OXFORD GENETICS

GENETICS, 2022, 221(3), iyac077
https://doi.org/10.1093/genetics/iyac077
Advance Access Publication Date: 10 May 2022
Investigation

Ongoing transposition in cell culture reveals the phylogeny of diverse *Drosophila S2* sublines

Shunhua Han [1], Guilherme B. Dias [1,2], Preston J. Basting [1], Michael G. Nelson [3], Sanjai Patel [3], Mar Marzo [3], Casey M. Bergman [1,2,*]

[1]Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA, ,
[2]Department of Genetics, University of Georgia, Athens, GA 30602, USA, ,
[3]Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, UK

*Corresponding author: Department of Genetics & Institute of Bioinformatics, University of Georgia, Davison Life Sciences Building, 120 E. Green St., Athens, GA 30602, USA. Email: cbergman@uga.edu

Abstract

Cultured cells are widely used in molecular biology despite poor understanding of how cell line genomes change in vitro over time. Previous work has shown that *Drosophila* cultured cells have a higher transposable element content than whole flies, but whether this increase in transposable element content resulted from an initial burst of transposition during cell line establishment or ongoing transposition in cell culture remains unclear. Here, we sequenced the genomes of 25 sublines of *Drosophila* S2 cells and show that transposable element insertions provide abundant markers for the phylogenetic reconstruction of diverse sublines in a model animal cell culture system. DNA copy number evolution across S2 sublines revealed dramatically different patterns of genome organization that support the overall evolutionary history reconstructed using transposable element insertions. Analysis of transposable element insertion site occupancy and ancestral states support a model of ongoing transposition dominated by episodic activity of a small number of retrotransposon families. Our work demonstrates that substantial genome evolution occurs during long-term *Drosophila* cell culture, which may impact the reproducibility of experiments that do not control for subline identity.

Keywords: *Drosophila*; transposable element; copy number variation; genome evolution; cell culture

Introduction

Animal cell lines play vital roles in biology by providing an abundant source of material to study molecular processes and as cellular factories to express important biomolecules. Like all living systems, animal cell lines undergo genomic changes during routine propagation in vitro (Ruddle *et al.* 1958), leading to genetic diversity across time and laboratories that can lead to irreproducible research outcomes (Hughes *et al.* 2007). 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 sublines of long-term metazoan cell cultures (Junakovic *et al.* 1988; Di Franco *et al.* 1992; Ben-David *et al.* 2018; Liu *et al.* 2019), or how to identify and minimize the impact of such diversity (Hughes *et al.* 2007; Ben-David *et al.* 2018). 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 in vivo is known to vary across taxa (Lynch 2007).

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 (Potter *et al.* 1979; Ilyin *et al.* 1980). More recently, analysis of whole-genome sequence (WGS) data revealed between ~800 and ~3,000 nonreference TE insertions in different *Drosophila* cell lines (Rahman *et al.* 2015). The mechanisms that permit this proliferation of TEs in *Drosophila* cell lines are unknown, and are unexpected given the activity of small RNA-based pathways that regulate TE expression in somatic cells (Czech *et al.* 2008). Arkhipova *et al.* (1995) provided two non-mutually exclusive hypotheses to explain the proliferation of TEs in cell lines vs whole flies: ongoing transposition is more easily tolerated in cultured cells and is not as strongly selected against as it is in whole flies; or specific factors exist that regulate TE transposition, and their actions are altered significantly in cell culture.

In addition to the question of why proliferation of TEs occurs in *Drosophila* cell line genomes, it is unknown when TE proliferation occurred during cell line evolution. TE proliferation could be caused 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 (Echalier 1997). Di Franco *et al.* (1992) contrasted the stability of TE profiles among sublines of one of the oldest *Drosophila* cell lines (Kc) (Junakovic *et al.* 1988) 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

Received: January 10, 2022. **Accepted:** April 28, 2022

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. Subsequently, 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 subline of the OSS cell line (OSS_C) relative to a putative frozen progenitor subline (OSS_E). However, this conclusion is questioned by results reported in Han *et al.* (2021) showing that OSS_E is actually a misidentified version of a related cell line (OSC). More recently, Mariyappa *et al.* (2022) cultured S2R+ cells for 50 passages and showed relative stability of TE profiles for a subset of families, suggesting that proliferation of TEs during routine cell culture may not occur rapidly on short time scales. Providing definitive evidence showing that ongoing transposition occurs in cell culture over longer time scales is important because this process could lead to genomic variation among sublines that could impact functional studies and, more practically, provide useful markers for cell line identification and reconstruction of cell line evolutionary history (Han *et al.* 2021; Mariyappa *et al.* 2022).

Here, we contribute to the understanding of genome evolution during long-term animal cell culture using a large sample of sublines of *Drosophila* Schneider Line 2 (S2) cells, one of the most widely used non-mammalian cell culture systems (Bairoch 2018). S2 cells were established from embryonic tissue of an unmarked stock of Oregon-R flies in December 1969 (Schneider 1972) and are likely to be derived from macrophage-like hemocytes (Schneider 1972; Echalier 1997). Two other cell lines, S1 (August 1969) and S3 (February 1970), were derived from the same ancestral fly stock (Schneider 1972) and can serve as outgroups to analyze evolution in the S2 lineage (Lewerentz *et al.* 2022). Since their establishment, S2 cells have been distributed widely and grown more extensively than S1 or S3 cells (Lee *et al.* 2014). Many different sublines 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 sublines of S2 cells are unknown, as is the extent of their genomic or phenotypic diversity. At least one subtype of S2 cells, called S2R+ (for S2 receptor plus), is known to have distinct phenotypes from other S2 cell lines including expressing the Dfrizzled-1 and Dfrizzled-2 membrane proteins and having the desirable property of being more adherent to surfaces in tissue culture (Yanagawa *et al.* 1998). In addition to their ubiquity and diversity, S2 cells are a good model to study genome evolution in animal cell culture because of the wealth of prior biological knowledge in *D. melanogaster* and their relatively small genome size, which permits cost-effective whole-genome sequencing.

In this study, we report new WGS data for 25 sublines of S2 cells as well as the outgroup S1 and S3 cell lines. We analyze these data together with public WGS samples for S2R+ and mbn2 [recently shown by Han *et al.* (2021) to be a misidentified lineage of S2] and demonstrate that TE insertions provide abundant markers to reconstruct the evolutionary history of S2 sublines. These data reveal that publicly available S2 sublines form a monophyletic group defined by 2 major clades (A and B), and suggest that misidentification of available S2 cultures by other *Drosophila* cell lines is limited. Furthermore, we show that genome-wide copy number profiles support the major phylogenetic relationships among S2 sublines inferred using TE profiles. Using TE site occupancy and ancestral states, 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 has proliferated during S2 evolution, most of which are retrotransposons that do not encode a retroviral envelope (*env*) gene. Together, these results support the conclusions that TE proliferation in *Drosophila* somatic cell culture is primarily driven by an ongoing, episodic, cell-autonomous process that does not involve deregulation of global transpositional control mechanisms and that TE insertions provide useful markers of S2 subline identity and genome organization.

Materials and methods

Genome sequencing

We sequenced the genomes of 29 samples of S1, S2, or S3 cells to understand the genomic diversity and evolutionary relationships of publicly available sublines of S2 cells. Frozen stocks for each of these 29 samples were ordered from the 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 Supplementary Table 1. Cells were defrosted and 250 µl of the cell suspension was aliquoted and spun down for 5 min at 300 g. 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 3 batches, each of which contained an independent sample of S2-DRSC (DGRC-181) to identify any potential sample swaps within batches and to assess the reproducibility of phylogenetic clustering based on TE profiles. The triplicate samples of S2-DRSC were from the same freeze of this cell subline 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 <100 bp, and libraries were normalized and pooled prior to being sequenced on an Illumina HiSeq 2500 flow cell using a 101-bp paired-end layout.

In addition, we analyzed public WGS data for a sample of S2R+ (Han *et al.* 2022) and 3 samples of mbn2, a cell line which was recently shown to be a misidentified lineage of S2 cells (Han *et al.* 2021). A summary of the sequence data analyzed for each of the 33 samples in this study can be found in Supplementary Table 1.

Prediction of nonreference TE insertions

Nonreference TE insertions were detected in each sample using trimmed paired fastq sequences as input for the TEMP (Zhuang *et al.* 2014) module in McClintock (v2.0) (Nelson *et al.* 2017). We used TEMP to predict nonreference TEs based on previous results showing TEMP predictions are the least dependent on coverage and read length relative to other component methods in McClintock (Han *et al.* 2021). By default, McClintock filters predictions made by TEMP by requiring at least 1 read support on both sides of insertion and at least 10% TE allele frequency. The major sequences (chr2L, chr2R, chr3L, chr3R, chr4, chrM, chrY, and chrX) from the *D. melanogaster* dm6 assembly were used as a reference genome (Hoskins *et al.* 2015). The TE library used for McClintock analysis was a slightly modified version of the Berkeley *Drosophila* Genome Project canonical TE dataset described in Sackton *et al.* (2009) (https://github.com/bergmanlab/transposons/blob/master/releases/D_mel_transposon_sequence_set_v10.2.fa; accessed 2022 May 12).

Genome-wide nonreference TE predictions generated by McClintock were filtered to only include those in normal recombination regions (chrX: 405,967–20,928,973, chr2L: 200,000–20,100,000, chr2R: 6,412,495–25,112,477, chr3L: 100,000–21,906,900, chr3R: 4,774,278–31,974,278) using boundaries defined by Cridland *et al.* (2013) lifted over to dm6 coordinates, as in Han *et al.* (2021). Our analysis was restricted to normal recombination regions since low recombination regions have high reference TE content which reduces the ability to predict nonreference TE insertions (Bergman *et al.* 2006; Manee *et al.* 2018). We also excluded *INE-1* family from the subsequent analysis since this family has been reported to be inactive in *Drosophila* for millions of years (Singh and Petrov 2004; Wang *et al.* 2007). Filtered nonreference TE predictions were then clustered across genomic coordinates and samples. TEs predicted in different samples in the same cluster were required to directly overlap and be on the same strand. Clustered nonreference TE predictions were then filtered to exclude low-quality predictions by retaining nonreference TE loci with a single TE family per locus and one prediction per sample using the same criteria as in Han *et al.* (2021).

Phylogenetic analysis of cell subline samples using TE insertion profiles

Genome-wide nonreference TE predictions were then converted to a binary presence/absence matrix as input for phylogenetic analysis. Phylogenetic trees of cell sublines were built using Dollo parsimony in PAUP (v4.0a168) (Swofford 2003). 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 (Batzer and Deininger 2002; Han *et al.* 2021). "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) (Sukumaran and Holder 2010).

Copy number analysis of cell subline samples

BAM files generated by McClintock were used to generate copy number profiles for nonoverlapping windows of the dm6 genome using Control-FREEC (v11.6) (Boeva *et al.* 2012). 10-kb windows were used for Control-FREEC analyses unless specified otherwise. Windows with less than 85% mappability were excluded from the analysis based on mappability tracks generated by GEM (v1.315 beta) (Derrien *et al.* 2012). Baseline ploidy was set to diploid for S1 and tetraploid for all other samples, according to ploidy levels for S1, S2, S2R+, S3, and mbn2 cells estimated by Lee *et al.* (2014). The minimum and maximum expected values of the GC content were set to be 0.3 and 0.45, respectively.

Results

Genome-wide TE profiles reveal the evolutionary relationships among Schneider cell sublines

Previously, we showed that genome-wide TE profiles can be used to uniquely identify *Drosophila* cell lines and provide insight into the evolutionary history of clonally evolving sublines derived from the same cell line (Han *et al.* 2021). Here, we propose that TE profiles can also be used to infer the currently unknown evolutionary relationships for a large panel of diverse sublines originating from a widely used animal cell line, *Drosophila* S2 cells. We generated paired-end Illumina WGS data for a panel of 25 *Drosophila* S2 sublines from multiple lab origins (Supplementary Table 1), including triplicate samples of one subline (S2-DRSC) to act as an internal control, and for the S1 and S3 cell lines that were derived from the same ancestral fly stock (Oregon-R) as the S2 lineage (Schneider 1972). We also included a S2R+ subline from the *Drosophila* RNAi Screening Center (DRSC) reported in Han *et al.* (2022) and three mbn2 cell subline samples from Han *et al.* (2021) (Supplementary Table 1). mbn2 cells were originally reported to have a distinct origin (Gateff *et al.* 1980), but recent genomic analysis has shown that currently circulating mbn2 cells are a misidentified lineage of S2 cells (Han *et al.* 2021), although it remains unknown to which lineage mbn2 cells are most closely related. With the exception of four cell lines that were definitively reported to be cloned from single cells (S2R+-NPT005, S2R+-NPT017, S2R+-NPT050, and S2R+-NPT101) (Neumüller *et al.* 2012), we assume the majority of cell lines in this study to be polyclonal, even those that carry stably transfected plasmid-based transgenes maintained by resistance markers. Single nucleotide polymorphism (SNP) profiles revealed a similar pattern of low heterozygosity across the entire genome for all samples, implying that the original stock of Oregon-R used to independently establish the S1, S2, and S3 cell lines was effectively isogenic (Supplementary Fig. 1).

We predicted between 655 and 2,924 nonreference TE insertions in the euchromatic regions of these Schneider cell line samples using TEMP (Zhuang *et al.* 2014) (Supplementary Table 2). Each sample had a unique profile of nonreference TE insertions (Supplementary File 1). We performed phylogenetic analysis on genome-wide TE profiles of all Schneider cell line samples using the Dollo parsimony approach (Han *et al.* 2021). This approach fits the assumptions of the homoplasy-free nature of TE insertions (Shedlock and Okada 2000; Salem *et al.* 2003; Xing *et al.* 2005; Platt *et al.* 2015; Lammers *et al.* 2017, 2019) while also accommodating the false negative (FN) predictions inherent to short read-based TE detection methods (Nelson *et al.* 2017; Rishishwar *et al.* 2017; Vendrell-Mir *et al.* 2019). The most parsimonious tree revealed several expected patterns that suggest using TE profiles to infer the evolutionary relationship among Schneider cell lines is reliable (Fig. 1a; Supplementary File 2). First, most internal nodes have high bootstrap support. All weakly supported nodes are close to the terminal taxa, which presumably is due to the lack of phylogenetically informative TE insertions that differentiate very closely related sublines or sample replicates. Second, using a hypothetical ancestor representing the state without any nonreference insertions to root the tree, S1 and S3 cell lines were independently reconstructed as outgroups for the S2 sublines in the phylogeny, as expected based on their independent origin from the same ancestral fly stock (Schneider 1972). Third, replicate samples of S2-DRSC cluster as nearest taxa and form a monophyletic clade with 100% bootstrap support. Fourth, all samples from S2R+, which are sublines of S2 with unique phenotypic characteristics (Yanagawa *et al.* 1998), form a monophyletic clade with 100% bootstrap support. Finally, all mbn2 sublines form a monophyletic clade with 100% bootstrap support embedded within a monophyletic clade of S2 sublines that itself has 100% bootstrap support. These results suggest that TE profiles can be used to reliably infer the evolutionary relationship among diverse sublines of a widely used animal cell line, and that there is no evidence for any S2 sublines in our dataset being a misidentified non-S2 *Drosophila* cell lines.

The phylogeny of Schneider cell lines built using TE profiles revealed a major split in the history of S2 cell line evolution, resulting in two sister lineages which we labeled as "Clade A" and "Clade B" (Fig. 1). Clade A comprised one subclade containing all 7 S2R+ sublines and another subclade containing six S2 sublines,

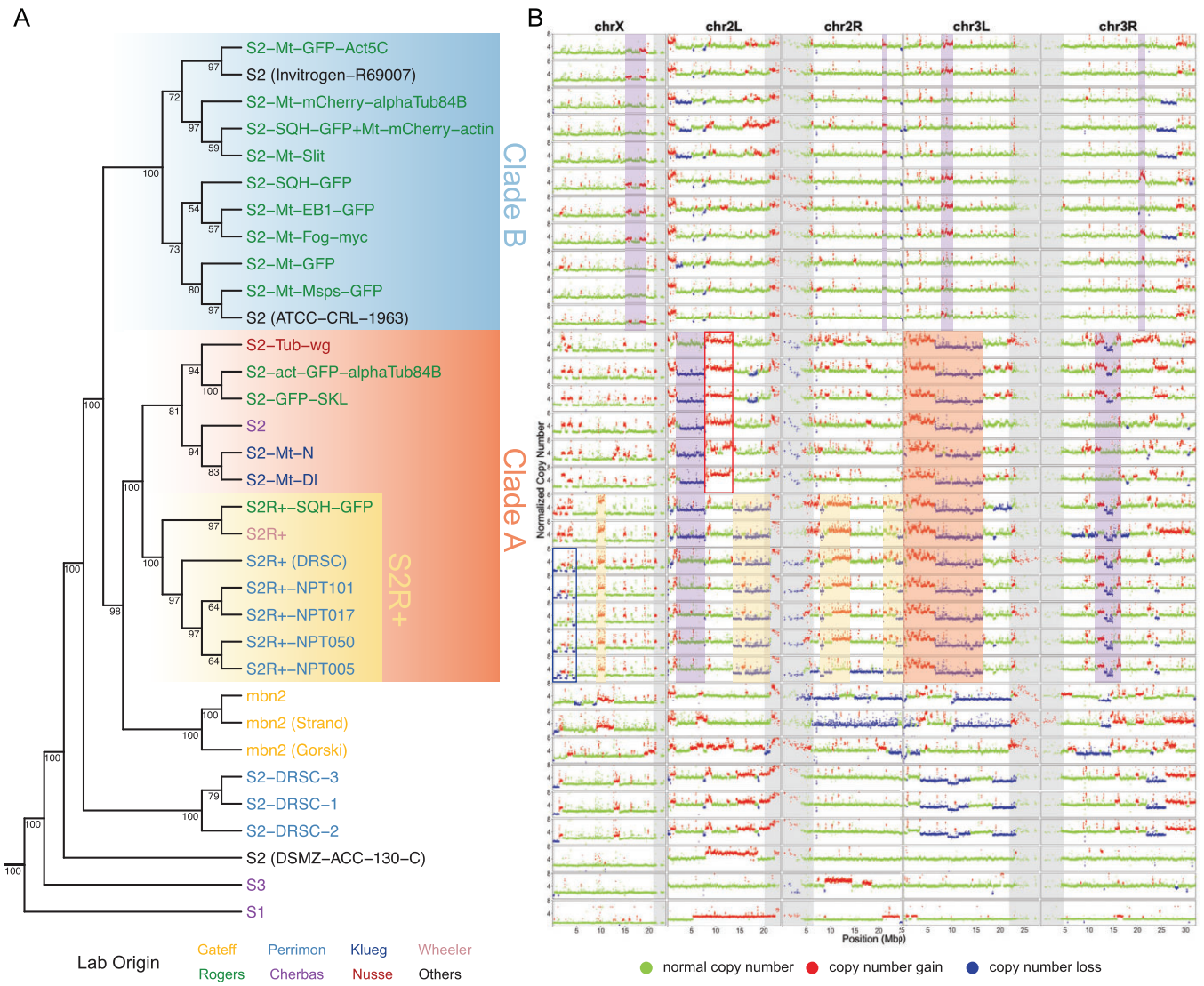

Fig. 1. TE and CNV profiles reveal the evolutionary relationship among S2 sublines. a) Dollo parsimony tree for a panel of 26 S2 sublines with diverse lab origins, two S1 and S3 sublines to serve as outgroups in the phylogeny, and three mbn2 sublines that were inferred to be misidentified S2 lines by Han *et al.* (2021). Replicate samples for S2-DRSC were also included. The phylogeny was constructed using genome-wide nonreference TE insertions predicted by TEMP (Zhuang *et al.* 2014). Percent bootstrap support is annotated below each node. 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 sublines were developed. b) Copy number profiles separated by chromosome arms for all samples included in panel a. Each data point represents normalized copy number (ratio×ploidy) for a given 10-kb window estimated by Control-FREEC (Boeva *et al.* 2012). 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 computed by Control-FREEC and baseline ploidy estimated by Lee *et al.* (2014). Red shading indicates CNVs that are exclusively shared by all S2 sublines in Clade A. Yellow shading indicates CNVs that are exclusively shared by S2R+ sublines. The red box represents CNVs on chromosome X that are exclusively shared by all S2 sublines in Clade A that are not S2R+. The blue box represents CNVs on chromosome arm 2L that are exclusively shared by S2R+ sublines from the Perrimon lab. Purple shading indicates CNVs that are exclusively shared by a subset of S2 sublines within Clade A or Clade B. Low recombination regions are shaded in gray.

one of which is the canonical S2 subline distributed by DGRC (DGRC-6). Clade B comprised 11 S2 sublines including sublines from Invitrogen and ATCC. The presence of S2 sublines in both Clade A and Clade B, but the presence of S2R+ sublines in Clade A, implies that the S2 cell line designation is paraphyletic (i.e. some S2 sublines are more closely related to S2R+ than they are to other S2 sublines). In some cases, Schneider cell lines from the same lab cluster together (e.g. S2R+ sublines from the Perrimon lab and S2 sublines from the Klueg lab, respectively). However, S2 sublines from the Rogers lab were placed in different major clades of the S2 phylogeny (three S2-sublines in Clade A, nine S2-sublines in Clade B, Fig. 1), demonstrating that the same lab can use divergent sublines of S2 from different major clades that have potentially different genome organization (see below).

The majority of S2 sublines we surveyed in this study were placed within Clade A and Clade B based on their TE profiles. However, two S2 sublines, S2-DRSC and S2 (DSMZ-ACC-130-C), were independently placed as outgroups for the two major clades of S2, suggesting that they are highly 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 subline from DGRC (i.e. DGRC-6) was previously not known. Our results suggest that S2-DRSC and S2 (DGRC-6) are not closely related sublines, which could explain

the phenotypic and functional differences between these two sublines reported in previous studies (Cherbas *et al.* 2011; Lee *et al.* 2014; Wen *et al.* 2014; Lee and Oliver 2015).

mbn2 sublines cluster in a monophyletic clade that is sister to Clade A (98% bootstrap support) but is clearly contained within a monophyletic lineage containing all S2 samples. This observation is consistent with previous results reported by Han *et al.* (2021) proposing that mbn2 is a misidentified S2 lineage. Han *et al.* (2021) showed that mbn2 clusters with S2-DRSC before clustering with S2R+. However, our results showed that the mbn2 clade clusters Clade A (containing S2R+ sublines) before clustering with S2-DRSC. We interpret this discrepancy as being caused by the sparse sampling and use of low coverage sequencing data for S2 and S2R+ from the modENCODE project in the previous study (Han *et al.* 2021), which led to insufficient signal to infer the evolutionary relationship of the mbn2 clade within S2 subline diversity.

Genome-wide copy number profiles correlate with history of S2 sublines

To further investigate potential genomic heterogeneity among Schneider cell lines and cross-validate our phylogenetic reconstruction based on TE profiles, we generated copy number profiles for all samples in our dataset (Fig. 1b) using Control-FREEC (Boeva *et al.* 2012). Two patterns in the copy number profiles generated suggested that our approach to characterize segmental variation in our cell sublines was robust. First, we observed a high concordance in copy number profiles for replicate samples of S2-DRSC (Fig. 1b). Second, copy number profiles we generated using our new data for S1, S2R+, S2-DRSC, and S3 are broadly consistent with profiles for these cell lines using data generated by the modENCODE project reported previously in Lee *et al.* (2014) (Supplementary Fig. 3).

Copy number profiles for S2 sublines revealed a substantial amount of segmental copy number variants (CNVs) among different clades in the S2 phylogeny (Fig. 1b). The major Clades A and B have distinct patterns of CNV variation, with S2 sublines in Clade A having many CNVs, while sublines in Clade B have very few CNVs throughout their genomes (Fig. 1b). CNVs that are exclusively shared by sublines in Clade A but not present in Clade B are readily apparent, such as the ~15-Mb copy number gains and losses on chromosome arm 3L (Fig. 1b, red shading). The 2 main subclades within Clade A are also distinguished by subclade-specific CNVs: several copy number gains and losses on chromosome X, arm 2L, and arm 2R are exclusively shared by all S2R+ sublines (Fig. 1b, yellow shading), while a ~5-Mb copy number gain on chromosome arm 2L is exclusively shared by non-S2R+ sublines (Fig. 1b, red box). Within the S2R+ clade, there are also copy number losses in the distal regions of chromosome X that are exclusively shared by S2R+ sublines from the Perrimon lab (Fig. 1b, blue box). Furthermore, S2-DRSC and S2 (DSMZ-ACC-130-C) have distinct copy number profiles that differ from other S2 sublines in Clade A and Clade B (Fig. 1b), supporting the inference based on TE profiles that these are divergent S2 lineages. Finally, CNV profiles for mbn2 samples have distinct copy number profiles that differ from all other S2 sublines, consistent with the interpretation that mbn2 cells are a divergent lineage of S2. In addition, we note that the abundance and diversity of CNVs in mbn2 sublines resembles the CNV diversity observed for S2 sublines in Clade A (Fig. 1b), the major S2 clade which the mbn2 is inferred to be most closely related to based on TE profiles.

We also observed some examples where reversals of CNVs may have arisen by somatic recombination (Han *et al.* 2021) or whole-chromosome aneuploidy events (Fig. 1b, purple shading). For example, S2R+, S2R+-SQH-GFP, and most S2 sublines in Clade A (except S2-Tub-wg) share a ~5-Mb copy number loss event in chromosome arm 2L (Fig. 1b). This pattern could be explained by a segmental deletion event occurring in the common ancestor of sublines in Clade A, followed by reversals of the deletion in S2-Tub-wg and in the common ancestor of S2R+ sublines from Perrimon lab through somatic recombination (Fig. 1b). In addition, a copy-number-loss event on the entire chromosome arm 2R can be observed for S2R+-NPT005 but not for other S2R+ sublines, which can be explained by a whole-arm aneuploidy event. Overall, these results suggest that copy number changes contribute to substantial diversity in genome organization among S2 sublines and that shared patterns of CNVs are broadly consistent with the evolutionary relationships among S2 sublines inferred from TE profiles (Fig. 1a).

Evidence for ongoing transposition during long-term S2 cell culture

In the absence of secondary events such as segmental deletion, we expect ancestral nonreference TE insertions from the original fly strain or that arose during cell line establishment to be clonally inherited by all descendant sublines. Ancestral TE insertions in regions without secondary copy-number-loss events should therefore not provide any phylogenetic signal. Thus, a simple model of TE proliferation during initial cell line establishment with no subsequent genome evolution cannot jointly explain: (1) the overall increase in TE abundance and (2) the phylogenetically informative nature of TE insertions in S2 cells. Two other contrasting models can however account for both features of the TE landscape in S2 genomes. Under the "Early transposition and subsequent deletion" model (Fig. 2a), the increase in TE abundance is caused by a massive proliferation of TEs during cell line establishment, with subsequent copy-number-loss events shared by descendent cell lines indirectly explaining the phylogenetic signal of genome-wide TE profiles. Under the "Ongoing transposition in cell culture" model (Fig. 2a), it is not necessary to invoke any TE proliferation during cell line establishment, and both the overall increase in TE abundance and phylogenetic signal of TE profiles result from the ongoing accumulation of TE insertions during routine cell culture that are inherited by descendent cell lines.

These alternative models can be distinguished by analyzing TE profiles in regions of the genome without shared copy-number-loss events. In regions without shared copy-number-loss events, the "Early transposition and subsequent deletion" model predicts that TE insertions will be shared by the majority of sublines and that TE profiles will not have strong phylogenetic signal to infer the evolutionary history of S2 sublines. In contrast, the "Ongoing transposition in cell culture" model predicts that very few TEs will be shared by all sublines, and that TE profiles in regions without copy-number-loss events will be able to reconstruct evolutionary history of S2 sublines in a similar manner as genome-wide TE profiles. To test these alternative models, we analyzed TE profiles in a ~15-Mb region in chromosome X that does not include significant copy number loss across all *bona fide* S2 sublines we surveyed (Supplementary Fig. 2b, purple shading). Our analysis revealed that the majority of TE insertions in regions of the X chromosome without shared copy-number-loss events are exclusive to one or a subset of S2 subline samples (Fig. 2b). Phylogenetic analysis of nonreference TE insertions in the same region of chromosome X generated a most parsimonious tree that has the same major topological features as the one built

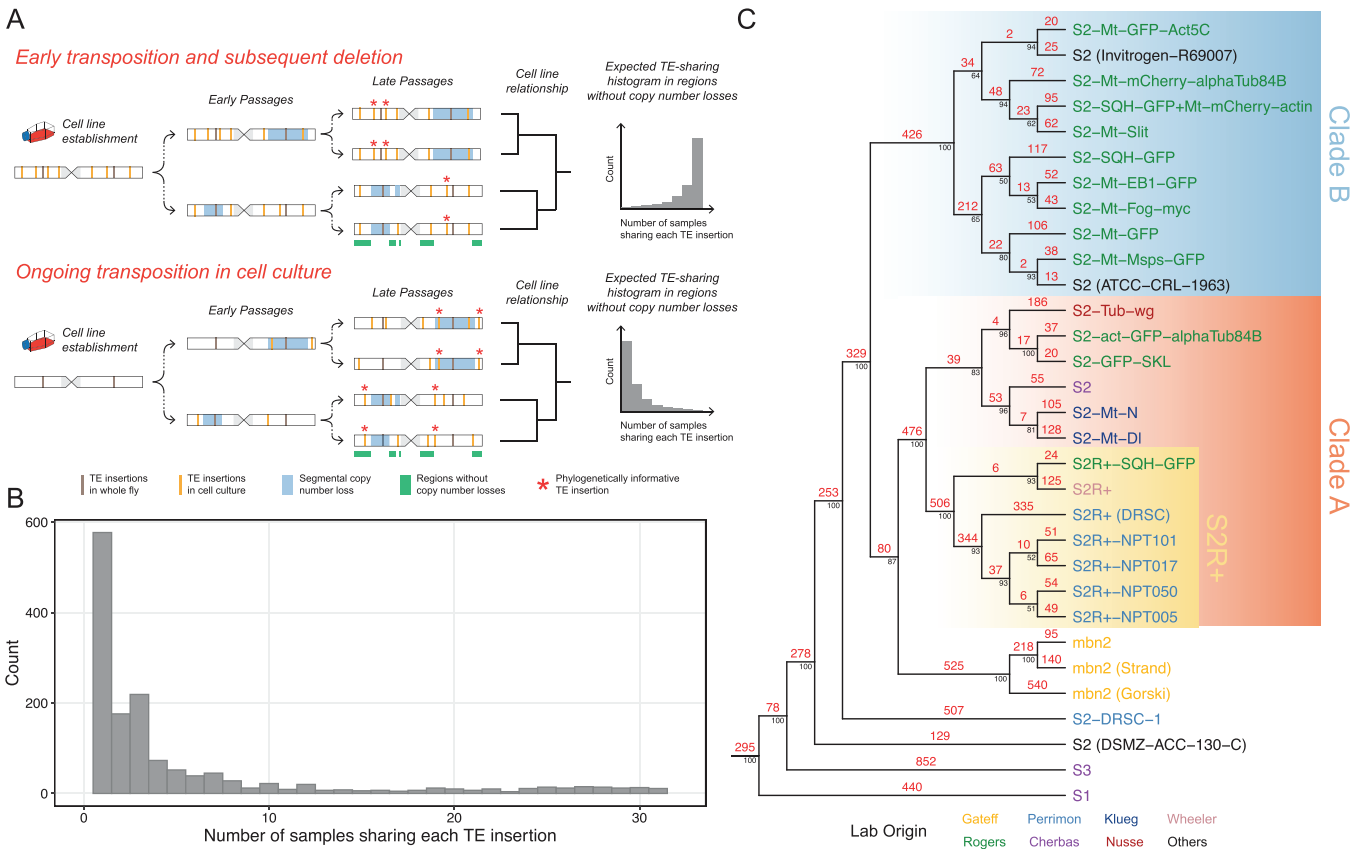

Fig. 2. TE profiles support ongoing transposition in S2 cell culture. a) Two hypotheses that could explain the mode of TE amplification in *Drosophila* S2 cell culture and how the resulting TE profiles could help infer the relationship among different cell sublines. Note that the schematic models represent genome-wide TE distributions combining all haplotypes. Therefore, given that S2 cells are tetraploid (Lee *et al.* 2014), a copy-number-loss event that occurred in one haplotype should only eliminate some TEs that are heterozygous in the affected region. b) Histogram shows the distribution of the number of *Drosophila* S2 subline samples that share each TE insertion in regions of chromosome X without major shared copy number losses (chrX: 500,000–20,928,973). c) Numbers of TE insertions on branches of the Dollo parsimony tree of 26 *Drosophila* S2 sublines constructed using nonreference TE predictions made by TEMP (Zhuang *et al.* 2014). Samples from S1, S3, and mbn2 cell lines were also included. The number of TE insertions estimated using ancestral state reconstruction was annotated in red above each branch. Percent bootstrap support was annotated in black below each node. 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 sublines were developed.

from genome-wide TE profiles (Supplementary Fig. 2a). Together, these results provide evidence against the "Early transposition and subsequent deletion" model and suggest that the genome-wide TE profiles used to infer evolutionary relationship of S2 sublines are contributed mainly by ongoing lineage-specific transposition during cell culture.

A subset of LTR retrotransposon families have episodically inserted during S2 cell line history

To gain additional insights into the dynamics of TE activity during the history of S2 cell line evolution, we mapped TE insertions on the phylogeny of *Drosophila* S2 sublines using ancestral state reconstruction based on the most parsimonious scenario of TE gain and loss under the Dollo model (Batzer and Deininger 2002; Ray *et al.* 2006; Han *et al.* 2021) (Fig. 2c). The Dollo model favors TE insertions to be gained once early in the phylogeny over parallel gains of TEs in different sublineages (Farris 1977) and is thus conservative with respect to the number of inferred transposition events on more terminal branches of the tree. The most parsimonious reconstruction of TE insertions mapped on the Schneider cell line phylogeny reveals a substantial number of TE insertions on branches at all depths in the phylogeny (Fig. 2c). For example, we observe over 250 TE insertions on each ancestral branch that split the divergent S2 lineages S2-DRSC and S2 (DSMZ-ACC-130-C) from the major S2 clades, and more than 400 TE insertions on the ancestral branches leading to both major Clades A and B. Likewise, more than 500 TE insertions are mapped on the ancestral branch leading to the S2R+ clade. This pattern of abundant insertion on most major internal branches of the phylogeny provides further support to the "Ongoing transposition in cell culture" model.

We then aggregated inferred TE insertions on each branch by TE family to visualize branch- and family-specific TE insertion profiles. This analysis revealed that only a subset of 125 recognized TE families in *D. melanogaster* contribute to the high transpositional activity in S2 cell culture (Fig. 3b; Supplementary File 3). The top 10 TE families with highest overall activities are all retrotransposons, including eight LTR retrotransposons (*blood*, *copia*, *297*, *3S18*, *1731*, *diver*, *mdg1*, and *17.6*) and two non-LTR retrotransposons (*jockey* and *Juan*). The majority of the most active TE families in S2 cells do not encode a retroviral *env* gene (8/10; 80%), with only the *297* and *17.6* Ty3/gypsy families having the potential to form infectious virus-like particles (Lerat and Capy 1999; Malik *et al.* 2000; Stefanov *et al.* 2012). This analysis also revealed that the pattern of TE family activity varies substantially on different branches of the S2 phylogeny (Fig. 3). For

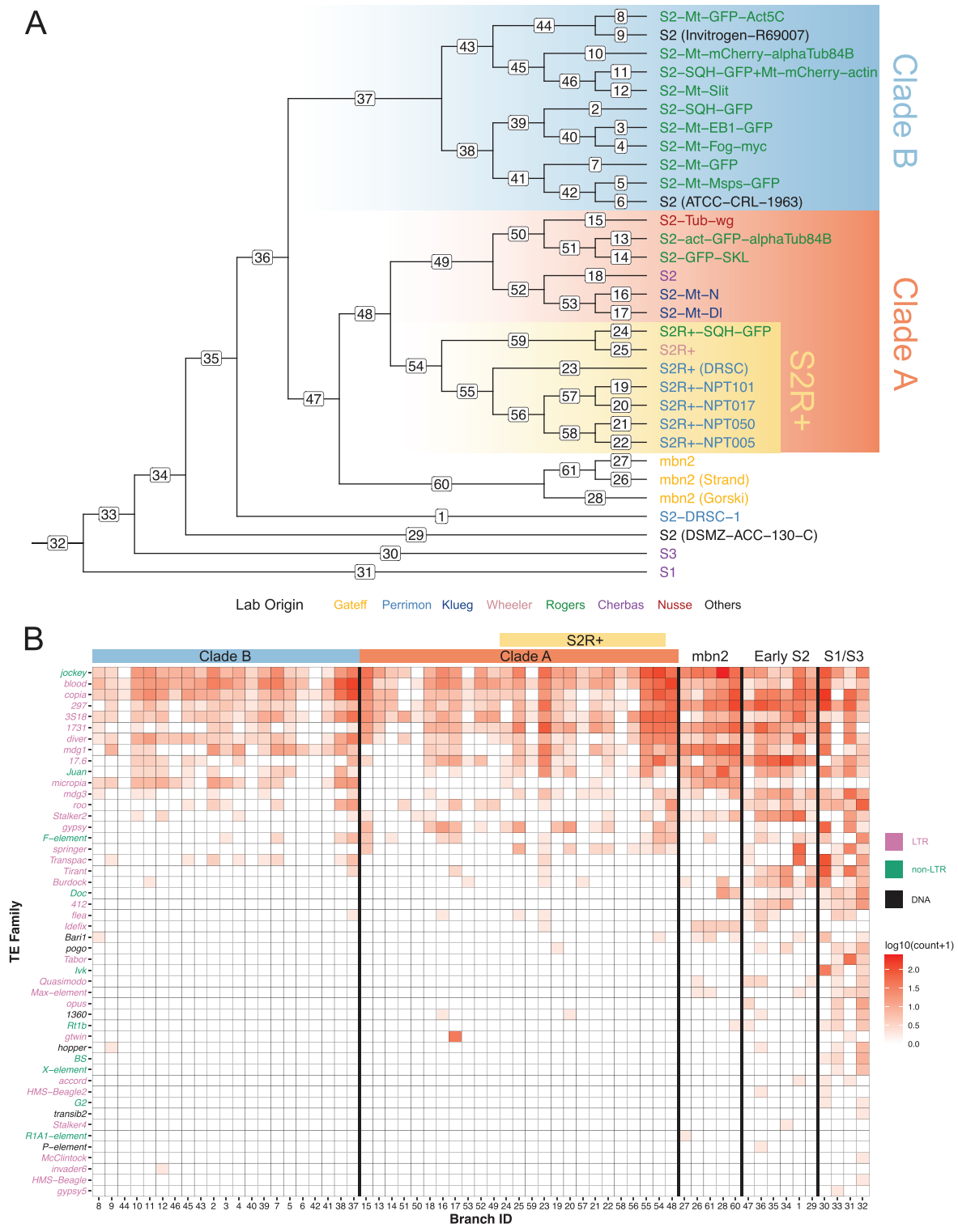

Fig. 3. Ongoing transposition in *Drosophila* S2 culture is contributed by a small subset of LTR retrotransposon families. a) Branch labelled Dollo parsimony tree including 26 *Drosophila* S2 sublines constructed using nonreference TE predictions made by TEMP (Zhuang *et al.* 2014). Samples from S1, S3, and mbn2 cell lines were also included. Taxa labels were colorized in the same way as Figs. 1 and 2c. Branch ID is annotated on each branch. b) Heatmap showing the number of estimated family-specific TE insertions on each branch of the tree in panel a. The heatmap is colorized by log-transformed [log10(count + 1)] number of gains per family per branch, sorted top to bottom by overall nonreference 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. TE family names were colorized by TE type.

example, families such as *17.6*, *297*, and *1731* have relatively high activity in branches prior to the split of Clades A and B (branch 33–36; "early S2") and in the early branches within Clade A (branch 48,49), but relatively low activity within Clade B. In contrast, families such as *jockey*, *blood*, and *3S18* have relatively low activity in "early S2" branches and relatively high activity across all branches within Clades A and B. We also observed TE family activity that is subline-specific, including the proliferation

of *gtwin* that occurred only in S2-Mt-Dl (Fig. 3), a subline of S2 that was transformed to express wild-type Delta from an inducible *metallothionein* promoter (FBtc0000152). Together, these results suggest that the increase in abundance of TEs during S2 cell culture is caused by a small subset of retrotransposon families, and that there have been episodic periods of family-specific transposition during the evolutionary history of S2 cells.

Discussion

Here, we used genome-wide TE profiles to reveal the evolutionary relationships and genomic diversity among a large panel of diverse *Drosophila* S2 sublines. Our TE-based phylogenetic analysis showed that all S2 sublines sampled form a single monophyletic clade that is an ingroup to the expected outgroup cell lines S1 and S3 (Schneider 1972; Lewerentz *et al.* 2022). This result suggests that no S2 subline in our dataset is a misidentified non-S2 *Drosophila* cell line, and implies relatively low rates of cross-contamination between S2 cells and other *Drosophila* cell lines for the sublines deposited in the DGRC by the research community. Our TE-based phylogeny also revealed two major clades of S2 sublines circulating in the research community (Clade A and Clade B), whose existence is supported by copy number profiles. Clade A includes all S2R+ sublines plus several S2 sublines, and is characterized by substantial copy number changes across the autosomes. Clade B includes only S2 sublines with mostly euploid genomes. These results imply that the "S2" subline designation is paraphyletic, and that there can be substantial genomic heterogeneity among sublines labeled as S2. We also found that some S2 sublines originating from the same lab were reconstructed in different major clades of S2, suggesting that heterogeneity in S2 genome content has the potential to influence experimental results within a single laboratory.

Our approach to clustering sublines of the same cell line using TE-based profiles has several advantages over using other types of genetic variation, especially given the fact that the phylogenetic signal that can resolve sublines of a clonallyevolving cell line is expected to be primarily haplotype-specific and therefore present as heterozygous variants. The biology of TE proliferation in *Drosophila* cell lines provides an abundant source of essentially homoplasy-free markers which can justifiably be encoded as presence/absence variants even in the face of polyploidy, segmental aneuploidy, and loss-of-heterozygosity. Furthermore, the Dollo parsimony approach can accommodate the FN predictions made by most short-read-based TE detection methods that are likely exacerbated by copy number variation across the S2 genome. In contrast, calling heterozygous SNP and small indel variants in a panel of polyploid samples with variable segmental aneuploidy is an unsolved bioinformatic challenge (Cooke *et al.* 2022), especially for intermediate coverage WGS data such as ours. Furthermore, there is no clear consensus concerning how to filter or encode heterozygous SNP variants in phylogenetic analysis (Lischer *et al.* 2014; Potts *et al.* 2014). Related challenges exist and are likely worse for other types of non-TE structural variants. Our finding that copy number profiles broadly support the TE-based phylogeny of S2 sublines suggests that the major Clades A and B we identify are not artifacts of our approach, and complements recent results showing that different types of genetic variation (SNP, TE, and local duplications) generate similar clustering of independently derived *Drosophila* cell line genomes (Lewerentz *et al.* 2022). Nevertheless, future work using other sources of genetic variation is worthwhile to cross-validate and resolve remaining uncertainties in the TE-based phylogeny of S2 sublines presented here, perhaps using extensions to methods developed for the analysis of single cell phylogenies (Kozlov *et al.* 2022).

The phylogeny of S2 sublines we infer also allows us to clarify the origin and unique phenotypes of S2R+ cells, a lineage of S2 cells whose increased adherence to tissue culture surfaces has led to its use in nearly 600 primary publications (FBtc0000150). S2R+ cells were first reported by Yanagawa *et al.* (1998) who showed that S2R+ cells are responsive to Wingless (Wg) signaling and expressed the Wg receptors Dfrizzled-1 and Dfrizzled-2, in contrast to S2 cells from the Nusse lab (presumably represented by a Clade A subline like S2-Tub-wg). Yanagawa *et al.* (1998) reported that the founding subline of the S2R+ lineage was obtained from Dr Tadashi Miyake Lab, who stated that these cells were "obtained directly from Dr. Schneider and stored frozen in his laboratory." This reported history has led the DGRC to conclude that S2R+ cells are "more similar to the original line established in the Schneider laboratory than any of the other S2 isolates in our collection" (https://dgrc.bio.indiana.edu/cells/S2Isolates; accessed 2022 May 12). In contrast to this reported history, our results place the S2R+ lineage as a derived clade inside Clade A, rather than at the base of the S2 phylogeny as would be expected if S2R+ cells were a basal lineage that reflects the original state of all S2 sublines. Furthermore, our results indicate that the increased adherence and Wg responsiveness of S2R+ cells are derived features, suggesting that they may have arisen as adaptations to propagation in cell culture. Further work will be necessary to understand the mechanisms that caused the in vitro evolution of these phenotypes, however preliminary analysis suggests that the gain of expression for Dfrizzled-1 and Dfrizzled-2 was not caused by increased copy number in the ancestor of S2R+ sublines, nor is the inferred lack of expression of these genes in other S2 isolates due to complete deletion of these loci (Supplementary Fig. 4).

Our phylogenetic hypothesis for the evolution of Schneider cell lines also allowed us to test competing models to explain the proliferation of TEs in *Drosophila* cell culture. Analysis of TE site occupancy in regions of the genome without shared copy number loss provided evidence against the "Early transposition and subsequent deletion" model while supporting the "Ongoing transposition in cell culture" model. Likewise, analysis of ancestral states provided additional evidence for the "Ongoing transposition in cell culture" model. One potential issue with our analysis of inferred TE ancestral states is the possibility of false-positive (FP) and FN nonreference TE predictions. In principle, a random FP prediction is unlikely to be shared by multiple cell samples and thus should only lead to a falsely reconstructed insertions on the terminal branches under the Dollo model. This suggests that the number of TE insertions reconstructed on the terminal branches of our trees may be overestimated. Conversely, a random FN would most likely lead to falsely reconstructed deletion on the terminal branch under the Dollo model. Thus, random FP and FN predictions should have a limited impact on our phylogenetic and ancestral state reconstruction analyses and thus not majorly affect the conclusion that there are substantial numbers of TE insertions on most internal branches of the tree, as expected under the "Ongoing transposition in cell culture" model. Furthermore, orthogonal support for the "Ongoing transposition in cell culture" model comes from a recent complementary study that found many haplotype-specific TE insertions in a S2R+ subline which occurred after initial cell line establishment and subsequent tetraploidization (Han *et al.* 2022).

Additionally, our ancestral state reconstruction analysis revealed that only a subset of TE families has high transpositional activity in S2 cell culture. Most active TE families in S2 cells

are retrotransposons that do not encode a functional retroviral *env* gene and thus are not likely to be capable of infecting another cell, suggesting that TE proliferation in *Drosophila* cell culture is mainly a cell-autonomous process. Furthermore, the fact that we do not observe activation of all TE families suggests transposition in S2 is not due to global deregulation of all TEs but is caused by some form of family- or class-specific regulation. The near-complete lack of DNA transposon activity during long-term S2 cell culture is notable in this regard, and suggests that differences in the mechanisms of RNA-based vs DNA-based transposition may provide clues to the factors regulating proliferation of specific TE families in S2 cells. Finally, our ancestral state reconstruction analysis revealed that transposition of active TE families in S2 culture is episodic. Some TE families such as *17.6*, *297*, and *1731* have relatively higher activities in the early stage of S2 evolution, while other families such as *jockey*, *blood*, and *3S18* were more active within both major clades of S2.

Our study leaves open a number of outstanding questions about the evolutionary processes governing genome evolution in *Drosophila* cell culture. More work is needed to understand the molecular mechanisms that permit TE proliferation in S2 cells and other *Drosophila* cell lines (Arkhipova *et al.* 1995). Our observation of family-specific, episodic TE activity during S2 cell line evolution favors changes in regulation of specific TE families over global relaxation of selection to explain TE proliferation in *Drosophila* cell culture. One possible mechanism to explain the family-specific, episodic TE activity in different sublines may be the variable presence of viruses, which are known to infect many *Drosophila* cell lines (Echalier 1997; Webster *et al.* 2015) and affect TE regulation in somatic tissues (Roy *et al.* 2020). Another open question is how TE insertions that arose in a single cell increase in frequency in cell culture sufficiently to be sampled and inherited by multiple sublineages of S2 cells. It is possible that some TE insertions may themselves cause adaptive mutations that cause clones carrying that TE insertion to rise in frequency. Alternatively, TE insertions could be neutral and rise in frequency by hitchhiking with adaptive mutations elsewhere in the genome, such as copy number changes in antiapoptosis or prosurvival driver genes (Lee *et al.* 2014). Increases in frequency of clones containing new TE insertions could also occur by nonadaptive events such as bottlenecks during passaging (especially for sublines that have undergone single-cell cloning) or population crashes during freeze-thaw cycles. Additionally, since we do not have information about the number of passages leading to each sample in our dataset, we cannot quantitatively relate how TE insertion or copy number changes occur as a function of evolutionary time. Thus, it is unclear if differences in the levels of genomic variability we observe among Clades A and B simply reflect the numbers of passages separating samples rather than intrinsic differences in genome stability in these clades. Future mutation accumulation experiments would be needed to estimate rates of transposition and copy number evolution in S2 cell culture and could help date the divergence time among major branches of the S2 tree. Finally, further studies on subline diversity for other *Drosophila* cell lines is needed to establish the generality of the results obtained from S2 cells, and to address the role of host genetic background on the rate and pattern of TE proliferation in *Drosophila* cell lines.

Overall, this study revealed ongoing somatic TE insertions and copy number changes as mechanisms for genome evolution in *Drosophila* S2 cell culture in the 50 years of its history since establishment (Schneider 1972). These results provide new insights into cell line genome evolution for a nonhuman metazoan species, and add to our understanding of the genomic and phenotypic heterogeneities that arise during cell culture that have been reported for the human HeLa (Liu *et al.* 2019) and MCF-7 cell lines (Ben-David *et al.* 2018). Together, these findings suggest that rapid genome evolution and subline heterogeneity are common features of animal cell lines evolving in vitro. Future work is needed to further characterize the rates and patterns of cell line genome evolution in a wider diversity of organisms to better understand how in vitro genome evolution changes affect cell line phenotypes and functional outcomes.

Data availability

Raw sequencing data generated in our study are available in the SRA under BioProject PRJNA603568. Supplementary material is available at figshare: https://doi.org/10.25386/genetics.18130898. Supplementary File 1 contains nonredundant BED files from McClintock runs using TEMP module on the dataset including 33 *Drosophila* cell line samples (reference TEs, *INE-1* insertions, and TEs in low recombination regions excluded). Supplementary File 2 contains clustered TE profiles in the format of binary presence/absence data matrix including 33 *Drosophila* cell line samples (reference TEs, *INE-1* insertions, and TEs in low recombination regions excluded). Supplementary File 3 includes data matrix of the number of nonreference TE insertion gain events per family on each branch of the most parsimonious tree used for the heatmap in Fig. 3b.

Acknowledgments

We thank the *Drosophila* Genomics Resource Center (supported by NIH grant 2P40OD010949) for cell lines, Stacey Holden and Andy Hayes (University of Manchester Genomic Technologies Core Facility) 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 and Daniel Mariyappa, Arthur Luhur, and Andrew Zelhof for comments on the manuscript.

Funding

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).

Conflicts of interest

None declared.

Literature cited

Arkhipova I, Lyubomirskaya N, Ilyin Y. Drosophila Retro transposons. Austin (TX): R.G. Landes Co; 1995.

Bairoch A. The Cellosaurus, a cell-line knowledge resource. J Biomol Tech. 2018;29(2):25–38.

Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nat Rev Genet. 2002;3(5):370–379.

Ben-David U, Siranosian B, Ha G, Tang H, Oren Y, Hinohara K, Strathdee CA, Dempster J, Lyons NJ, Burns R, *et al.* Genetic and transcriptional evolution alters cancer cell line drug response. Nature 2018;560(7718):325–330.

Bergman CM, Quesneville H, Anxolabehere D, Ashburner M. Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. Genome Biol. 2006;7(11):R112.

Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 2012;28(3):423–425.

Cherbas L, Willingham A, Zhang D, Yang L, Zou Y, Eads BD, Carlson JW, Landolin JM, Kapranov P, Dumais J, *et al.* The transcriptional diversity of 25 Drosophila cell lines. Genome Res. 2011;21(2): 301–314.

Cooke DP, Wedge DC, Lunter G. Benchmarking small-variant genotyping in polyploids. Genome Res. 2022;32(2):403–408.

Cridland JM, Macdonald SJ, Long AD, Thornton KR. Abundance and distribution of transposable elements in two Drosophila QTL mapping resources. Mol Biol Evol. 2013;30(10):2311–2327.

Czech B, Malone CD, Zhou R, Stark A, Schlingeheyde C, Dus M, Perrimon N, Kellis M, Wohlschlegel JA, Sachidanandam R, *et al.* An endogenous small interfering RNA pathway in Drosophila. Nature 2008;453(7196):798–802.

Derrien T, Estelle J, Sola SM, Knowles DG, Raineri E, Guigo R, Ribeca P. Fast computation and applications of genome mappability. PLoS One 2012;7(1):e30377.

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

Echalier G. Drosophila Cells in Culture. San Diego (CA): Academic Press; 1997.

Farris JS. Phylogenetic analysis under Dollo's law. Syst Biol. 1977; 26(1):77–88.

Gateff E, Gissmann L, Shrestha R, Plus N, Pfister H, Schroder J, Hausen H. 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). Amsterdam: Elsevier/North Holland Biomedical Press, 1980. p. 517–533.

Han S, Basting PJ, Dias GB, Luhur A, Zelhof AC, Bergman CM. Transposable element profiles reveal cell line identity and loss of heterozygosity in Drosophila cell culture. Genetics 2021;219: iyab113.

Han S, Dias GB, Basting PJ, Viswanatha R, Perrimon N, Bergman CM. Local assembly of long reads enables phylogenomics of transposable elements in a polyploid cell line. bioRxiv. 2022; https://doi.org/10.1101/2022.01.04.471818.

Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, Booth BW, Pfeiffer BD, George RA, Svirskas R, *et al.* The Release 6 reference sequence of the *Drosophila melanogaster* genome. Genome Res. 2015;25(3):445–458.

Hughes P, Marshall D, Reid Y, Parkes H, Gelber C. The costs of using unauthenticated, over-passaged cell lines: how much more data do we need? BioTechniques 2007;43(5):575–584.

Ilyin YV, Chmeliauskaite VG, Ananiev EV, Georgiev GP. Isolation and characterization of a new family of mobile dispersed genetic elements, mdg3, in *Drosophila melanogaster*. Chromosoma 1980; 81(1):27–53.

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

Kozlov A, Alves JM, Stamatakis A, Posada D. CellPhy: accurate and fast probabilistic inference of single-cell phylogenies from scDNA-seq data. Genome Biol. 2022;23(1):37.

Lammers F, Blumer M, Ruckle C, Nilsson MA. Retrophylogenomics in rorquals indicate large ancestral population sizes and a rapid radiation. Mob DNA. 2019;10:5.

Lammers F, Gallus S, Janke A, Nilsson MA. Phylogenetic conflict in bears identified by automated discovery of transposable element insertions in low-coverage genomes. Genome Biol Evol. 2017; 9(10):2862–2878.

Lee H, McManus CJ, Cho DY, Eaton M, Renda F, Somma MP, Cherbas L, May G, Powell S, Zhang D, *et al.* DNA copy number evolution in Drosophila cell lines. Genome Biol. 2014;15(8):R70.

Lee H, Oliver B. Drosophila cell lines to model selection for aneuploid states. J Down Syndrome Chromosome Abnormalities. 2015;2: 1–4.

Lerat E, Capy P. Retrotransposons and retroviruses: analysis of the envelope gene. Mol Biol Evol. 1999;16(9):1198–1207.

Lewerentz J, Johansson AM, Larsson J, Stenberg P. Transposon activity, local duplications and propagation of structural variants across haplotypes drive the evolution of the Drosophila S2 cell line. BMC Genomics 2022;23(1):276.

Lischer HE, Excoffier L, Heckel G. Ignoring heterozygous sites biases phylogenomic estimates of divergence times: implications for the evolutionary history of *Microtus* voles. Mol Biol Evol. 2014;31(4): 817–831.

Liu Y, Mi Y, Mueller T, Kreibich S, Williams EG, Van Drogen A, Borel C, Frank M, Germain PL, Bludau I, *et al.* Multi-omic measurements of heterogeneity in HeLa cells across laboratories. Nat Biotechnol. 2019;37(3):314–322.

Lynch M. The Origins of Genome Architecture. 1st ed. Sunderland (MA): Sinauer Associates Inc; 2007.

Malik HS, Henikoff S, Eickbush TH. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res. 2000;10(9):1307–1318.

Manee MM, Jackson J, Bergman CM. Conserved noncoding elements influence the transposable element landscape in Drosophila. Genome Biol Evol. 2018;10(6):1533–1545.

Mariyappa D, Rusch DB, Han S, Luhur A, Overton D, Miller DFB, Bergman CM, Zelhof AC. A novel transposable element-based authentication protocol for Drosophila cell lines. G3 (Bethesda) 2022;12:jkab403.

Nelson MG, Linheiro RS, Bergman CM. McClintock: an integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data. G3 (Bethesda) 2017;7: 2749–2762.

Neumüller RA, Wirtz-Peitz F, Lee S, Kwon Y, Buckner M, Hoskins RA, Venken KJT, Bellen HJ, Mohr SE, Perrimon N. Stringent analysis of gene function and protein–protein interactions using fluorescently tagged genes. Genetics 2012;190(3):931–940.

Platt RN, Zhang Y, Witherspoon DJ, Xing J, Suh A, Keith MS, Jorde LB, Stevens RD, Ray DA. Targeted capture of phylogenetically informative Ves SINE insertions in genus *Myotis*. Genome Biol Evol. 2015;7(6):1664–1675.

Potter SS, Brorein WJ, Dunsmuir P, Rubin GM. Transposition of elements of the 412, copia and 297 dispersed repeated gene families in *Drosophila*. Cell 1979;17(2):415–427.

Potts AJ, Hedderson TA, Grimm GW. Constructing phylogenies in the presence of intra-individual site polymorphisms (2ISPs) with a focus on the nuclear ribosomal cistron. Syst Biol. 2014;63(1):1–16.

Rahman R, Chirn G-w, Kanodia A, Sytnikova YA, Brembs B, Bergman CM, Lau NC. Unique transposon landscapes are pervasive across *Drosophila melanogaster* genomes. Nucleic Acids Res. 2015;43(22):10655–10672.

Ray DA, Xing J, Salem AH, Batzer MA. SINEs of a nearly perfect character. Syst Biol. 2006;55(6):928–935.

Rishishwar L, Marino-Ramirez L, Jordan IK. Benchmarking computational tools for polymorphic transposable element detection. Brief Bioinformatics 2017;18:908–918.

Roy M, Viginier B, Saint-Michel E, Arnaud F, Ratinier M, Fablet M. Viral infection impacts transposable element transcript amounts in Drosophila. Proc Natl Acad Sci USA. 2020;117(22):12249–12257.

Ruddle FH, Berman L, Stulberg CS. Chromosome analysis of five longterm cell culture populations derived from non-leukemic human peripheral blood (Detroit strains). Cancer Res. 1958;18(9):1048–1059.

Sackton TB, Kulathinal RJ, Bergman CM, Quinlan AR, Dopman EB, Carneiro M, Marth GT, Hartl DL, Clark AG. Population genomic inferences from sparse high-throughput sequencing of two populations of *Drosophila melanogaster*. Genome Biol Evol. 2009;1:449–465.

Salem AH, Ray DA, Xing J, Callinan PA, Myers JS, Hedges DJ, Garber RK, Witherspoon DJ, Jorde LB, Batzer MA. Alu elements and hominid phylogenetics. Proc Natl Acad Sci USA. 2003;100(22):12787–12791.

Schneider I. Cell lines derived from late embryonic stages of *Drosophila melanogaster*. J Embryol Exp Morphol. 1972;27(2):353–365.

Shedlock AM, Okada N. SINE insertions: powerful tools for molecular systematics. Bioessays 2000;22(2):148–160.

Singh ND, Petrov DA. Rapid sequence turnover at an intergenic locus in Drosophila. Mol Biol Evol. 2004;21(4):670–680.

Stefanov Y, Salenko V, Glukhov I. Drosophila errantiviruses. Mob Genet Elements. 2012;2(1):36–45.

Sukumaran J, Holder MT. DendroPy: a Python library for phylogenetic computing. Bioinformatics 2010;26(12):1569–1571.

Swofford D. PAUP: Phylogenetic Analysis Using Parsimony (and Other Methods). Sunderland (MA): Sinauer Associates; 2003.

Sytnikova YA, Rahman R, Chirn G-W, Clark JP, Lau NC. Transposable element dynamics and PIWI regulation impacts lncRNA and gene expression diversity in Drosophila ovarian cell cultures. Genome Res. 2014;24(12):1977–1990.

Vendrell-Mir P, Barteri F, Merenciano M, González J, Casacuberta JM, Castanera R. A benchmark of transposon insertion detection tools using real data. Mob DNA. 2019;10:53.

Wang J, Keightley PD, Halligan DL. 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. 2007;65(6):627–639.

Webster CL, Waldron FM, Robertson S, Crowson D, Ferrari G, Quintana JF, Brouqui J-M, Bayne EH, Longdon B, Buck AH, *et al.* The discovery, distribution, and evolution of viruses associated with *Drosophila melanogaster*. PLoS Biol. 2015;13(7):e1002210.

Wen J, Mohammed J, Bortolamiol-Becet D, Tsai H, Robine N, Westholm JO, Ladewig E, Dai Q, Okamura K, Flynt AS, *et al.* Diversity of miRNAs, siRNAs, and piRNAs across 25 Drosophila cell lines. Genome Res. 2014;24(7):1236–1250.

Xing J, Wang H, Han K, Ray DA, Huang CH, Chemnick LG, Stewart CB, Disotell TR, Ryder OA, Batzer MA. A mobile element based phylogeny of Old World monkeys. Mol Phylogenet Evol. 2005;37(3):872–880.

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

Zhuang J, Wang J, Theurkauf W, Weng Z. TEMP: a computational method for analyzing transposable element polymorphism in populations. Nucleic Acids Res. 2014;42(11):6826–6838.

Communicating editor: J. Bateman