Peluso, Fritz J. Sedlazeck, Maria Nattestad, Gregory T. Concepcion, Alicia Clum, Christopher Dunn, Ronan O'Malley, Rosa Figueroa-Balderas, Abraham Morales-Cruz, Grant R. Cramer, Massimo Delledonne, Chongyuan Luo,

Joseph R. Ecker, Dario Cantu, David R. Rank, and Michael C. Schatz. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods*, 13(12):1050–1054, December 2016.

[58] Chong Chu, Rebeca Borges-Monroy, Vinayak V. Viswanadham, Soohyun Lee, Heng Li, Eunjung Alice Lee, and Peter J. Park. Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat Commun*, 12(1):3836, June 2021. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Subject_term: Cancer genomics;Genome informatics;Genomic instability;Machine learning;Software Subject_term_id: cancer-genomics;genome-informatics;genomic-instability;machine-learning;software.

[59] Nancy L. Craig. *Mobile DNA II*. ASM Press, Washington, D.C., 2002. 2001045975 edited by Nancy L. Craig ... [et al.] Mobile DNA 2. Mobile DNA two. ill. (some col.) ; 29 cm. Includes bibliographical references and index.

[60] Julie M. Cridland, Stuart J. Macdonald, Anthony D. Long, and Kevin R. Thornton. Abundance and distribution of transposable elements in two Drosophila QTL mapping resources. *Mol Biol Evol*, 30(10):2311–2327, October 2013.

[61] D. A. Currie, M. J. Milner, and C. W. Evans. The growth and differentiation in vitro of leg and wing imaginal disc cells from Drosophila melanogaster. *Development*, 102(4):805–814, April 1988.

[62] Vittorio Defendi, R. E. Billingham, Willys K. Silvers, and Paul Moorhead. Immunological and karyological criteria for identification of cell lines. *JNCI: Journal of the National Cancer Institute*, 25(2):359–385, August 1960.

[63] D. DeFranco, O. Schmidt, and D. Soll. Two control regions for eukaryotic tRNA gene transcription. *Proc Natl Acad Sci USA*, 77(6):3365–8, June 1980.

[64] P. Deininger and A. Roy-Engel. Mobile Elements in Animals and Plants. In N. Craig, editor, *Mobile DNA II*, pages 1074–1092. ASM Press, Washington, D.C., 2002.

[65] Thomas Derrien, Jordi Estelle, Santiago Marco Sola, David G. Knowles, Emanuele Raineri, Roderic Guigo, and Paolo Ribeca. Fast Computation and Applications of Genome Mappability. *PLOS ONE*, 7(1):e30377, January 2012.

[66] S. E. Devine and J. D. Boeke. Integration of the yeast retrotransposon Ty1 is targeted to regions upstream of genes transcribed by RNA polymerase III. *Genes Dev*, 10(5):620–633, March 1996.

[67] C. Di Franco, C. Pisano, F. Fourcade-Peronnet, G. Echalier, and N. Junakovic. Evidence for de novo rearrangements of Drosophila transposable elements induced by the passage to the cell culture. *Genetica*, 87(2):65–73, 1992.

[68] Eric Disdero and Jonathan Filee. LoRTE: Detecting transposon-induced genomic variants using low coverage PacBio long read sequences. *Mob DNA*, 8:5, April 2017.

[69] G. Echalier and A. Ohanessian. Isolement, en cultures in vitro, de lignees cellulaires diploides de Drosophila melanogaster. *Comptes rendus hebdomadaires des seances de l'Academie des Sciences*, 268:1771–1773, 1969.

[70] Guy Echalier. *Drosophila Cells in Culture*. Academic Press, San Diego, Calif., February 1997.

[71] Christophe J. Echeverri and Norbert Perrimon. High-throughput RNAi screening in cultured cells: a user's guide. *Nat Rev Genet*, 7(5):373–384, May 2006.

[72] Adam C. English, William J. Salerno, and Jeffrey G. Reid. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics*, 15(1):180, June 2014.

[73] Adam D. Ewing. Transposable element detection from whole genome sequence data. *Mob DNA*, 6:24, 2015.

[74] Adam D. Ewing, Nathan Smits, Francisco J. Sanchez-Luque, Jamila Faivre, Paul M. Brennan, Sandra R. Richardson, Seth W. Cheetham, and Geoffrey J. Faulkner. Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. *Molecular Cell*, November 2020.

[75] James S. Farris. Phylogenetic Analysis Under Dollo's Law. *Systematic Biology*, 26(1):77–88, March 1977.

[76] Anna-Sophie Fiston-Lavier, Maite G. Barron, Dmitri A. Petrov, and Josefa Gonzalez. T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. *Nucleic Acids Res*, 43(4):e22–e22, February 2015.

[77] Anna-Sophie Fiston-Lavier, Matthew Carrigan, Dmitri A. Petrov, and Josefa González. T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. *Nucleic Acids Res*, 39(6):e36, March 2011.

[78] Jullien M. Flynn, Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, and Arian F. Smit. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA*, 117(17):9451–9457, April 2020.

[79] D. K. Ford and G. Yerganian. Observations on the chromosomes of Chinese hamster cells in tissue culture. *J Natl Cancer Inst*, 21(2):393–425, August 1958.

[80] S. M. Gartler. Genetic markers as tracers in cell culture. *Natl Cancer Inst Monogr*, 26:167–195, September 1967.

[81] E. Gateff. [New mutants report.]. *Drosophila Information Service*, 52:4–5, 1977.

[82] E. Gateff, L. Gissmann, R. Shrestha, N. Plus, H. Pfister, J. Schroder, and H.Z. Hausen. Characterization of two tumorous blood cell lines of Drosophila melanogaster and the viruses they contain. *Invertebrate Systems in Vitro Fifth International Conference on Invertebrate Tissue Culture, Rigi-Kaltbad, Switzerland, 1979*, pages 517–533, 1980.

[83] D. A. Gilbert, Y. A. Reid, M. H. Gail, D. Pee, C. White, R. J. Hay, and S. J. O'Brien. Application of DNA fingerprints for cell-line individualization. *Am J Hum Genet*, 47(3):499–514, September 1990.

[84] Josefa Gonzalez, Kapa Lenkov, Mikhail Lipatov, J. Michael Macpherson, and Dmitri A. Petrov. High rate of recent transposable element-induced adaptation in Drosophila melanogaster. *Plos Biol*, 6(10):e251, October 2008.

[85] Clément Goubert, Laurent Modolo, Cristina Vieira, Claire ValienteMoro, Patrick Mavingui, and Matthieu Boulesteix. De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). *Genome Biol Evol*, 7(4):1192–1205, March 2015.

[86] G. Gremme, S. Steinbiss, and S. Kurtz. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(03):645–656, June 2013.

[87] Rui Guo, Yan-Ran Li, Shan He, Le Ou-Yang, Yiwen Sun, and Zexuan Zhu. RepLong: de novo repeat identification using long read sequencing data. *Bioinformatics*, 34(7):1099–1107, April 2018.

[88] Shunhua Han, Preston J Basting, Guilherme B Dias, Arthur Luhur, Andrew C Zelhof, and Casey M Bergman. Transposable element profiles reveal cell line identity and loss of heterozygosity in Drosophila cell culture. *Genetics*, (iyab113):(in press), July 2021.

[89] M Hattori, S Kuhara, O Takenaka, and Y Sakaki. L1 family of repetitive DNA sequences in primates may be derived from a sequence encoding a reverse transcriptase-related protein. *Nature*, 321(6070):625–8, 1986.

[90] Ericka R Havecker, Xiang Gao, and Daniel F Voytas. The diversity of LTR retrotransposons. *Genome biology*, 5(6):225, January 2004.

[91] L. Hayflick and P. S. Moorhead. The serial cultivation of human diploid cell strains. *Exp Cell Res*, 25:585–621, December 1961.

[92] David Heller and Martin Vingron. SVIM: structural variant identification using mapped long reads. *Bioinformatics*, 35(17):2907–2915, September 2019.

[93] W.F. Hink. A compilation of invertebrate cell lines and culture media. In Karl Maramorosch, editor, *Invertebrate Tissue Culture*, pages 319–369. Academic Press, January 1976.

[94] Serge P. J. M. Horbach and Willem Halffman. The ghosts of HeLa: How cell line misidentification contaminates the scientific literature. *PLOS ONE*, 12(10):e0186281, October 2017.

[95] Roger A. Hoskins, Joseph W. Carlson, Kenneth H. Wan, Soo Park, Ivonne Mendez, Samuel E. Galle, Benjamin W. Booth, Barret D. Pfeiffer, Reed A. George, Robert Svirskas, Martin Krzywinski, Jacqueline Schein, Maria Carmela Accardo, Elisabetta Damia, Giovanni Messina, María Méndez-Lago, Beatriz de Pablos, Olga V. Demakova, Evgeniya N. Andreyeva, Lidiya V. Boldyreva, Marco Marra, A. Bernardo Carvalho, Patrizio Dimitri, Alfredo Villasante, Igor F. Zhimulev, Gerald M. Rubin, Gary H. Karpen, and Susan E. Celniker. The Release 6 reference sequence of the Drosophila melanogaster genome. *Genome Res*, 25(3):445–458, March 2015.

[96] Benjamin E. Housden, Hilary E. Nicholson, and Norbert Perrimon. Synthetic lethality screens using RNAi in combination with CRISPR-based knockout in Drosophila cells. *Bio Protoc*, 7(3), February 2017.

[97] Yaqing Huang, Yuehong Liu, Congyi Zheng, and Chao Shen. Investigation of cross-contamination and misidentification of 278 widely used tumor cell lines. *PLOS One*, 12(1):e0170384, 2017.

[98] Peyton Hughes, Damian Marshall, Yvonne Reid, Helen Parkes, and Cohava Gelber. The costs of using unauthenticated, over-passaged cell lines: how much more data do we need? *BioTechniques*, 43(5):575–584, November 2007.

[99] Thomas Hummel and Christian Klämbt. P-element mutagenesis. *Methods Mol Biol*, 420:97–117, January 2008.

[100] Yurii V. Ilyin, Valerija G. Chmeliauskaite, Eugenii V. Ananiev, and Georgii P. Georgiev. Isolation and characterization of a new family of mobile dispersed genetic elements, mdg3, in Drosophila melanogaster. *Chromosoma*, 81(1):27–53, November 1980.

[101] Zamin Iqbal, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet*, 44(2):226–232, February 2012.

[102] H. Ji, D. P. Moore, M. A. Blomberg, L. T. Braiterman, D. F. Voytas, G. Natsoulis, and J. D. Boeke. Hotspots for unselected Ty1 transposition events on yeast chromosome III are near tRNA genes and LTR sequences. *Cell*, 73(5):1007–1018, June 1993.

[103] Tao Jiang, Bo Liu, and Yadong Wang. rMETL: sensitive and fast mobile element insertion detection with long read realignment. *bioRxiv*, page 421560, September 2018.

[104] Tao Jiang, Yongzhuang Liu, Yue Jiang, Junyi Li, Yan Gao, Zhe Cui, Yadong Liu, Bo Liu, and Yadong Wang. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol*, 21(1):189, August 2020.

[105] N. Junakovic, C. Di Franco, M. Best-Belpomme, and G. Echalier. On the transposition of copia-like nomadic elements in cultured Drosophila cells. *Chromosoma*, 97(3):212–218, November 1988.

[106] Joshua S Kaminker, Casey M Bergman, Brent Kronmiller, Joseph Carlson, Robert Svirskas, Sandeep Patel, Erwin Frise, David A Wheeler, Suzanna E Lewis, Gerald M Rubin, Michael Ashburner, and Susan E Celniker. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol*, 3(12):research0084, 2002.

[107] V. V. Kapitonov and J. Jurka. Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci USA*, 98(15):8714–9, 2001. 0027-8424 Journal Article.

[108] Kazutaka Katoh and Daron M. Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, 30(4):772–780, April 2013.

[109] Gurvinder Kaur and Jannette M. Dufour. Cell lines. *Spermatogenesis*, 2(1):1–5, January 2012.

[110] H. H. Kazazian and J. V. Moran. The impact of L1 retrotransposons on the human genome. *Nat Genet*, 19(1):19–24, May 1998.

[111] Haig H. Kazazian. Mobile Elements: Drivers of Genome Evolution. *Science*, 303(5664):1626–1632, March 2004.

[112] Thomas M. Keane, Kim Wong, and David J. Adams. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics*, 29(3):389–390, February 2013.

[113] M.G. Kidwell, T. Frydryk, and J.B. Novy. The hybrid dysgenesis potential of Drosophila melanogaster strains of diverse temporal and geographical natural origins. *Drosophila Information Service*, 59:63–69, 1983.

[114] Elizabeth G. King, Chris M. Merkes, Casey L. McNeil, Steven R. Hoofer, Saunak Sen, Karl W. Broman, Anthony D. Long, and Stuart J. Macdonald. Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. *Genome Res*, 22(8):1558–1566, August 2012.

[115] Robert Kofler, Andrea J. Betancourt, and Christian Schlotterer. Sequencing of pooled DNA samples (pool-seq) uncovers complex dynamics of transposable element insertions in Drosophila melanogaster. *PLOS Genet*, 8(1):e1002487, January 2012.

[116] Robert Kofler, Daniel Gomez-Sanchez, and Christian Schlotterer. PoPoolationTE2: comparative population genomics of transposable elements using pool-seq. *Mol Biol Evol*, 33(10):2759–2764, October 2016.

[117] Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5):540–546, May 2019.

[118] Sergey Koren, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*, 27(5):722–736, 2017.

[119] Shunichi Kosugi, Yukihide Momozawa, Xiaoxi Liu, Chikashi Terao, Michiaki Kubo, and Yoichiro Kamatani. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol*, 20(1):117, June 2019.

[120] Jesse H. Krijthe. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation, 2015.

[121] Evgenia V. Kriventseva, Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Renata Dias, Felipe A. Simão, and Evgeny M. Zdobnov. OrthoDB v10: sampling the diversity

of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res*, 47(D1):D807–D811, January 2019.

[122] Fritjof Lammers, Moritz Blumer, Cornelia Ruckle, and Maria A. Nilsson. Retrophylogenomics in rorquals indicate large ancestral population sizes and a rapid radiation. *Mob DNA*, 10(1):5, January 2019.

[123] Fritjof Lammers, Susanne Gallus, Axel Janke, and Maria A. Nilsson. Phylogenetic conflict in bears identified by automated discovery of transposable element insertions in low-coverage genomes. *Genome Biol Evol*, 9(10):2862–2878, October 2017.

[124] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L.

Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[125] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, March 2009.

[126] Aoife Larkin, Steven J Marygold, Giulia Antonazzo, Helen Attrill, Gilberto dos Santos, Phani V Garapati, Joshua L Goodman, L Sian Gramates, Gillian Millburn, Victor B Strelets, Christopher J Tabone, Jim Thurmond, and FlyBase Consortium. FlyBase: updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Res*, 49(D1):D899–D907, January 2021.

[127] Nelson C. Lau, Nicolas Robine, Raquel Martin, Wei-Jen Chung, Yuzo Niki, Eugene Berezikov, and Eric C. Lai. Abundant primary piRNAs, endo-siRNAs, and microRNAs in a Drosophila ovary cell line. *Genome Res*, 19(10):1776–1785, October 2009.

[128] Pascal Leblanc, Bernard Dastugue, and Chantal Vaury. The Integration Machinery of ZAM, a Retroelement from Drosophila melanogaster, Acts as a Sequence-Specific Endonuclease. *J Virol*, 73(8):7061–7064, August 1999.

[129] Hangnoh Lee, C. Joel McManus, Dong-Yeon Cho, Matthew Eaton, Fioranna Renda, Maria Patrizia Somma, Lucy Cherbas, Gemma May, Sara Powell, Dayu Zhang, Lijun Zhan, Alissa Resch, Justen Andrews, Susan E. Celniker, Peter Cherbas, Teresa M. Przytycka, Maurizio Gatti, Brian Oliver, Brenton Graveley, and David MacAlpine. DNA copy number evolution in Drosophila cell lines. *Genome Biol*, 15(8):R70, August 2014.

[130] Hangnoh Lee and Brian Oliver. Drosophila cell lines to model selection for aneuploid states. *Journal of Down Syndrome & Chromosome Abnormalities*, 2(1):1–4, December 2015.

[131] Michael Levine. The Chromosome-Number in Cancer Tissue of Man, of Rodent, of Bird and in Crown Gall Tissue of Plants. *The Journal of Cancer Research*, 14(3):400–425, August 1930.

[132] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, November 2011.

[133] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, page 1303.3997, March 2013.

[134] Heng Li. seqtk, 2015.

[135] Heng Li. wgsim, 2015.

[136] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[137] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[138] May M. Y. Liang-Chu, Mamie Yu, Peter M. Haverty, Julie Koeman, Janet Ziegle, Marie Lee, Richard Bourgon, and Richard M. Neve. Human biosample authentication using the high-throughput, cost-effective snptrace system. *PLOS ONE*, 10(2):e0116218, February 2015.

[139] Raquel S. Linheiro and Casey M. Bergman. Whole genome resequencing reveals natural target site preferences of transposable elements in Drosophila melanogaster. *PLOS One*, 7(2):e30008, February 2012.

[140] Mikhail Lipatov, Kapa Lenkov, Dmitri A. Petrov, and Casey M. Bergman. Paucity of chimeric gene-transposable element transcripts in the Drosophila melanogaster genome. *BMC Biol*, 3:24, January 2005.

[141] Yansheng Liu, Yang Mi, Torsten Mueller, Saskia Kreibich, Evan G. Williams, Audrey Van Drogen, Christelle Borel, Max Frank, Pierre-Luc Germain, Isabell Bludau, Martin Mehnert, Michael Seifert, Mario Emmenlauer, Isabel Sorg, Fedor Bezrukov, Frederique Sloan Bena, Hu Zhou, Christoph Dehio, Giuseppe Testa, Julio Saez-Rodriguez, Stylianos E. Antonarakis, Wolf-Dietrich Hardt, and Ruedi Aebersold. Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nat Biotechnol*, 37(3):314–322, March 2019.

[142] Jon R. Lorsch, Francis S. Collins, and Jennifer Lippincott-Schwartz. Fixing problems with cell lines. *Science*, 346(6216):1452–1453, December 2014.

[143] D D Luan, M H Korman, J L Jakubczak, and T H Eickbush. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, 72(4):595–605, February 1993.

[144] Arthur Luhur, Kristin M. Klueg, and Andrew C. Zelhof. Generating and working with Drosophila cell cultures: Current challenges and opportunities. *Wiley Interdiscip Rev Dev Biol*, 8(3):e339, 2019.

[145] Michael Lynch. *The Origins of Genome Architecture*. Sinauer Associates Inc, Sunderland, Mass, 1st edition edition, March 2007.

[146] Laurens van der Maaten. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, 15(93):3221–3245, 2014.

[147] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[148] C. MacDonald. Development of new cell lines for animal cell biotechnology. *Crit Rev Biotechnol*, 10(2):155–178, 1990.

[149] Trudy F C Mackay, Stephen Richards, Eric A Stone, Antonio Barbadilla, Julien F Ayroles, Dianhui Zhu, Sònia Casillas, Yi Han, Michael M Magwire, Julie M Cridland, Mark F Richardson, Robert R H Anholt, Maite Barron, Crystal Bess, Kerstin Petra Blankenburg, Mary Anna Carbone, David Castellano, Lesley Chaboub, Laura Duncan, Zeke Harris, Mehwish Javaid, Joy Christina Jayaseelan, Shalini N Jhangiani, Katherine W Jordan, Fremiet Lara, Faye Lawrence, Sandra L Lee, Pablo Librado, Raquel S Linheiro, Richard F Lyman, Aaron J Mackey, Mala Munidasa, Donna Marie Muzny, Lynne Nazareth, Irene Newsham, Lora Perales, Ling-Ling Pu, Carson Qu, Miquel Ramia, Jeffrey G Reid, Stephanie M Rollmann, Julio Rozas, Nehad Saada, Lavanya Turlapati, Kim C Worley, Yuan-Qing Wu, Akihiko Yamamoto, Yiming Zhu, Casey M Bergman, Kevin R Thornton, David Mittelman, and Richard A Gibbs. The Drosophila melanogaster Genetic Reference Panel. *Nature*, 482(7384):173–178, February 2012.

[150] R. A. MacLeod, W. G. Dirks, Y. Matsuo, M. Kaufmann, H. Milch, and H. G. Drexler. Widespread intraspecies cross-contamination of human tumor cell lines arising at source. *Int J Cancer*, 83(4):555–563, November 1999.

[151] Medhat Mahmoud, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J. Sedlazeck. Structural variant calling: the long and the short of it. *Genome Biol*, 20(1):246, November 2019.

[152] Claude Maisonhaute, David Ogereau, Aurélie Hua-Van, and Pierre Capy. Amplification of the 1731 LTR retrotransposon in Drosophila melanogaster cultured cells: origin of neocopies and impact on the genome. *Gene*, 393(1-2):116–126, May 2007.

[153] H. S. Malik and T. H. Eickbush. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Res*, 11(7):1187–1197, July 2001.

[154] Manee M Manee, John Jackson, and Casey M Bergman. Conserved noncoding elements influence the transposable element landscape in Drosophila. *Genome Biol Evol*, 10(6):1533–1545, May 2018.

[155] Daniel Mariyappa, Douglas B. Rusch, Shunhua Han, Arthur Luhur, Danielle Overton, David F. B. Miller, Casey M. Bergman, and Andrew C. Zelhof. A novel transposable element based authentication protocol for Drosophila cell lines. *bioRxiv*, page 2021.08.16.456580, August 2021.

[156] Guillaume Marçais, Arthur L. Delcher, Adam M. Phillippy, Rachel Coston, Steven L. Salzberg, and Aleksey Zimin. MUMmer4: A fast and versatile genome alignment system. *PLOS Comput Biol*, 14(1):e1005944, January 2018.

[157] J. R. Masters, J. A. Thomson, B. Daly-Burns, Y. A. Reid, W. G. Dirks, P. Packer, L. H. Toji, T. Ohno, H. Tanabe, C. F. Arlett, L. R. Kelland, M. Harrison, A. Virmani, T. H. Ward, K. L. Ayres, and P. G. Debenham. Short tandem repeat profiling provides an international reference standard for human cell lines. *Proc Natl Acad Sci USA*, 98(14):8012–8017, July 2001.

[158] Alice McCarthy. Third generation DNA sequencing: Pacific Biosciences' single molecule real time technology. *Chem Biol*, 17(7):675–676, July 2010.

[159] B McClintock. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci USA*, 36(6):344–55, June 1950.

[160] Rajiv C. McCoy, Ryan W. Taylor, Timothy A. Blauwkamp, Joanna L. Kelley, Michael Kertesz, Dmitry Pushkarev, Dmitri A. Petrov, and Anna-Sophie Fiston-Lavier. Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements. *PLOS ONE*, 9(9):e106689, September 2014.

[161] Torrin L. McDonald, Weichen Zhou, Christopher P. Castro, Camille Mumm, Jessica A. Switzenberg, Ryan E. Mills, and Alan P. Boyle. Cas9 targeted enrichment of mobile elements using nanopore sequencing. *Nat Commun*, 12(1):3586, June 2021. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Subject_term: DNA sequencing;Genetic techniques;Transposition Subject_term_id: dna-sequencing;genetic-techniques;transposition.

[162] Jason R. Miller, Sergey Koren, Kari A. Dilley, Derek M. Harkins, Timothy B. Stockwell, Reed S. Shabman, and Granger G. Sutton. A draft genome sequence for the Ixodes scapularis cell line, ISE6. *F1000Res*, 7, March 2018.

[163] Jason R. Miller, Sergey Koren, Kari A. Dilley, Vinita Puri, David M. Brown, Derek M. Harkins, Françoise Thibaud-Nissen, Benjamin Rosen, Xiao-Guang Chen, Zhijian Tu, Igor V. Sharakhov, Maria V. Sharakhova, Robert Sebra, Timothy B. Stockwell, Nicholas H. Bergman, Granger G. Sutton, Adam M. Phillippy, Peter M. Piermarini, and Reed S. Shabman. Analysis of the Aedes albopictus C6/36 genome provides insight into cell line utility for viral propagation. *Gigascience*, 7(3):1–13, March 2018.

[164] S. A. Miller, D. D. Dykes, and H. F. Polesky. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res*, 16(3):1215–1215, February 1988.

[165] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534, May 2020.

[166] Akio Miyao, Mariko Nakagome, Takako Ohnuma, Harumi Yamagata, Hiroyuki Kanamori, Yuichi Katayose, Akira Takahashi, Takashi Matsumoto, and Hirohiko

Hirochika. Molecular spectrum of somaclonal variation in regenerated rice revealed by whole-genome sequencing. *Plant Cell Physiol*, 53(1):256–264, January 2012.

[167] Mourdas Mohamed, Nguyet Thi-Minh Dang, Yuki Ogyama, Nelly Burlet, Bruno Mugat, Matthieu Boulesteix, Vincent Mérel, Philippe Veber, Judit Salces-Ortiz, Dany Severac, Alain Pélisson, Cristina Vieira, François Sabot, Marie Fablet, and Séverine Chambeyron. 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, August 2020.

[168] Tabrez A. Mohammad, Yun S. Tsai, Safwa Ameer, Hung-I Harry Chen, Yu-Chiao Chiu, and Yidong Chen. CeL-ID: cell line identification using RNA-seq data. *BMC Genomics*, 20(1):81, February 2019.

[169] Stephanie E. Mohr, Yanhui Hu, Kevin Kim, Benjamin E. Housden, and Norbert Perrimon. Resources for Functional Genomics Studies in Drosophila melanogaster. *Genetics*, 197(1):1–18, May 2014.

[170] E. Montgomery, B. Charlesworth, and C. H. Langley. A test for the role of natural selection in the stabilization of transposable element copy number in a population of Drosophila melanogaster. *Genet Res*, 49(1):31–41, 1987. 0016-6723 Journal Article.

[171] Violette Morales, Tobias Straub, Martin F. Neumann, Gabrielle Mengus, Asifa Akhtar, and Peter B. Becker. Functional integration of the histone acetyltransferase MOF into the dosage compensation complex. *The EMBO Journal*, 23(11):2258–2268, June 2004.

[172] Maria Nattestad, Sara Goodwin, Karen Ng, Timour Baslan, Fritz J. Sedlazeck, Philipp Rescheneder, Tyler Garvin, Han Fang, James Gurtowski, Elizabeth Hutton, Elizabeth Tseng, Chen-Shan Chin, Timothy Beck, Yogi Sundaravadanam, Melissa Kramer, Eric Antoniou, John D. McPherson, James Hicks, W. Richard McCombie, and Michael C. Schatz. Complex rearrangements and oncogene amplifications revealed by long-read

DNA and RNA sequencing of a breast cancer cell line. *Genome Res*, 28(8):1126–1135, August 2018.

[173] Maria Nattestad and Michael C. Schatz. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics*, 32(19):3021–3023, 2016.

[174] Lidia Nefedova and Alexander Kim. Mechanisms of LTR-Retroelement Transposition: Lessons from Drosophila melanogaster. *Viruses*, 9(4):81, April 2017.

[175] Michael G. Nelson, Raquel S. Linheiro, and Casey M. Bergman. McClintock: an integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data. *G3*, 7:2749–2762, August 2017.

[176] W. A. Nelson-Rees, D. W. Daniels, and R. R. Flandermeyer. Cross-contamination of cells in culture. *Science*, 212(4493):446–452, April 1981.

[177] Yuzo Niki, Takafumi Yamaguchi, and Anthony P. Mahowald. Establishment of stable cell lines of Drosophila germ-line stem cells. *Proc Natl Acad Sci USA*, 103(44):16325–16330, October 2006.

[178] Petr Novak, Pavel Neumann, and Jiri Macas. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, 11(1):378+, July 2010.

[179] S. U. O'Brien, G. Kleiner, R. Olson, and J. E. Shannon. Enzyme polymorphisms as genetic signatures in human cell cultures. *Science*, 195(4284):1345–1348, March 1977.

[180] H. Ogura. Chromosome variation in plant tissue culture. In Y. P. S. Bajaj, editor, *Somaclonal Variation in Crop Improvement I*, Biotechnology in Agriculture and Forestry, pages 49–84. Springer, Berlin, Heidelberg, 1990.

[181] Yukiteru Ono, Kiyoshi Asai, and Michiaki Hamada. PBSIM2: a simulator for long-read sequencers with a novel generative model of quality scores. *Bioinformatics*, 37(5):589–595, March 2021.

[182] Naoki Osada, Arihiro Kohara, Toshiyuki Yamaji, Noriko Hirayama, Fumio Kasai, Tsuyoshi Sekizuka, Makoto Kuroda, and Kentaro Hanada. The Genome Landscape of the African Green Monkey Kidney-Derived Vero Cell Line. *DNA Research*, 21(6):673–683, December 2014.

[183] Walther Parson, Romana Kirchebner, Roswitha Mühlmann, Kathrin Renner, Anita Kofler, Stefan Schmidt, and Reinhard Kofler. Cancer cell line identification by short tandem repeat profiling: power and limitations. *FASEB J*, 19(3):434–436, March 2005.

[184] David J. Peel and Martin J. Milner. The diversity of cell morphology in cloned cell lines derived from Drosophila imaginal discs. *Roux's Arch Dev Biol*, 198(8):479–482, June 1990.

[185] Valentina Peona, Mozes P. K. Blom, Luohao Xu, Reto Burri, Shawn Sullivan, Ignas Bunikis, Ivan Liachko, Tri Haryoko, Knud A. Jønsson, Qi Zhou, Martin Irestedt, and Alexander Suh. Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Mol Ecol Resour*, 21(1):263–286, January 2021.

[186] Dmitri A. Petrov, Anna-Sophie Fiston-Lavier, Mikhail Lipatov, Kapa Lenkov, and Josefa Gonzalez. Population genomics of transposable elements in Drosophila melanogaster. *Mol Biol Evol*, 28(5):1633–1644, May 2011.

[187] Roy N. Platt, Yuhua Zhang, David J. Witherspoon, Jinchuan Xing, Alexander Suh, Megan S. Keith, Lynn B. Jorde, Richard D. Stevens, and David A. Ray. Targeted Capture of Phylogenetically Informative Ves SINE Insertions in Genus Myotis. *Genome Biol Evol*, 7(6):1664–1675, June 2015.

[188] Martin O. Pollard, Deepti Gurdasani, Alexander J. Mentzer, Tarryn Porter, and Manjinder S. Sandhu. Long reads: their purpose and place. *Hum Mol Genet*, 27(R2):R234–R241, August 2018.

[189] S. Steven Potter, William J. Brorein, Pamela Dunsmuir, and Gerald M. Rubin. Transposition of elements of the 412, copia and 297 dispersed repeated gene families in Drosophila. *Cell*, 17(2):415–427, June 1979.

[190] A. L. Price, N. C. Jones, and P. A. Pevzner. De novo identification of repeat families in large genomes. *Bioinformatics*, 21 Suppl 1:i351–i358, 2005. 1367-4803 Journal Article.

[191] E. Pritham. Transposable elements and factors influencing their success in eukaryotes. *J Hered*, 100(5):648–55, September 2009.

[192] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010.

[193] Reazur Rahman, Gung-wei Chirn, Abhay Kanodia, Yuliya A. Sytnikova, Björn Brembs, Casey M. Bergman, and Nelson C. Lau. Unique transposon landscapes are pervasive across Drosophila melanogaster genomes. *Nucleic Acids Res*, 43(22):10655–10672, December 2015.

[194] David A. Ray, Jinchuan Xing, Abdel-Halim Salem, and Mark A. Batzer. SINEs of a nearly perfect character. *Syst Biol*, 55(6):928–935, December 2006.

[195] Gabriel E. Rech, Santiago Radío, Sara Guirao-Rico, Laura Aguilera, Vivien Horvath, Llewellyn Green, Hannah Lindstadt, Véronique Jamilloux, Hadi Quesneville, and Josefa González. Population-scale long-read sequencing uncovers transposable elements contributing to gene expression variation and associated with adaptive signatures in Drosophila melanogaster. *bioRxiv*, page 2021.10.08.463646, January 2021.

[196] C. Ress, M. Holtmann, U. Maas, J. Sofsky, and A. Dorn. 20-Hydroxyecdysone-induced differentiation and apoptosis in the Drosophila cell line, l(2)mbn. *Tissue and Cell*, 32(6):464–477, December 2000.

[197] Arang Rhie, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, William Chow, Arkarachai Fungtammasan, Gregory L. Gedman, Lindsey J. Cantin, Francoise Thibaud-Nissen, Leanne Haggerty, Chul Lee, Byung June Ko, Juwan Kim, Iliana Bista, Michelle Smith, Bettina Haase, Jacquelyn Mountcastle, Sylke Winkler, Sadye Paez, Jason Howard, Sonja C. Vernes, Tanya M. Lama, Frank Grutzner, Wesley C. Warren, Christopher Balakrishnan, Dave Burt, Julia M. George, Mathew Biegler, David Iorns, Andrew Digby, Daryl Eason, Taylor Edwards, Mark Wilkinson, George Turner, Axel Meyer, Andreas F. Kautt, Paolo Franchini, H. William Detrich, Hannes Svardal, Maximilian Wagner, Gavin J. P. Naylor, Martin Pippel, Milan Malinsky, Mark Mooney, Maria Simbirsky, Brett T. Hannigan, Trevor Pesout, Marlys Houck, Ann Misuraca, Sarah B. Kingan, Richard Hall, Zev Kronenberg, Jonas Korlach, Ivan Sović, Christopher Dunn, Zemin Ning, Alex Hastie, Joyce Lee, Siddarth Selvaraj, Richard E. Green, Nicholas H. Putnam, Jay Ghurye, Erik Garrison, Ying Sims, Joanna Collins, Sarah Pelan, James Torrance, Alan Tracey, Jonathan Wood, Dengfeng Guan, Sarah E. London, David F. Clayton, Claudio V. Mello, Samantha R. Friedrich, Peter V. Lovell, Ekaterina Osipova, Farooq O. Al-Ajli, Simona Secomandi, Heebal Kim, Constantina Theofanopoulou, Yang Zhou, Robert S. Harris, Kateryna D. Makova, Paul Medvedev, Jinna Hoffman, Patrick Masterson, Karen Clark, Fergal Martin, Kevin Howe, Paul Flicek, Brian P. Walenz, Woori Kwak, Hiram Clawson, Mark Diekhans, Luis Nassar, Benedict Paten, Robert H. S. Kraus, Harris Lewin, Andrew J. Crawford, M. Thomas P. Gilbert, Guojie Zhang, Byrappa Venkatesh, Robert W. Murphy, Klaus-Peter Koepfli, Beth Shapiro, Warren E. Johnson, Federica Di Palma, Tomas Margues-Bonet, Emma C. Teeling, Tandy Warnow, Jennifer Marshall Graves, Oliver A. Ryder, David Hausler, Stephen J.

O'Brien, Kerstin Howe, Eugene W. Myers, Richard Durbin, Adam M. Phillippy, and Erich D. Jarvis. Towards complete and error-free genome assemblies of all vertebrate species. *bioRxiv*, page 2020.05.22.110833, May 2020.

[198] Lavanya Rishishwar, Leonardo Marino-Ramirez, and I. King Jordan. Benchmarking computational tools for polymorphic transposable element detection. *Brief Bioinformatics*, 18(6):908–918, November 2017.

[199] Sofia M. C. Robb, Lu Lu, Elizabeth Valencia, James M. Burnette, Yutaka Okumoto, Susan R. Wessler, and Jason E. Stajich. The use of RelocaTE and unassembled short reads to produce high-resolution snapshots of transposable element generated diversity in rice. *G3*, 3(6):949–957, June 2013.

[200] Stephen L. Rogers, Gregory C. Rogers, David J. Sharp, and Ronald D. Vale. Drosophila EB1 is important for proper assembly, dynamics, and positioning of the mitotic spindle. *J Cell Biol*, 158(5):873–884, September 2002.

[201] Stephen L. Rogers, Ursula Wiedemann, Nico Stuurman, and Ronald D. Vale. Molecular requirements for actin-based lamella formation in Drosophila S2 cells. *J Cell Biol*, 162(6):1079–1088, September 2003.

[202] Sushmita Roy, Jason Ernst, Peter V. Kharchenko, Pouya Kheradpour, Nicolas Negre, Matthew L. Eaton, Jane M. Landolin, Christopher A. Bristow, Lijia Ma, Michael F. Lin, Stefan Washietl, Bradley I. Arshinoff, Ferhat Ay, Patrick E. Meyer, Nicolas Robine, Nicole L. Washington, Luisa Di Stefano, Eugene Berezikov, Christopher D. Brown, Rogerio Candeias, Joseph W. Carlson, Adrian Carr, Irwin Jungreis, Daniel Marbach, Rachel Sealfon, Michael Y. Tolstorukov, Sebastian Will, Artyom A. Alekseyenko, Carlo Artieri, Benjamin W. Booth, Angela N. Brooks, Qi Dai, Carrie A. Davis, Michael O. Duff, Xin Feng, Andrey A. Gorchakov, Tingting Gu, Jorja G. Henikoff, Philipp Kapranov, Renhua Li, Heather K. MacAlpine, John Malone, Aki

Minoda, Jared Nordman, Katsutomo Okamura, Marc Perry, Sara K. Powell, Nicole C. Riddle, Akiko Sakai, Anastasia Samsonova, Jeremy E. Sandler, Yuri B. Schwartz, Noa Sher, Rebecca Spokony, David Sturgill, Marijke van Baren, Kenneth H. Wan, Li Yang, Charles Yu, Elise Feingold, Peter Good, Mark Guyer, Rebecca Lowdon, Kami Ahmad, Justen Andrews, Bonnie Berger, Steven E. Brenner, Michael R. Brent, Lucy Cherbas, Sarah C. R. Elgin, Thomas R. Gingeras, Robert Grossman, Roger A. Hoskins, Thomas C. Kaufman, William Kent, Mitzi I. Kuroda, Terry Orr-Weaver, Norbert Perrimon, Vincenzo Pirrotta, James W. Posakony, Bing Ren, Steven Russell, Peter Cherbas, Brenton R. Graveley, Suzanna Lewis, Gos Micklem, Brian Oliver, Peter J. Park, Susan E. Celniker, Steven Henikoff, Gary H. Karpen, Eric C. Lai, David M. MacAlpine, Lincoln D. Stein, Kevin P. White, and Manolis Kellis. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, 330(6012):1787–1797, December 2010.

[203] Jue Ruan and Heng Li. Fast and accurate long-read assembly with wtdbg2. *Nat Methods*, 17(2):155–158, February 2020.

[204] G. M. Rubin, M. G. Kidwell, and P. M. Bingham. The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. *Cell*, 29(3):987–94, 1982. Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. United states.

[205] F. H. Ruddle, L. Berman, and C. S. Stulberg. Chromosome analysis of five longterm cell culture populations derived from non-leukemic human peripheral blood (Detroit strains). *Cancer Res*, 18(9):1048–1059, October 1958.

[206] Joseph Russo, Andrew W. Harrington, and Mindy Steiniger. Antisense Transcription of Retrotransposons in Drosophila: The Origin of Endogenous Small Interfering RNA Precursors. *Genetics*, November 2015.

[207] Timothy B. Sackton, Rob J. Kulathinal, Casey M. Bergman, Aaron R. Quinlan, Erik B. Dopman, Mauricio Carneiro, Gabor T. Marth, Daniel L. Hartl, and Andrew G. Clark. Population genomic inferences from sparse high-throughput sequencing of two populations of Drosophila melanogaster. *Genome Biol Evol*, 1:449–65, January 2009.

[208] Kuniaki Saito, Sachi Inagaki, Toutai Mituyama, Yoshinori Kawamura, Yukiteru Ono, Eri Sakota, Hazuki Kotani, Kiyoshi Asai, Haruhiko Siomi, and Mikiko C. Siomi. A regulatory circuit for piwi by the large Maf gene traffic jam in Drosophila. *Nature*, 461(7268):1296–1299, October 2009.

[209] Abdel-Halim Salem, David A. Ray, Jinchuan Xing, Pauline A. Callinan, Jeremy S. Myers, Dale J. Hedges, Randall K. Garber, David J. Witherspoon, Lynn B. Jorde, and Mark A. Batzer. Alu elements and hominid phylogenetics. *Proc Natl Acad Sci USA*, 100(22):12787–12791, October 2003.

[210] Christos Samakovlis, Bengt Asling, Hans G. Boman, Elisabeth Gateff, and Dan Hultmark. In vitro induction of cecropin genes — an immune response in a Drosophila blood cell line. *Biochemical and Biophysical Research Communications*, 188(3):1169–1175, November 1992.

[211] W. F. Scherer, J. T. Syverton, and G. O. Gey. Studies on the propagation in vitro of poliomyelitis viruses. IV. Viral multiplication in a stable strain of human malignant epithelial cells (strain HeLa) derived from an epidermoid carcinoma of the cervix. *J Exp Med*, 97(5):695–710, May 1953.

[212] Patrick S. Schnable, Doreen Ware, Robert S. Fulton, Joshua C. Stein, Fusheng Wei, Shiran Pasternak, Chengzhi Liang, Jianwei Zhang, Lucinda Fulton, Tina A. Graves, Patrick Minx, Amy Denise Reily, Laura Courtney, Scott S. Kruchowski, Chad Tomlinson, Cindy Strong, Kim Delehaunty, Catrina Fronick, Bill Courtney, Susan M. Rock, Eddie Belter, Feiyu Du, Kyung Kim, Rachel M. Abbott, Marc Cotton, Andy Levy,

Pamela Marchetto, Kerri Ochoa, Stephanie M. Jackson, Barbara Gillam, Weizu Chen, Le Yan, Jamey Higginbotham, Marco Cardenas, Jason Waligorski, Elizabeth Applebaum, Lindsey Phelps, Jason Falcone, Krishna Kanchi, Thynn Thane, Adam Scimone, Nay Thane, Jessica Henke, Tom Wang, Jessica Ruppert, Neha Shah, Kelsi Rotter, Jennifer Hodges, Elizabeth Ingenthron, Matt Cordes, Sara Kohlberg, Jennifer Sgro, Brandon Delgado, Kelly Mead, Asif Chinwalla, Shawn Leonard, Kevin Crouse, Kristi Collura, Dave Kudrna, Jennifer Currie, Ruifeng He, Angelina Angelova, Shanmugam Rajasekar, Teri Mueller, Rene Lomeli, Gabriel Scara, Ara Ko, Krista Delaney, Marina Wissotski, Georgina Lopez, David Campos, Michele Braidotti, Elizabeth Ashley, Wolfgang Golser, HyeRan Kim, Seunghee Lee, Jinke Lin, Zeljko Dujmic, Woojin Kim, Jayson Talag, Andrea Zuccolo, Chuanzhu Fan, Aswathy Sebastian, Melissa Kramer, Lori Spiegel, Lidia Nascimento, Theresa Zutavern, Beth Miller, Claude Ambroise, Stephanie Muller, Will Spooner, Apurva Narechania, Liya Ren, Sharon Wei, Sunita Kumari, Ben Faga, Michael J. Levy, Linda McMahan, Peter Van Buren, Matthew W. Vaughn, Kai Ying, Cheng-Ting Yeh, Scott J. Emrich, Yi Jia, Ananth Kalyanaraman, An-Ping Hsia, W. Brad Barbazuk, Regina S. Baucom, Thomas P. Brutnell, Nicholas C. Carpita, Cristian Chaparro, Jer-Ming Chia, Jean-Marc Deragon, James C. Estill, Yan Fu, Jeffrey A. Jeddeloh, Yujun Han, Hyeran Lee, Pinghua Li, Damon R. Lisch, Sanzhen Liu, Zhijie Liu, Dawn Holligan Nagel, Maureen C. McCann, Phillip SanMiguel, Alan M. Myers, Dan Nettleton, John Nguyen, Bryan W. Penning, Lalit Ponnala, Kevin L. Schneider, David C. Schwartz, Anupma Sharma, Carol Soderlund, Nathan M. Springer, Qi Sun, Hao Wang, Michael Waterman, Richard Westerman, Thomas K. Wolfgruber, Lixing Yang, Yeisoo Yu, Lifang Zhang, Shiguo Zhou, Qihui Zhu, Jeffrey L. Bennetzen, R. Kelly Dawe, Jiming Jiang, Ning Jiang, Gernot G. Presting, Susan R. Wessler, Srinivas Aluru, Robert A. Martienssen, Sandra W. Clifton, W. Richard McCombie, Rod A. Wing, and Richard K. Wilson. The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956):1112–1115, November 2009.

[213] Imogene Schneider. Cell lines derived from late embryonic stages of Drosophila melanogaster. *J Embryol Exp Morphol*, 27(2):353–365, April 1972.

[214] M. D. Schug, T. F. Mackay, and C. F. Aquadro. Low mutation rates of microsatellite loci in Drosophila melanogaster. *Nat Genet*, 15(1):99–102, January 1997.

[215] Yuri B. Schwartz, Tatyana G. Kahn, David A. Nix, Xiao-Yong Li, Richard Bourgon, Mark Biggin, and Vincenzo Pirrotta. Genome-wide analysis of Polycomb targets in Drosophila melanogaster. *Nature Genetics*, 38(6):700–705, June 2006.

[216] Fritz J. Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C. Schatz. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods*, 15(6):461–468, June 2018.

[217] A. M. Shedlock and N. Okada. SINE insertions: powerful tools for molecular systematics. *Bioessays*, 22(2):148–160, February 2000.

[218] Grzegorz Sienski, Derya Dönertas, and Julius Brennecke. Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell*, 151(5):964–980, November 2012.

[219] Felipe A. Simao, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, October 2015.

[220] Nadia D. Singh and Dmitri A. Petrov. Rapid sequence turnover at an intergenic locus in Drosophila. *Molecular Biology and Evolution*, 21(4):670–80, April 2004.

[221] Mikiko C. Siomi, Kuniaki Saito, and Haruhiko Siomi. How selfish retrotransposons are silenced in Drosophila germline and somatic cells. *FEBS Letters*, 582(17):2473–2478, July 2008.

[222] R. Keith Slotkin and Robert Martienssen. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet*, 8(4):272–85, April 2007.

[223] Arian FA Smit. The origin of interspersed repeats in the human genome. *Curr Opin Genetics Dev*, 6(6):743–748, December 1996.

[224] Caiti Smukowski Heil, Kira Patterson, Angela Shang-Mei Hickey, Erica Alcantara, and Maitreya J Dunham. Transposable element mobilization in interspecific yeast hybrids. *Genome Biol Evol*, 13(3):evab033, February 2021.

[225] Charlotte Soneson, Michael I. Love, and Mark D. Robinson. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res*, 4:1521, December 2015.

[226] Sarah Song, Katia Nones, David Miller, Ivon Harliwong, Karin S. Kassahn, Mark Pinese, Marina Pajic, Anthony J. Gill, Amber L. Johns, Matthew Anderson, Oliver Holmes, Conrad Leonard, Darrin Taylor, Scott Wood, Qinying Xu, Felicity Newell, Mark J. Cowley, Jianmin Wu, Peter Wilson, Lynn Fink, Andrew V. Biankin, Nic Waddell, Sean M. Grimmond, and John V. Pearson. qpure: a tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles. *PLOS ONE*, 7(9):e45835, September 2012.

[227] Craig E. Stanley and Rob J. Kulathinal. Genomic signatures of domestication on neurogenetic genes in Drosophila melanogaster. *BMC Evolutionary Biology*, 16(1):6, January 2016.

[228] Marcus Stoiber, Susan Celniker, Lucy Cherbas, Ben Brown, and Peter Cherbas. Diverse Hormone Response Networks in 41 Independent Drosophila Cell Lines. *G3*, 6(3):683–694, March 2016.

[229] Jeet Sukumaran and Mark T. Holder. DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, 26(12):1569–1571, June 2010.

[230] D.L. Swofford. *PAUP*: phylogenetic analysis using parsimony (* and other methods).* Sinauer Associates, Sunderland, Massachusetts, 2003.

[231] Yuliya A. Sytnikova, Reazur Rahman, Gung-wei Chirn, Josef P. Clark, and Nelson C. Lau. Transposable element dynamics and PIWI regulation impacts lncRNA and gene expression diversity in Drosophila ovarian cell cultures. *Genome Res*, 24(12):1977–1990, December 2014.

[232] Lorenzo Tattini, Romina D'Aurizio, and Alberto Magi. Detection of genomic structural variants from next-generation sequencing data. *Front Bioeng Biotechnol*, 3:92, June 2015.

[233] K. Ui, R. Ueda, and T. Miyake. Cell lines from imaginal discs of Drosophila melanogaster. *In Vitro Cell Dev Biol*, 23(10):707–711, October 1987.

[234] V. V. Vagin, A. Sigova, C. Li, H. Seitz, V. Gvozdev, and P. D. Zamore. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science*, 313(5785):320–4, 2006. 1095-9203 (Electronic) Journal article.

[235] Pol Vendrell-Mir, Fabio Barteri, Miriam Merenciano, Josefa González, Josep M. Casacuberta, and Raúl Castanera. A benchmark of transposon insertion detection tools using real data. *Mob DNA*, 10(1):53, December 2019.

[236] Raghuvir Viswanatha, Zhongchi Li, Yanhui Hu, and Norbert Perrimon. Pooled genome-wide CRISPR screening for basal and context-specific fitness gene essentiality in Drosophila cells. *eLife*, 7:e36333, July 2018.

[237] C. Vitte and O. Panaud. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice Oryza sativa L. *Mol Biol Evol*, 20(4):528–540, April 2003.

[238] Jun Wang, Peter D. Keightley, and Daniel L. Halligan. Effect of divergence time and recombination rate on molecular evolution of Drosophila INE-1 transposable elements

and other candidates for neutrally evolving sites. *J Mol Evol*, 65(6):627, December 2007.

[239] Robert M. Waterhouse, Mathieu Seppey, Felipe A. Simão, Mosè Manni, Panagiotis Ioannidis, Guennadi Klioutchnikov, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol*, 35(3):543–548, March 2018.

[240] Neil I. Weisenfeld, Vijay Kumar, Preyas Shah, Deanna M. Church, and David B. Jaffe. Direct determination of diploid genome sequences. *Genome Res*, 27(5):757–767, 2017.

[241] Jiayu Wen, Jaaved Mohammed, Diane Bortolamiol-Becet, Harrison Tsai, Nicolas Robine, Jakub O. Westholm, Erik Ladewig, Qi Dai, Katsutomo Okamura, Alex S. Flynt, Dayu Zhang, Justen Andrews, Lucy Cherbas, Thomas C. Kaufman, Peter Cherbas, Adam Siepel, and Eric C. Lai. Diversity of miRNAs, siRNAs, and piRNAs across 25 Drosophila cell lines. *Genome Res*, 24(7):1236–1250, July 2014.

[242] Travis J. Wheeler, Jody Clements, Sean R. Eddy, Robert Hubley, Thomas A. Jones, Jerzy Jurka, Arian F. A. Smit, and Robert D. Finn. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res*, 41(Database issue):D70–82, January 2013.

[243] Thomas Wicker, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, Etienne Paux, Phillip SanMiguel, and Alan H. Schulman. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*, 8(12):973–982, December 2007.

[244] Jinchuan Xing, Hui Wang, Kyudong Han, David A. Ray, Cheney H. Huang, Leona G. Chemnick, Caro-Beth Stewart, Todd R. Disotell, Oliver A. Ryder, and Mark A. Batzer.

A mobile element based phylogeny of Old World monkeys. *Mol Phylogenet Evol*, 37(3):872–880, December 2005.

[245] S. Yanagawa, J. S. Lee, and A. Ishimoto. Identification and characterization of a novel line of Drosophila Schneider S2 cells that respond to wingless signaling. *J Biol Chem*, 273(48):32353–32359, November 1998.

[246] Nachen Yang, Satyam P. Srivastav, Reazur Rahman, Qicheng Ma, Gargi Dayama, Madoka Chinen, Elissa P. Lei, Michael Rosbash, and Nelson C. Lau. Transposable element landscape changes are buffered by RNA silencing in aging Drosophila. *bioRxiv*, page 2021.01.08.425853, January 2021.

[247] Andrew D Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, José Carlos Marugán, Carla Cummins, Claire Davidson, Kamalkumar Dodiya, Reham Fatima, Astrid Gall, Carlos Garcia Giron, Laurent Gil, Tiago Grego, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G Izuogu, Sophie H Janacek, Thomas Juettemann, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, Jose Gonzalez Martinez, Thomas Maurel, Mark McDowall, Aoife McMahon, Shamika Mohanan, Benjamin Moore, Michael Nuhn, Denye N Oheh, Anne Parker, Andrew Parton, Mateus Patricio, Manoj Pandian Sakthivel, Ahamed Imran Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, Mira Sycheva, Marek Szuba, Kieron Taylor, Anja Thormann, Glen Threadgold, Alessandro Vullo, Brandon Walts, Andrea Winterbottom, Amonida Zadissa, Marc Chakiachvili, Bethany Flint, Adam Frankish, Sarah E Hunt, Garth IIsley, Myrto Kostadima, Nick Langridge, Jane E Loveland, Fergal J Martin, Joannella Morales, Jonathan M Mudge, Matthieu Muffato, Emily Perry, Magali Ruffier, Stephen J Trevanion, Fiona Cunningham, Kevin L Howe, Daniel R Zerbino, and Paul Flicek. Ensembl 2020. *Nucleic Acids Res*, 48(D1):D682–D688, January 2020.

[248] Mamie Yu, Suresh K. Selvaraj, May M. Y. Liang-Chu, Sahar Aghajani, Matthew Busse, Jean Yuan, Genee Lee, Franklin Peale, Christiaan Klijn, Richard Bourgon, Joshua S. Kaminker, and Richard M. Neve. A resource for cell line authentication, annotation and quality control. *Nature*, 520(7547):307–311, April 2015.

[249] Tianxiong Yu, Xiao Huang, Shengqian Dou, Xiaolu Tang, Shiqi Luo, William E Theurkauf, Jian Lu, and Zhiping Weng. A benchmark and an algorithm for detecting germline transposon insertions and measuring de novo transposon insertion frequencies. *Nucleic Acids Res*, 49(8):e44, May 2021.

[250] Sophie Zaaijer, Assaf Gordon, Daniel Speyer, Robert Piccone, Simon Cornelis Groen, and Yaniv Erlich. Rapid re-identification of human samples using portable DNA sequencing. *eLife*, 6:e27798, November 2017.

[251] John G. Zampella, Nemanja Rodić, Wan Rou Yang, Cheng Ran Lisa Huang, Jane Welch, Veena P. Gnanakkan, Toby C. Cornish, Jef D. Boeke, and Kathleen H. Burns. A map of mobile DNA insertions in the NCI-60 human cancer cell panel. *Mob DNA*, 7(1):20, October 2016.

[252] Vanessa Zanni, Angéline Eymery, Michael Coiffet, Matthias Zytnicki, Isabelle Luyten, Hadi Quesneville, Chantal Vaury, and Silke Jensen. Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. *Proc Natl Acad Sci USA*, 110(49):19842–19847, December 2013.

[253] Yu Zhang, John H. Malone, Sara K. Powell, Vipul Periwal, Eric Spana, David M. Macalpine, and Brian Oliver. Expression in aneuploid Drosophila S2 cells. *PLOS biology*, 8(2):e1000320+, February 2010.

[254] Xuefang Zhao, Ryan L. Collins, Wan-Ping Lee, Alexandra M. Weber, Yukyung Jun, Qihui Zhu, Ben Weisburd, Yongqing Huang, Peter A. Audano, Harold Wang, Mark Walker, Chelsea Lowther, Jack Fu, Mark B. Gerstein, Scott E. Devine, Tobias

Marschall, Jan O. Korbel, Evan E. Eichler, Mark J. P. Chaisson, Charles Lee, Ryan E. Mills, Harrison Brand, and Michael E. Talkowski. Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *The American Journal of Human Genetics*, 0(0), March 2021.

[255] Grace X. Y. Zheng, Billy T. Lau, Michael Schnall-Levin, Mirna Jarosz, John M. Bell, Christopher M. Hindson, Sofia Kyriazopoulou-Panagiotopoulou, Donald A. Masquelier, Landon Merrill, Jessica M. Terry, Patrice A. Mudivarti, Paul W. Wyatt, Rajiv Bharadwaj, Anthony J. Makarewicz, Yuan Li, Phillip Belgrader, Andrew D. Price, Adam J. Lowe, Patrick Marks, Gerard M. Vurens, Paul Hardenbol, Luz Montesclaros, Melissa Luo, Lawrence Greenfield, Alexander Wong, David E. Birch, Steven W. Short, Keith P. Bjornson, Pranav Patel, Erik S. Hopmans, Christina Wood, Sukhvinder Kaur, Glenn K. Lockwood, David Stafford, Joshua P. Delaney, Indira Wu, Heather S. Ordonez, Susan M. Grimes, Stephanie Greer, Josephine Y. Lee, Kamila Belhocine, Kristina M. Giorda, William H. Heaton, Geoffrey P. McDermott, Zachary W. Bent, Francesca Meschi, Nikola O. Kondov, Ryan Wilson, Jorge A. Bernate, Shawn Gauby, Alex Kindwall, Clara Bermejo, Adrian N. Fehr, Adrian Chan, Serge Saxonov, Kevin D. Ness, Benjamin J. Hindson, and Hanlee P. Ji. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, 34(3):303–311, March 2016.

[256] Bo Zhou, Steve S. Ho, Stephanie U. Greer, Noah Spies, John M. Bell, Xianglong Zhang, Xiaowei Zhu, Joseph G. Arthur, Seunggyu Byeon, Reenal Pattni, Ishan Saha, Yiling Huang, Giltae Song, Dimitri Perrin, Wing H. Wong, Hanlee P. Ji, Alexej Abyzov, and Alexander E. Urban. Haplotype-resolved and integrated genome analysis of the cancer cell line HepG2. *Nucleic Acids Res*, 47(8):3846–3861, May 2019.

[257] Bo Zhou, Steve S. Ho, Stephanie U. Greer, Xiaowei Zhu, John M. Bell, Joseph G. Arthur, Noah Spies, Xianglong Zhang, Seunggyu Byeon, Reenal Pattni, Noa Ben-

Efraim, Michael S. Haney, Rajini R. Haraksingh, Giltae Song, Hanlee P. Ji, Dimitri Perrin, Wing H. Wong, Alexej Abyzov, and Alexander E. Urban. Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Res*, 29(3):472–484, March 2019.

[258] Weichen Zhou, Sarah B. Emery, Diane A. Flasch, Yifan Wang, Kenneth Y. Kwan, Jeffrey M. Kidd, John V. Moran, and Ryan E. Mills. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res*, 48(3):1146–1163, February 2020.

[259] Jiali Zhuang, Jie Wang, William Theurkauf, and Zhiping Weng. TEMP: a computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Res*, 42(11):6826–6838, June 2014.