Regulation of Gene Expression by Putatively Adaptive Transposable Element (TE) Insertions

Ellison, Christopher[1a]; Rele, Chinmay P.[1b*]

Abstract

Transposable elements (TEs) make up significant portions of eukaryotic genomes. TEs are generally considered to be genomic parasites, however recent studies in a variety of eukaryotes suggest that TEs may have a significant effect on gene expression and regulation in these species. We used *Drosophila melanogaster* in order to study the effect of TE insertions on chromatin structure as well as gene expression. In order to identify adaptive TEs involved in the regulation of gene expression, we used command-line programs to do genomic arithmetic and find the distance of these TEs to enhancer regions. We predict that at least some of the TEs studied have altered the chromatin state, chromosome conformation, and gene expression near their insertion site , but due to inefficient methods, this postulation is not yet confirmed. We hope to alleviate these drawbacks and improve the study for future work.

[1]Department of Genetics, Rutgers University, New Brunswick, New Jersey, USA
[a]Assistant Professor, Department of Genetics, Rutgers University
[b]Undergraduate Student, Department of Genetics, Rutgers University
[*]For more information or access to files, please contact chinmay.rele@rutgers.edu

Introduction

According to the Selfish DNA Hypothesis [*Dimitri &Junakovic 1999*], transposons, or transposable elements (TEs) are like selfish chromosomal parasites that "live" in and spread within genomes. They generate excess DNA to add into the genome, in which case they are known as DNA transposons [*Gonzalez & Petrov 2009*], or make RNA intermediates that are to be reverse transcribed back into DNA and then added into the genome. These types of TEs are known as retrotransposons [*Chuong et. al, 2016*].

TEs were originally studied in corn by McClintock [*McClintock, 1956*], but since, the field has moved to using many non-model and model organisms, including *Drosophila melanogaster*, to study their effect. Flies are used due to the fact that it is very easy to rear *D. melanogaster* and observe their life stages, it is very cheap, and they do not need as long a time as mice to complete a generation [*Demerec & Kaufman 1996*]. About 10% of all base sequence in *Drosophila melanogaster* consist of moderately repeated TEs [*Finnegan & Fawcett, 1986*]. TE activity has been studied in *D. melanogaster* embryos and they are often highly expressed and replicated during *Drosophila* embryogenesis [*Batut et al., 2013*].

Due to the fact that TEs insert randomly into the genome [*Ding & Lipshitz 1994*], most of them are bound to be deleterious to the organism (if they insert into a functional gene) or neutral (if they insert within the introns or intergenic regions), but sometimes advantageous (if they are inserted within a regulatory region and function as a promoter to genes that are needed in high quantities) [*Gonzalez & Petrov 2009*]. Most TE insertions are either deleterious or neutral, and thus are under purifying selection to be removed naturally from the population [*Chuong et. al, 2016*]. Advantageous insertions will spread through the population via positive selection [*Gonzalez et. al. 2008*]. *Gonzalez et. al.* (*2008*) used a population genetic approach to search for TEs that have experienced recent positive selection in *Drosophila melanogaster* population(s) from Africa. Out of 1572 TE insertions, they identified 13 that they think are regulatory and adaptive. In a separate study, *Braverman et al. (2005)* identified 20 adaptive TE insertions in *D. melanogaster* strains from North American and African populations by focusing on insertions using the non-equilibrium transposition model. Finally, *Petrov et al. (2003)* studied *D. melanogaster* strains from African and North American populations and identified 2 putatively

adaptive insertions. Together, these studies have found a total of 35 TE insertions whose population frequency and pattern of nucleotide variation are consistent with them experiencing recent positive selection.

Transposition events during development could be a potent factor affecting gene expression and evolution [*Demerec & Kaufman 1996*]. Many of the 35 adaptive *Drosophila* TEs described above use host cell machinery to transcribe themselves and replicate during embryogenesis. These transcribed TEs might be able to enhance or suppress expression of neighboring genes, which could manifest in new phenotypes [*Ding & Lipshitz, 1994*].

Most TE insertions reduce gene expression as the host tries to silence them by depositing a suppressive histone modification [*Hollister et al., 2009*]. These modifications can spread into nearby genes and in turn reduce their expression. An alternate way that TEs can affect nearby gene expression is by carrying regulatory sequences that either change the time or spatial pattern of expression [*Cavalieri et al., 2008*]. This may also be the case for genes that are not adjacent or close to the TE insertion in question. Such regulatory sequences are known as enhancers and, in theory, these TE insertions have a much higher likelihood of producing a novel phenotype and thus a greater likelihood of being adaptive [*Mateo et al., 2014*]. There are two possible ways to examine whether putatively adaptive TEs act as enhancers, (1) if they are associated with a histone modification that is known to mark enhancer regions of euchromatin, or (2) if the chromosomal conformation is organized in such a way that the TE is near the promoter of a distant gene.

Chromatin states represent the epigenetic status of a certain region of the chromosome and are defined by Chromatin-Associated Proteins (CAPs) and post-translational modifications (PTMs) [Baker, 2011]. Histone modifications such as methylation, ubiquitination, phosphorylation, acetylation and sumoylation are common PTMs. Histone PTMs can affect gene expression by recruiting histone modifiers or altering chromosome structure in order to expose or conceal sequences [Bannister & Kouzarides, 2011]. The H3K4me1 histone modification has been previously shown to be a useful marker of enhancers [*Bonn et al., 2012*].

If we know the chromatin state of a particular region, we can tell whether that region is being expressed at a particular time during the cell cycle. Based on the time

that the chromosome state was measured, the types of genes expressed at the time can be estimated [*Alberts et al., 2002*].

Chromosome conformation capture-on-chip (4C) allows the inference of inter- and intra-chromosomal interactions between a targeted region and the rest of the chromosome [*Simonis et. al, 2006*]. Earlier studies assume that TE insertions only affect expression of genes closest to them on the chromosome, but they do not account for the fact that the chromosomes are folded in 3D space. Gene loci may have a 3D arrangement that differs from their linear organization on the 2D chromosome due to folding of chromatin.

The field still doesn't know why and how TEs affect transcription of other genes. We hypothesize that the TEs compiled by *Barrón et al. (2014)* (from [*Braverman et al., 2005*] [*González et al., 2008*] [*Petrov et al., 2003*]) are adaptive because they modify embryonic gene expression. Here we have used three complementary approaches to address this hypothesis: (1) we used previously published ChIP-seq data to identify genomic regions with the H3K4me1 histone modification in Drosophila embryos and determined whether adaptive TE insertions tend to reside near these marks; (2) we assessed whether the Drosophila Genetic Reference Panel (DGRP) population will be a useful resource for studying chromosome conformation in a genomic region with and without an adaptive TE insertion by determining the frequencies of these adaptive TEs in the population; and (3) we developed a modified 4C protocol based on the UMI-4C approach to assess chromosome conformation at these TE insertion sites.

We show that several of the adaptive TEs are likely to act as enhancers based on the histone modification data and that the frequencies of many of the adaptive TEs in the DGRP population are ideal for performing the 4C experiment on individuals with and without a given TE insertion. Finally, we have developed a modified 4C protocol and identified specific steps that need to be optimized in order to produce a high-concentration sequencing library.

These results will help us understand the mechanism by which TEs influence gene expression, and determine whether they can modify expression of genes that are near or far away from their insertion site.

Results

R.1 Candidate Enhancer TEs

As mentioned previously, one way that a TE insertion might result in an adaptive phenotype is by modifying the spatial and temporal pattern of gene expression. Such an effect is produced by regulatory sequences known as enhancers and we sought to determine whether any of the TE insertions that were previously identified as being adaptive, contained histone modifications consistent with them being enhancers.

To test this, we used H3K4me1 expressed sites from *Bonn et al. (2012)* and calculated the distance between each adaptive TE and the nearest H3K4me1 histone PTM enhancer region. Out of the 35 TE insertions, only 17 of them were within 5000 basepairs of a H3K4me1 histone PTM enhancer region. This means that these 17 TE insertions are much more likely to act as enhancers.

R.2 Frequency of Putatively Adaptive TEs in DGRP

A signature for putatively adaptive TEs is a high frequency in the population. Consequently, those that are less adaptive would have a lower frequency in the population. The association of a specific chromosomal conformation or interaction to the presence or absence of a TE means that there is some correlation between the TE and that interaction. To study this, we needed strains that have a particular TE and those that do not. In order to test the frequencies of putatively adaptive TEs as compiled by *Barrón et al. (2014)* in the DGRP strains provided by *Rahman et al. (2015)*, we had to run intersectBed with multiple parameters.

The files for DGRP were only for deletions, i.e., they only showed the TEs that have been excised or were never present in the first place. Thus, we had to invert the frequencies by subtracting them from 1. The data received is in {*Appendix 04*}. A visual representation of the data is illustrated in {*Fig02.*}. There were some TEs that were still present in a large number of strains, which also reflect the TE matches in the two strains studied. There were 4 TEs that were present (not depleted) from all the DGRP strains.

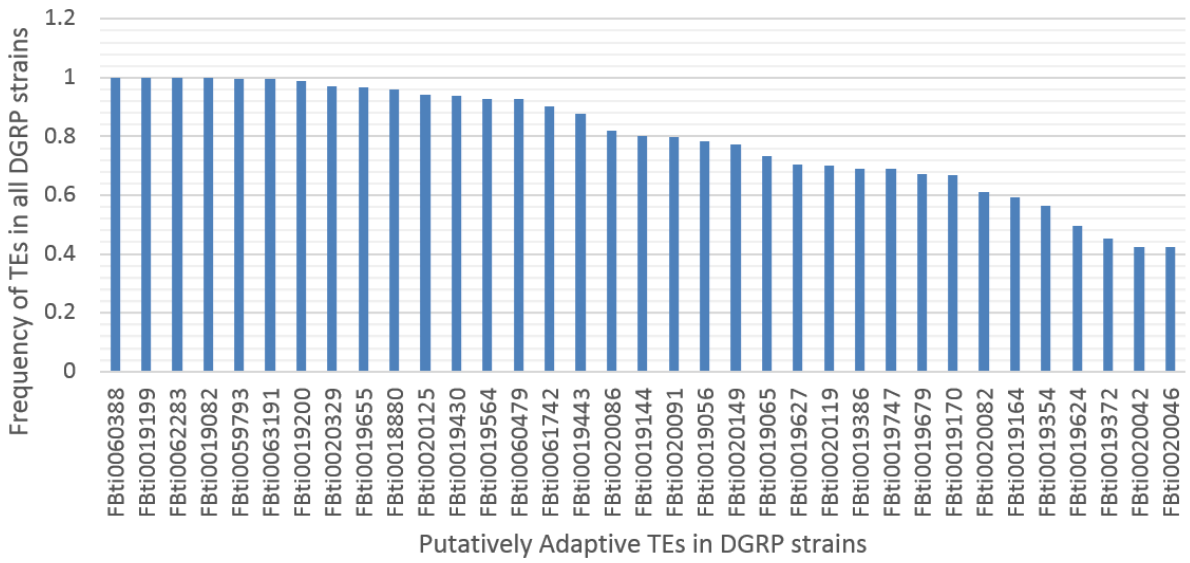


Fig02. Frequency of Putatively Adaptive TEs in all DGRP strains.
A visual representation of the frequencies of studied Putatively adaptive TEs in all DGRP strains from TIDAL fly. The range from 1.0 (FBti0060388) to 0.423729 (FBti0020046). Quantitative frequencies for each TE are given in {*Appendix 04*}.

Of the two strains from DGRP that were analyzed (25203 and 25819) , only 8 TEs were present in one strain and absent in the other {*Appendix 07*}.

R.3 4C Protocol

We need to know what the presence or absence of each of these putatively adaptive TEs mean in terms of chromosomal interactions. In order to do this, we used 4C, which is an effective assay to determine the proximity of the TE to a locus on the chromosome in a 3D spatial arrangement. 4C is a procedure where the chromosome conformation is fixed and then sequenced to see which parts of the chromosome interact or are near other parts, thus exerting some sort of effect, either in increased or decreased expression of that region.

Each step of the 4C protocol as well as effective sequencing relied on having a high amount of DNA and maintaining of this amount during the assay. This high amount of DNA was measured via the concentration of the sample as well as its mass.

The current 4C protocol used had some issues with it. For example, the amount of dsDNA that was retrieved at every step seemed to be diminishing exponentially [*Table 01*].

	Concentration	Volume	Total dsDNA
Post-Digestion	6.3 ng/μL	558 μL	3,704 ng
Post-Ligation	35 ng/μL	50 μL	1,750 ng
Post-Sonication	2.53 ng/μL	45 μL	113.85 ng

Table 01. Total weight of dsDNA after key steps in 4C Protocol
Clearly see that though careful measures have been taken, the concentration of dsDNA has gone down after key steps in the UMI-4C protocol by Schwartzman et al. (2016).

The most drastic of this decrease in amount of dsDNA is post-sonication of the sample.

Discussion

We were able to design 4C primers for about half of the putatively adaptive TEs mentioned by *Barrón et al. (2014)*. This was due to a combination of not having a DpnII site close (<250 bp) to the TE, another TE being relatively close to it and thus possibly confounding our results, some of the designed primer sets hybridizing multiple times to the genome and thus not giving us specificity, and finally the distance of primers that only hybridized once to either the TE or the DpnII site. This brought the shortlist from 35 insertions to 18, of which we only chose to use 9 in order to validate the protocol. The 9 that we chose not to use are still good candidates for future work. We prioritized these 9 TE insertions from 18 due to their proximity to an H3K4me1 PTM site. TEs that are closer to this PTM have a much higher likelihood of acting as direct enhancers. These TEs are likely to influence gene expression. These primers are on a priority list for future 4C work that we will do.

Our results suggest that there may be a problem with our modified 4C protocol. Due to the rapid decline in the amount of dsDNA post-sonication, we assume that we lost most of the sample during or after that step. According to a gel we ran post-ligation, the size of the DNA seemed to hover above the *1kb Plus DNA Ladder*, which would put it at about 13,000+ bp. Thus, we assumed that due to this size, as well as the absence of proteins holding the dsDNA together, it would behave much like genomic DNA (gDNA). Thus, we used the sonication protocol and *Covaris S2* settings for gDNA.

However, in order to confirm the size of the fragments post-sonication, we ideally should have run a gel. However, we didn't run the gel and assumed that it cut the dsDNA in fragments of ~0.4–0.5 kbp. We carried on with the 4C protocol, and the AmpureXP beads were supposed to bind dsDNA that were > 400 basepairs. However, due to the possibly incorrect sonication parameters, we were unable to salvage the amount of dsDNA we needed in order to sequence it. During the regular bead cleanup to move the dsDNA onto the beads and then into the eluent (EB), we likely lost DNA that was less than 400 basepairs. Ideally, we should have run the gel at this point (post end repair) to estimate the size of the dsDNA fragments after sonication.

Future Directions

From the experiments we conducted, there were three outstanding issues that have yet to be resolved.

The first being the low sample concentration post-sonication. This could possibly be fixed by running a time series on the sonicator by first sonicating the sample for a set time, removing an aliquot, sonicating some more, taking another aliquot etc… We will also run a gel for these to see which aliquot gives us the desired strand length. Also, after using those settings, we will run a small sample on a gel to confirm that the DNA is at the correct concentration and length. The second issue was that of only being able to design primers effectively for 18 TEs out of 35. This was primarily restricted due to the fact that the DpnII site was mentioned in the protocol. We will try to use another RE, possibly EcoRI or BamHI, or a combination of REs that work with individual TEs. We might also test using DNase and MNase to digest the chromatin, which have less defined cut sites, meaning that we could potentially use any primer location for these enzymes.

The last issue was finding adaptive TE insertions which are present in one strain and absent in another {*Appendix 07*}. In order to add more columns into this and find out exactly which TEs are present and absent in each of the strains, we plan to run the same algorithm we did for 25203 and 25189 with all of the rest by compiling the databases and then running them together. With this, we can then estimate which TE is present in about half of the strains and then study it to see how much its effect on chromatin state in that strain differs from other strains.

Methods

M.01 Identification of TE and primers

First, each TE compiled by *Barrón et. al (2017)* was searched in FlyBase, coordinates and closest DpnII site noted down in {*Appendix 01*} along with chromosome, starts and stop coordinate, closest external DpnII site, distances to and locations of closest upstream and downstream TEs and 5′ and/or 3′ primers designed for each TE that we chose to study. Each primary DpnII site was observed to be no more than 1kb upstream or downstream of the TE insertion. The main criteria for primers that were to be synthesized were (1) distance to the closest DpnII site should be no more than 150-200 bp upstream or downstream from the 5′ or 3′ end of the TE respectively; (2) if the DpnII site closest to the 5′ end of the TE is within the TE, then it must be no more than 25 bp downstream from the 5′ end; (3) two non-overlapping primers must be able to be synthesized with the parameters covered in '*M.2 Primer Synthesis*'; (4) the primers must be unique and not hybridize multiple times to the genome; and (5) the distance of the closest primer must be ~50-75 bp away from the DpnII site and <200 bp away from the TE.

M.02 Primer synthesis

All primer sequences were designed using GBrowse on flybase.org {*http://flybase.org/cgi-bin/gbrowse2/dmel/*} and the PrimerQuest tool {*https://eu.idtdna.com/Primerquest/Home/Index*}. Primers were designed according to their distance from the closest DpnII site (primary DpnII) from the 5′ end of the TE. All sequences mentioned below were taken using GBrowse, dumped as FASTA and then input in the PrimerQuest tool.

If the primary DpnII site was upstream of the 5′ end of the TE, then two types of primers were possible, Forward (F) and Reverse (R) primers. Only one set (i.e. either F primers or R primers, but not both) were required for 4C. Both sets were designed so that their locations could be compared and the optimal set selected for the experiment. Fs were designed from input sequence consisting of either: 1) the 200 bp sequence fragment upstream of the primary DpnII site, or 2) all the bases between

the primary DpnII site and another DpnII site upstream of the TE. Design parameters were set for 'Forward Only – Sequencing' in PrimerQuest and the *variation of primer location* parameter was set to high. The minimum primer T_m was set to 55°C and the minimum GC content was set to 30%. All other parameters were left as defaults. Two non-overlapping primers were chosen. The downstream (D) primer was closer to the DpnII site and provided specificity using nested PCR [*Boon et al., 2002*] [*Lee, 1994*], Rs were made by selecting the design parameter as 'Reverse Only – Sequencing', and all the other parameters were left the same as for Fs {Fig01.A}.

For TEs where the closest DpnII site was within the TE, no Rs could be made as they would be within the TE. For primers directly upstream of the TE all the same parameters were selected as mentioned above for Fs {Fig1.B}.

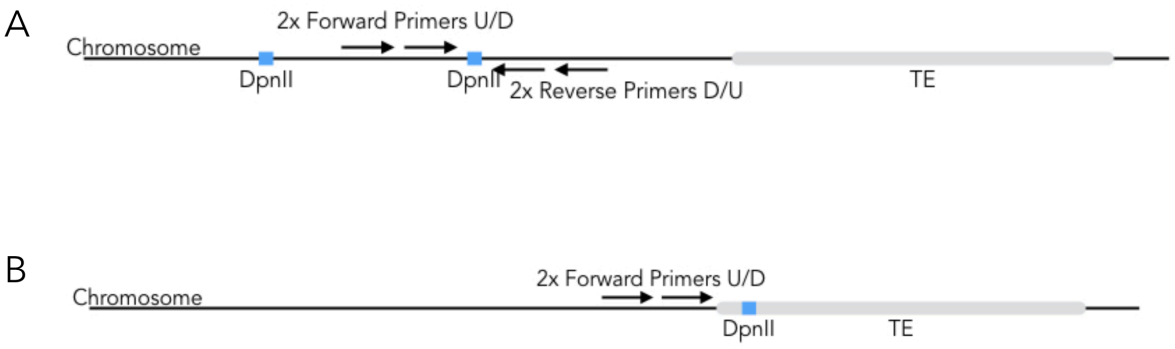


Fig01. Selection of primer locations based on distance of closest DpnII site to 5' end of TE
(A) If the DpnII site was upstream of the 5' end of the TE, then selection of forward and reverse primers was done where at least 200 bp could be used without interference of the TE or another DpnII site. (B) If the DpnII site was located within the TE, then only a forward primer was made that ended a few bases upstream of the 5' TE end.

M.03 Evaluation of primers

Unique primers were sought after as it would mean more specificity towards TE selection and extraction. This was done using Bowtie v1.2 and a maximum of 2 mismatches per read [`-v 2 -a`] on macOS Sierra v10.12.3 (16D32). A file of primer sequences was saved as a FASTA file and aligned to *Drosophila melanogaster* genome Dmel R6.14. Using the Bowtie output file *BOWTIE_OUTPUT*, the number

of alignments was calculated for each primer on the Unix command line: [`cut -f 1 *BOWTIE_OUTPUT* | sort | uniq -c`].

M.04 Embryo Collection

In order to make embryo plates, a mixture of molasses, food agar, water and 10% Tegosept solution in EtOH was made by vigorously boiling the mixture multiple times and cooling before use. Yeast paste was added to the plates so that the adult flies could feed. The recipe for the plates and the yeast paste is detailed in {*Appendix 02*}.

A rough contraption of a wider sieve and finer membrane was made where the wider sieve was placed above the smaller-holed membrane. Flies were anesthetized using CO_2, and the embryo collection plate was removed from the cage. 1x PBS was added to the plate and was agitated with a brush. The mixture of embryos, grown flies and yeast was placed in the sieves and further washed with PBS [*Rothwell and Sullivan, 2007*]. The embryos were dechorionated by washing in bleach for 1.5 minutes and followed by 2 water baths to wash out the bleach. They were weighed and then frozen using dry ice [*Kiehart et al., 2007*]. They were further stored in a freezer at -80 °C.

M.05 Nuclei Preparation

Special care was taken during preparation of nuclei as formaldehyde is a mutagen, heptane is highly flammable and phenylmethane sulfonyl fluoride (PMSF) is toxic and severely corrosive to oculi.

Membranes with ~1.5 grams of frozen embryos were retrieved, which were then quickly transferred to a 50 mL conical tube with 10 mL cross linking solution and 30 mL of heptane. The embryos were shaken off and the membrane was recovered. The tube was vigorously shaken using a vortex at RT for 15 mins.

The embryos were pelleted via centrifugation in a 4 °C centrifuge at 500 G for a minute, the supernatant was replaced with 30 mL stop solution and shaken vigorously for a minute to stop cross-linking. The embryos were pelleted again at 500 G for another minute. The supernatant was carefully decanted and the pellet was washed with 50 mL of PBT. The solution was washed again and then resuspended in 15 mL

cold PBT with protease inhibitors and 1 mM of PMSF. The suspension was then transferred to a 15 mL Dounce homogenizer.

It was homogenized using a loose-fitting pestle on ice. The lysate was transferred to a 15 mL conical tube and centrifuged at 400 G for 1 minute at 4 °C in order to precipitate the vitelline membrane and other large debris. The supernatant was decanted into fresh tube and was spun at 1,100 G for 10 mins at 4 °C. The cell pellet formed was resuspended in 15 mL of cold cell lysis buffer and supplemented with protease inhibitors and 1 mM PMSF.

The cell pellet was homogenized again using 20 strokes of the homogenizer and a tight-fitting pestle in the homogenizer on ice. The sample was then transferred to 10 1.5 mL LoBind tubes. The samples were then centrifuged at 2,000 G for 4 mins at 4 °C to pellet the nuclei. The supernatant was discarded, the tubes were flash-frozen and then stored at -80 °C. A detailed recipe for solvents is given in {*Appendix 08*}.

M.06 DpnII 4C Protocol

After applying TE selection criteria [*M.1 Identification of TEs and Primers*], we could only design primers for a few TEs. Even for the TEs were primers should have been designed for, there was conflicts with the primer design criteria [*M.2 Primer Synthesis*]. Out of all the 35 TEs that were intended to be studied, we could only design effective primers for 18 of the,

19 out of these were 250 bp away from the closest DpnII site and were kept in order to further study them. Out of the remaining 16, 1 was included in the assay due to an error and 2 more were included as they were very close to a histone PTM that enables sequences to act as enhancers. Out of this total of 22 TEs that were to be included in the assay, 2 had a combination of the primers hybridizing multiple times to the chromosome, or the complete inability to synthesize primers according to the synthesis criteria mentioned in [*M.2 Primer Synthesis*]. Two of these 20 TEs that meet all the criteria mentioned above (FBti0062283 and FBti0019082) were removed as primers could not be synthesized for them.

This left us with 18 TEs out of which only 9 were chosen to continue with the assay as they distance between the primers and both the TE and the DpnII site were small. This was done in order prioritize those that were near a H3K4me1 histone PTM and

likely to act as enhancers. TEs whose primers were near both, the TE as well as the DpnII site were chosen so that there would be less error in Illumina sequencing of the short segments. TE insertions near other transposons were discarded as there would be a non-integral effect on the chromatin state near the insertion. They would have cumulatively played a role and be affected by the H3K4me1 histone PTMs.

According to the 4C protocol outlined by Simonis et. al. (2006) and Schwartzman et. al. (2016), the DpnII restriction endonuclease was used as it had a fairly frequent restriction site.

The nuclei pellet was first thawed on ice. Then 450 µL of *UltraPure* H_2O was added along with 60 µL of 10x DpnII Reaction Reaction Buffer (#B0543S) from *NEB* and 15 µL of 10% SDS. Mixture was incubated at 37°C for 60 mins at 900 rpm. After this, 75 µL of 20% TX-100 was added and incubated again at 37°C for 60 mins at 900 rpm.

The chromatin was then digested using HC DpnII. For the first digest, 200 units of HC DpnII from *NEB* was added and left in the thermomixer at 37°C for 120 mins. For the second digest, 200U of HC DpnII from *NEB* was added and left in the thermomixer overnight. The third digest was the same as the first one.

For the ligation step of the UMI-4C protocol [*Schwartzman et. al., 2016*], the tube from the digestion was centrifuged for 60s at 2,500 G at room temperature. The supernatant was discarded and the pellet was resuspended in 150 µL of *UltraPure* H_2O. 300 µL of AMPure XP beads were added and mixed by pipetting. The solution was left to incubate at RT for 5 mins and then placed on a magnet for 2 mins. The supernatant was discarded and the beads were washed twice with 80% EtOH. The beads were then spun at 500g for 10s and the residual EtOH was removed and then air-dried for 2 mins. This will be called a bead cleanup. The beads were then resuspended in 87 µL of *UltraPure* H_2O, after which 10 µL of 10x T4 ligase buffer, 2.5 µL of 20% TX-100, and 200U of *NEB T4 DNA ligase* was added to the tube. This solution was gently mixed and left on the ThermoMixer at 16 °C overnight without mixing.

The pellet mixture was then centrifuged for 60s at 2,500G at RT. The supernatant was removed and the pellet was resuspended in 400 µL of 1x *NEBuffer 2*. 40 µL of 10% SDS and RNase A each were added to the mixture and it was incubated for 45 mins. After incubation, 40 µL of proteinase K (20mg/mL) was added and then

incubated for an additional 7 hrs at 60 °C at 900 rpm. After which, 3 µL of GlycoBlue, 50 µL of 3M Sodium Acetate (pH = 5.2) and 550 µL of isopropanol were added. The solution was mixed by inverting ~20 times and then incubated overnight at -80 °C.

The next day, the mixture was centrifuged for 30 mins at 4 °C at max speed. The supernatant was removed and the pellet was resuspended in 100 µL of nuclease free water. An additional 100 µL of magnetic beads were added and mixed well by vortexing. The mixture was incubated at RT for 5 mins and then placed on a magnetic stand for 2 mins. A bead cleanup was done and the beads were then resuspended in 50 µL of EB. Beads were then incubated at RT for 5 mins, collected via a magnet and then the eluent was transferred to a fresh 1.5 mL LoBind tube.

Before PCR could be done, we needed to fragment the ligated the DNA, repair the fragment ends and ligate on sequence adapters. In order to sonicate the sample, it was first transferred to a *Covaris™* tube and put in a *Covaris™ Sonicator S2* with the following settings: intensity (5); duty cycle (5%); cycles per burst (200); time (35s). 42.5 µL of this sonicated sample were transferred to a 1.5mL LoBind tube, and 5 µL of *NEBNext End Repair Reaction Buffer* (10x) and 2.5 µL of *NEBNext End Repair Enzyme Mix* were added. The mixture was pipetted to mix, and incubated at 20 °C for 30 mins. 110 µL of magnetic beads were mixed in and the mixture as left to incubate for 10 mins at RT.

Another round of bead cleanup was done and then were resuspended in 42.5 µL of EB. To 41.5 µL of the sample, 5 µL of NEBuffer 2 (10x), 0.5 µL dATP (10mM) and 3 µL Klenov Fragment (3′ → 5′ exo-) (NEB M0212L) were added and incubated at 37 °C for 20 mins. 2.5 µL of Quick CIP (M0508S) was added and the mixture was incubated for 10 mins at 37 °C. 102 µL of beads were added and incubated for 10 mins. A bead cleanup was done and the beads were resuspended in 23 µL of EB. They were mixed and the mixture was incubated at RT for 2 mins. Tube was placed on a magnetic stand for 2 mins and 22 µL of eluent was moved to a new tube. It was then kept in a -20 °C freezer overnight.

To this mixture, 25 µL of 2xLigation buffer, 0.5 µL of Indexed Adapter 4 {*Appendix 06*} and 2.5 µL of NEB Ligase M2200L was added and incubated at 25 °C for 15 mins. The mixture was then transferred to a PCR tube and incubated at 95 °C for 2 mins. 60 µL of beads were added to the PCR tube and incubated at RT for 10 mins. 0.5 mL of EtOH was used for a bead cleanup. The beads were then resuspended in 47 µL of EB

and left to incubate at RT for 2 mins. The tube was placed on the magnetic stand for 4 mins and 45 µL of the eluent was transferred to a LoBind tube and stored at -20 °C.

M.07 Gel Preparation

A 1% agarose gel was prepared by adding 0.35g of Agarose to 35 mL of 1x TAE Buffer, after which 3.5 µL of SYBR safe was added. It was run using a 12 µL aliquot of the third digest. To this aliquot, 2 µL of 20mg/mL of Proteinase K was added and the mixture was incubated at 65°C for 30 mins. The DNA was purified using Qiaquick PCR purification kit and 5 µL of this was run on the gel.

A gel was run twice to estimate the size of the fragments after certain steps in the protocol like post-digestion and post-sonication.

M.08 Concentration Check

The concentration was checked by taking 1 µL of the sample, 99 µL of 1x TE Buffer and 100 µL of 1x dsDNA dye provided by Promega with its Quantus™ Fluorometer.

M.09 DGRP analysis

We needed to know how many of the putatively adaptive TEs were absent or depleted in the DGRP wild strains. To do this, we used the intersect and groupby operators for bedtools or intersectBed and groupBy. The shell for the script run in Mac Terminal Version 2.7.2 (388.1) to get the retention frequencies of the TEs in the DGRP flies is given in {*Appendix 03*}; the file 'TE_to_DGRP_freq_retention.txt' attained at the end of the analysis is given in {*Appendix 04*}. For this, all multiple datasets within the DGRP website were combined to a single file {all_DGRP_sorted.bed *Appendix 05*}. 177 was the total number of unique RAL strains as provided by TIDAL-FLY.

M.10 Enhancer Analysis (H3K4me1)

We needed to know which of our putatively adaptive TEs were near enhancer regions and could possibly augment gene expression. To do this, we used the closest operator for bedtools or closestBed. H3K4me1 is a PTM that is associated with enhancer sequences. [*Bonn et al., 2012*]. The shell for the script to get the distance between these TEs and the closest H3K4me1 PTM is given in {*Appendix 05*}.

Acknowledgements

We would like to thank Dr. Dibyendu Kumar and his associates for sonication as well as other fragment size analysis procedures.

We would also like to specially thank Weihuan Cao for constant input with improving procedures, preparing embryos, extracting nuclei and generally rearing the fly strains that have been used for the study.

References

Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. Chromosomal DNA and Its Packaging in the Chromatin Fiber. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26834/

Aminetzach, Y. T., Macpherson, J. M., & Petrov, D. A. (2005). Pesticide Resistance via Transposition-Mediated Adaptive Gene Truncation in Drosophila. Science, 309(5735), 764–767. https://doi.org/10.1126/science.1112699

Baker, M. (2011). Making sense of chromatin states. Nat Methods, 8(9), 717–722. https://doi.org/10.1038/nmeth.1673

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Research, 21(3), 381–395. https://doi.org/10.1038/cr.2011.22

Barrón, M. G., Fiston-Lavier, A.-S., Petrov, D. A., & González, J. (2014). Population Genomics of Transposable Elements in Drosophila. Annual Review of Genetics, 48(1), 561–581. https://doi.org/10.1146/annurev-genet-120213-092359

Batut, P., Dobin, A., Plessy, C., Carninci, P., & Gingeras, T. R. (2013). High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. Genome Research, 23(1), 169–180. https://doi.org/10.1101/gr.139618.112

Blumenstiel, J. P., Chen, X., He, M., & Bergman, C. M. (2014). An Age-of-Allele Test of Neutrality for Transposable Element Insertions. Genetics, 196(2), 523–538. https://doi.org/10.1534/genetics.113.158147

Bonev Boyan, & Cavalli Giacomo. (2016). Organization and function of the 3D genome. Nature Review Genetics, 17, 661–678. https://doi.org/10.1038/nrg.2016.112

Boon, N., Windt, W., Verstraete, W., & Top, E. M. (2002). Evaluation of nested PCR-DGGE (denaturing gradient gel electrophoresis) with group-specific 16S rRNA primers for the analysis of bacterial communities from different wastewater treatment plants. FEMS Microbiology Ecology, 39(2), 101–112. https://doi.org/10.1111/j.1574-6941.2002.tb00911.x

Bonn, S., Zinzen, R. P., Girardot, C., Gustafson, E. H., Perez-Gonzalez, A., Delhomme, N., … Furlong, E. E. M. (2012). Tissue-specific analysis of chromatin state identifies

temporal signatures of enhancer activity during embryonic development. Nature Genetics, 44(2), 148–156. https://doi.org/10.1038/ng.1064

Braverman, J. M., Lazzaro, B. P., Aguadé, M., & Langley, C. H. (2005). DNA Sequence Polymorphism and Divergence at the erect wing and suppressor of sable Loci of Drosophila melanogaster and D. simulans. Genetics, 170(3). Retrieved from http://www.genetics.org/content/170/3/1153

Cavalieri, V., Di Bernardo, M., Anello, L., & Spinelli, G. (2008). cis-Regulatory sequences driving the expression of the Hbox12 homeobox-containing gene in the presumptive aboral ectoderm territory of the Paracentrotus lividus sea urchin embryo. Developmental Biology, 321(2), 455–469. https://doi.org/10.1016/j.ydbio.2008.06.006

Chuong, E. B., Elde, N. C., & Feschotte, C. (2016). Regulatory activities of transposable elements: from conflicts to benefits. Nature Reviews. Genetics, advance on(2), 71–86. https://doi.org/10.1038/nrg.2016.139

Demerec, M., and B. P. Kaufman. *Drosophila Guide: Introduction to the Genetics and Cytology of Drosophila melanogaster*. 10th ed. Washington D.C.: Carnegie Institution of Washnigton, 1996. 1-28. Print.

Dimitri, P. (1999). Revising the selfish {DNA} hypothesis new evidence on accumulation of transposable elements in heterochromatin The bulk of the eukaryotic genome is composed of families of repetitive sequences that are genetically silent. Trends in Genetics, 15(4), 123–124. https://doi.org/10.1016/S0168-9525(99)01711-4

Ding, D., & Lipshitz, H. (1994). Spatially regulated expression of retroviruses-like transposons during Drosophila melanogaster embryogenesis. Genet. Res., 64(1994), 167.

Feschotte, C. (2008). The contribution of transposable elements ot the evolution of regulatory networks. Nature Reviews Genetics, 9(5), 397–405. https://doi.org/10.1038/nrg2337.The

Finnegan, D. J., & Fawcett, D. H. (1986). Transposable elements in Drosophila melanogaster. Oxford Surveys on Eukaryotic Genes, 3, 1–62. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2855303

George, J. A., DeBaryshe, P. G., Traverse, K. L., Celniker, S. E., & Pardue, M. Lou. (2006). Genomic organization of the Drosophila telomere retrotransposable elements. Genome Research, 16(10), 1231–1240. https://doi.org/10.1101/gr.5348806

González, J., & Petrov, D. A. (2009). The adaptive role of transposable elements in the Drosophila genome. Gene, 448(2), 124–133. https://doi.org/10.1016/j.gene.2009.06.008

González, J., Karasov, T. L., Messer, P. W., Petrov, D. A., Lemaitre, B., & Chambers, K. (2010). Genome-Wide Patterns of Adaptation to Temperate Environments Associated with Transposable Elements in Drosophila. PLoS Genetics, 6(4), e1000905. https://doi.org/10.1371/journal.pgen.1000905

González, J., Lenkov, K., Lipatov, M., Macpherson, J. M., & Petrov, D. A. (2008). High rate of recent transposable element-induced adaptation in Drosophila melanogaster. PLoS Biology, 6(10), 2109–2129. https://doi.org/10.1371/journal.pbio.0060251

Hollister, J. D., & Gaut, B. S. (2009). Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Research, 19(8), 1419–28. https://doi.org/10.1101/gr.091678.109

Kiehart, D. P., Crawford, J. M., & Montague, R. A. (2007). Collection, dechorionation, and preparation of Drosophila embryos for quantitative microinjection. CSH Protocols, 2007, pdb.prot4717. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21357063

Kofler, R., Betancourt, A. J., Schlötterer, C., Sander, C., & Roth, F. (2012). Sequencing of Pooled DNA Samples (Pool-Seq) Uncovers Complex Dynamics of Transposable Element Insertions in Drosophila melanogaster. PLoS Genetics, 8(1), e1002487. https://doi.org/10.1371/journal.pgen.1002487

Lee, I.-M. (1994). Use of Mycoplasmalike Organism (MLO) Group-Specific Oligonucleotide Primers for Nested-PCR Assays to Detect Mixed-MLO Infections in a Single Host Plant. Phytopathology, 84(6), 559. https://doi.org/10.1094/Phyto-84-559

Mateo, L., Ullastres, A., González, J., Parajuli, B., & Perez-Miller, S. (2014). A Transposable Element Insertion Confers Xenobiotic Resistance in Drosophila. PLoS Genetics, 10(8), e1004560. https://doi.org/10.1371/journal.pgen.1004560

McClintock, B. (1956). Controlling elements and the gene. Cold Spring Harbor Symposia on Quantitative Biology, 21, 197–216. https://doi.org/10.1101/SQB.1956.021.01.017

Petrov, D. A. (2003). Size Matters: Non-LTR Retrotransposable Elements and Ectopic Recombination in Drosophila. Molecular Biology and Evolution, 20(6), 880–892. https://doi.org/10.1093/molbev/msg102

Rothwell, W. F., & Sullivan, W. (2007). Drosophila embryo collection. CSH Protocols, 2007(9), pdb.prot4825. https://doi.org/10.1101/PDB.PROT4825

Schwartzman, O., Mukamel, Z., Oded-Elkayam, N., Olivares-Chauvet, P., Lubling, Y., Landan, G., … Tanay, A. (2016). UMI-4C for quantitative and targeted chromosomal contact profiling. Nature Methods, 13(8), 685–91. https://doi.org/10.1038/nmeth.3922

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., … de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nature Genetics, 38(11), 1348–1354. https://doi.org/10.1038/ng1896

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics, 123(3), 585–95. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2513255

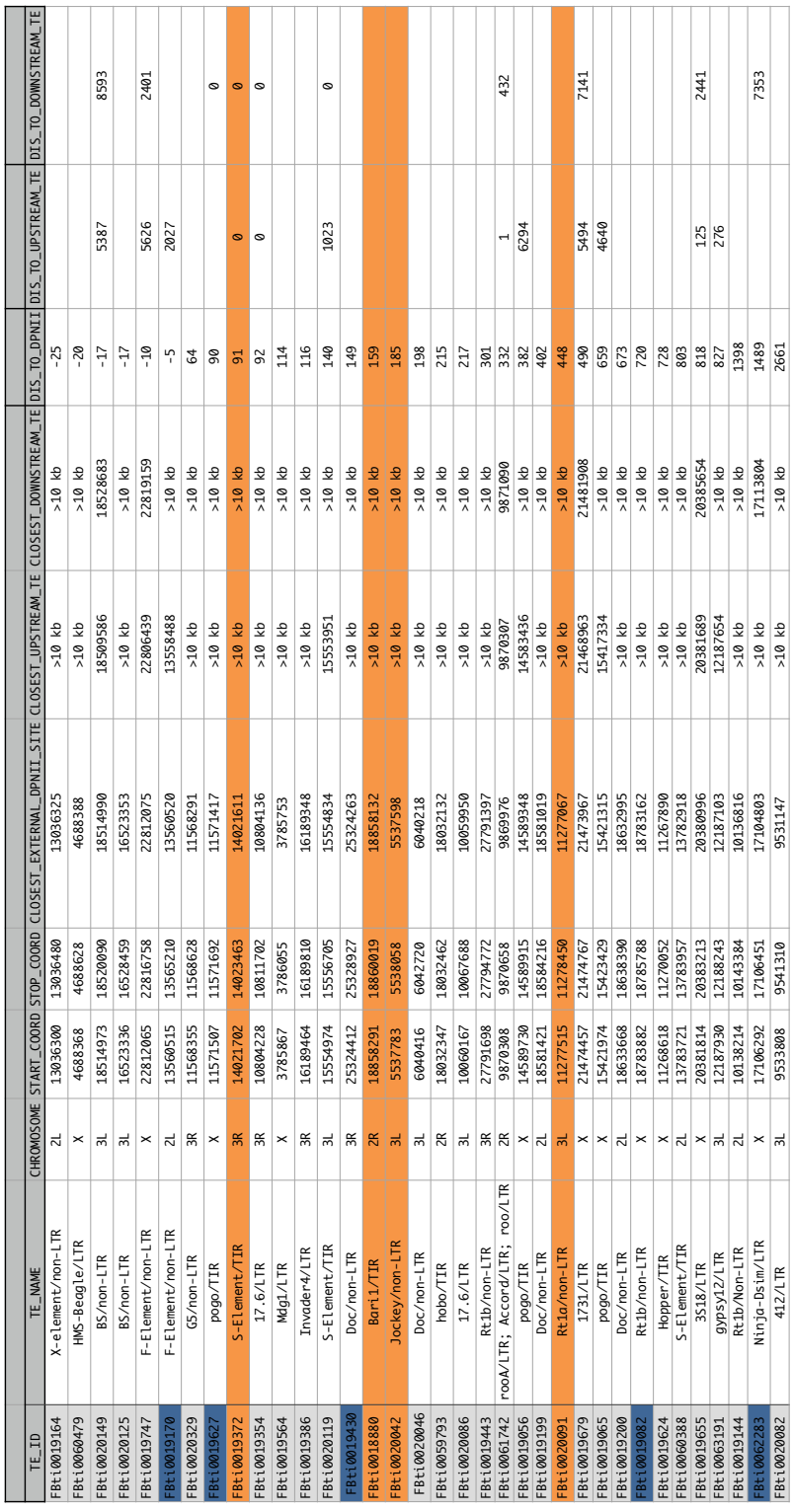
Wang, X., Weigel, D., & Smith, L. M. (2013). Transposon Variants and Their Effects on Gene Expression in Arabidopsis. PLoS Genetics, 9(2). https://doi.org/10.1371/journal.pgen.1003255

Wei, G., Wei, L., Zhu, J., Zang, C., Hu-Li, J., Yao, Z., … Zhao, K. (2009). Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. Immunity, 30(1), 155–67. https://doi.org/10.1016/j.immuni.2008.12.009

Zaratiegui, M. (2017). Cross-Regulation between Transposable Elements and Host DNA Replication. Viruses, 9(3), 57. https://doi.org/10.3390/v9030057

Appendix

Appendix 01: Table of Studied TEs.



TE_ID	TE_NAME	CHROMOSOME	START_COORD	STOP_COORD	CLOSEST_EXTERNAL_DPNII_SITE	CLOSEST_UPSTREAM_TE	CLOSEST_DOWNSTREAM_TE	DIS_TO_DPNII	DIS_TO_UPSTREAM_TE	DIS_TO_DOWNSTREAM_TE
FBti0019164	X-element/non-LTR	2L	13036300	13036480	13036325	>10 kb	>10 kb	-25		
FBti0060479	HMS-Beagle/LTR	X	4688368	4688628	4688388	>10 kb	>10 kb	-20		
FBti0020149	BS/non-LTR	3L	18514973	18520090	18514990	18509586	18528683	-17	5387	8593
FBti0020125	BS/non-LTR	3L	16523336	16528459	16523353	>10 kb	>10 kb	-17		
FBti0019747	F-Element/non-LTR	X	22812065	22816758	22812075	22806439	22819159	-10	5626	2401
FBti0019170	F-Element/non-LTR	2L	13560515	13565210	13560520	13558488	>10 kb	-5	2027	
FBti0020329	GS/non-LTR	3R	11568355	11568628	11568291	>10 kb	>10 kb	64		
FBti0019627	pogo/TIR	X	11571507	11571692	11571417	>10 kb	>10 kb	90		0
FBti0019372	S-Element/TIR	3R	14021702	14023463	14021611	>10 kb	>10 kb	91	0	0
FBti0019354	17.6/LTR	3R	10804228	10811702	10804136	>10 kb	>10 kb	92	0	0
FBti0019564	Mdg1/LTR	X	3785867	3786055	3785753	>10 kb	>10 kb	114		
FBti0019386	Invader4/LTR	3R	16189464	16189810	16189348	>10 kb	>10 kb	116		
FBti0020119	S-Element/TIR	3L	15554974	15556705	15554834	15553951	>10 kb	140	1023	0
FBti0019430	Doc/non-LTR	3R	25324412	25328927	25324263	>10 kb	>10 kb	149		
FBti0018880	Bari1/TIR	2R	18858291	18860019	18858132	>10 kb	>10 kb	159		
FBti0020042	Jockey/non-LTR	3L	5537783	5538058	5537598	>10 kb	>10 kb	185		
FBti0020046	Doc/non-LTR	3L	6040416	6042720	6040218	>10 kb	>10 kb	198		
FBti0059793	hobo/TIR	2R	18032347	18032462	18032132	>10 kb	>10 kb	215		
FBti0020086	17.6/LTR	3L	10060167	10067688	10059950	>10 kb	>10 kb	217		
FBti0019443	Rt1b/non-LTR	3R	27791698	27794772	27791397	>10 kb	>10 kb	301		
FBti0061742	rooA/LTR; Accord/LTR; roo/LTR	2R	9870308	9870658	9869976	9870307	9871090	332	1	432
FBti0019056	pogo/TIR	X	14589730	14589915	14589348	14583436	>10 kb	382	6294	
FBti0019199	Doc/non-LTR	2L	18581421	18584216	18581019	>10 kb	>10 kb	402		
FBti0020091	Rt1a/non-LTR	3L	11277515	11278450	11277067	>10 kb	>10 kb	448		
FBti0019679	1731/LTR	X	21474457	21474767	21473967	21468963	21481908	490	5494	7141
FBti0019065	pogo/TIR	X	15421974	15423429	15421315	15417334	>10 kb	659	4640	
FBti0019200	Doc/non-LTR	2L	18633668	18638390	18632995	>10 kb	>10 kb	673		
FBti0019082	Rt1b/non-LTR	X	18783882	18785788	18783162	>10 kb	>10 kb	720		
FBti0019624	Hopper/TIR	X	11268618	11270052	11267890	>10 kb	>10 kb	728		
FBti0060388	S-Element/TIR	2L	13783721	13783957	13782918	>10 kb	>10 kb	803		
FBti0019655	3S18/LTR	X	20381814	20383213	20380996	20381689	20385654	818	125	2441
FBti0063191	gypsy12/LTR	3L	12187930	12188243	12187103	12187654	>10 kb	827	276	
FBti0019144	Rt1b/Non-LTR	2L	10138214	10143384	10136816	>10 kb	>10 kb	1398		
FBti0062283	Ninja-Dsim/LTR	X	17106292	17106451	17104803	>10 kb	17113804	1489		7353
FBti0020082	412/LTR	3L	9533808	9541310	9531147	>10 kb	>10 kb	2661		

Appendix 01a. List of TEs studied
List of TEs studied with TE_ID, TE Name, Chromosome, Start and stop sites, closest external DpnII site, Closest upstream and downstream TEs, Distance to DpnII, upstream and downstream TE. The row highlighted in orange show good candidates based on location of the primers and DpnII sites relative to the TE insertion. The TEs highlighted in blue represent those for which 3' primers were tried. Some cells are left blank in the columns 'Dis_to_Upstream_TE' and 'Dis_to_Downstream_TE' as they were more than 10kb away.

TE_ID	Primers	5' End of TE								3' End of TE							
		F/R	Sequence1 U	f	Sequence2 D	f	Dis_Primer/DpnII	Dis_TE/Primer	F/R	Sequence1 U	f	Sequence2 D	f	Dis_Primer/DpnII	Dis_TE/Primer
FBti0019164	MANY	F	TAAGAGATGCGGAACGAAAGAG	2	GATGTTAAGCAGAGTCGGTTAGT	1	76	51							
		R			~~WITHIN TE~~										
FBti0060479	MANY	F	CGAGAAGTCAGGGTGTTTCTAC	2	CACACATCCACGCACAATAAAG	3		20							
		R			~~WITHIN TE~~										
FBti0020149	MANY	F	TCTATTTGAGAGCGGAACATAAGA	1	GTTTCTTAACAACAAGCGCATCTA	1	51	34							
		R	~~WITHIN TE~~												
FBti0020125	MANY	F	TCATTCGGCCTCTGAAAGTATG	1	GGTCCGCAAACAATTGAACTAA	1	36	19							
		R	~~WITHIN TE~~												
FBti0019747	MANY	F	CATCAAAGTAAGTTCACGATGCTAAA	1	CATGCCTTACAGCTAAATACATTAAATAC	1		40							
		R	~~WITHIN TE~~												
FBti0019170	ONE/MANY	F	~~ALL OVERLAPPING~~						F	~~ALL OVERLAPPING~~					
		R	~~WITHIN TE~~						R	~~CLOSE TO DpnII~~					
FBti0020329	MANY	F	GGGTTTGCCACGAAACATAC	1	GTATACCAATAATAAGTAACGTC	1		68							
		R	~~NO ASSAYS FOUND~~												
FBti0019627	MANY	F	GTCTCACGGACCTCATGAAATA	1	GCCTGAGTAGACCAGAACATAAA	1			F	AGCAAACCGTCTTACAATCTCT	1	GGTTGGCCATATTCTGGTGTATAG	1	8	12
		R	TCAGACTTAAAGGCAGCCATC	1	ATGGCTGCCTTTAAGTCTGAC	1		23	R	~~CLOSE TO DpnII~~					
FBti0019372	MANY	F	CACCATTTGCGACGACATTG	2	CTAGCGTGGGCATCACTAAT	1	13	104							
		R	GACATCCATGGGTTTCGATGA	1	CAAACCCAGTTGCATAGTCATAAA	1	0	27							
FBti0019354	ONE/MANY	F	CTCGGGAACTCTTGGTAAACTC	1	TGCGCCGACAAGGAATTAT	2	55	147							
		R	CACTCCAGTTATTAAGGGTCATT	1	TTTATTGGTAAGAATCCGCGTCA	1	3	41							
FBti0019564	MANY	F	CGTATGGAGTGTGGGTGTTATC	1	AAAGCCCTTCGGTCACAATC	1	33	147							
		R	GTTAGAAACGGTTGGCTTGTTG	1	ATGTCAGAGACAACTCACCTTG	1	26	37							
FBti0019386	MANY	F	CCAGCTGTTCAGACTTTCAGATA	1	GAAGTGGCAAATCATCGTTTCA	1	28	144							
		R	GCTGCAGGTTCTTTCCAATATC	1	GTGCTCAAGACCCTAATGGAA	1	8	10							
FBti0020119	ONE/MANY	F	TGAAATGTAACAGTTGTGCTTTAGG	1	GGCAGAAGCATCAAACAATTAAGA	1	48	153							
		R	~~ALL OVERLAPPING~~												
FBti0019430	MANY	F	~~CLOSE TO ANOTHER DpnII SITE~~						F	CGTGGCAACTTGACCATGTA	1	GCAACATATTAGTGGCGTAGTTG	1	13	33
		R	GTTTGGAGCCAAGAGCTATGA	2	TACATGGTGCGTTCGATTTCG	3		11	R	CTTTGACGTGGCAACTTGAC	1	TATTAGTGGCGTAGTTGTCACC	1	9	27
FBti0018880	MANY	F	TCGTATTCCACACGGCAATG	1	TTGGCAACAAATTGAGTGAGAG	1		201							
		R	TGTTCACTTTACTTTCGTGGATTG	1	CGTGTTGGTAATTCGAGTTCTTAAT	1	16	19							
FBti0020042	MANY	F	GTGAGCCTAACACCAGTCTTT	1	CTGCTTTACTGCTTGCTGTG	1		221							
		R	GCACCTCTTCATCTAGCAACA	1	GAAACACGAACTCATAGTAATGGC	1	43	35							
FBti0020046	MANY	F	~~CLOSE TO ANOTHER DpnII SITE~~												
		R	GACCCTCAAACTGGGAACATAA	1	GTAACAGCTTGGTTCTGCATTC	1	10	57							
FBti0059793	MANY	F	CTTCTGCCCAAACCATTTATGT	1	CGCATCGTGCACAGGAAA	1	134	349							
		R	GTCAGATGCTGCACTCGGAAA	1	CCCTCACACAAACACAACAC	1	60	29							
FBti0020086	MANY	F	CAAAGTCGAAAGCCAGCAAA	7	AGAGGGACAGAGTCACATACA	1	120	337							
		R	AGCGGAGAGTGGAAGAGT	70	GCATCTCGTTTCGTTTCGTTTAG	1	3	86							
FBti0019443															
FBti0061742															
FBti0019056															
FBti0019199															
FBti0020091	MANY	F	TCCAAGCCTTTGGGCTAATC	1	AGTATTCCGAGCGAAAGAAAGT	1	19	467							
		R	TCCCGTTTCCCAATTGGATTT	1	GGTCTCCAGTTGAGCTGATAAA	1	62	287							
FBti0019679															
FBti0019065															
FBti0019200															
FBti0019082									F	~~ALL OVERLAPPING~~					
									R	~~CLOSE TO DpnII~~					
FBti0019624															
FBti0060388															
FBti0019655															
FBti0063191															
FBti0019144															
FBti0062283									F	~~ALL OVERLAPPING~~					
									R	~~CLOSE TO DpnII~~					
FBti0020082															

Appendix 01b. List of TEs studied

List of TEs studied with TE_ID, 5' and 3' primer sequences and hybridization sites on D.mel chromosome. Some cells were left blank as those TEs did not have a DpnII site near enough to have a site that could be easily studied. The numbers next to the primer sequences shown indicate the hybridization sites on the whole chromosome. Those highlighted in blue mean that that set of primers could not be used for the assay. Things that are crossed out signify (for those TEs that were to be studied) primers that could not be synthesized for the reasons underlying. The F and R next to the sequence of the primers signify whether they are the Forward or the Reverse primers for that TE.

Appendix 02: Embryo Collection Plate and Yeast Paste Recipe

PLATES		YEAST PASTE	
0.56 L	dH_2O	45 mL	dH_2O
88 mL	Molasses	272 µL	Propionic acid
22 g	Food Agar	30 g	Active Dry Yeast
9.5 mL	10% Togasept		

Appendix 02. Detailed recipes
Detailed recipes for Embryo Collection plates and Yeast paste used in protocol. Data in this study was not dependent on the exact concentrations and makeup of these items.

Appendix 03: Shell Script for 'TE_to_DGRP_freq_retention.txt'

```
# Downloaded files from http://www.bio.brandeis.edu/laulab/Tidal_Fly/Tidal_Fly_Home.html
#  Downloaded by clicking "here (1.60 GB)." at bottom of page.
#  Put it in working folder.
gunzip Tidal_Fly_v1Archive20150930.zip
cd Tidal_Fly_v1Archive20150930
gunzip DGRP_flies.zip
cd DGRP_flies
#  Searched for "*Depletion_Annotated_TEonly.bed"
find . -name "*Depletion_Annotated_TEonly.bed" | mv ~/LINUX/DGRP_analysis
cd ~/LINUX/DGRP_analysis
less *Depletion_Annotated_TEonly.bed > RAL.bed
sort -k1,1 -k2,2n RAL.bed > RAL_sorted.bed
#  Had a file called "TE_sorted.bed" that had presorted list of TEs in .bed format.
awk '{print chr$1"\t"$2"\t"$3"\t"$4}' TE_sorted.bed > TE_chr.bed
bedtools intersect -wo -a TE_chr.bed -b RAL_sorted_unfixed.bed | head
awk '($3-$2)>0{print$1"\t"$2"\t"$3"\t"$4}' RAL_sorted_unfixed.bed > RAL_unsorted.bed
awk '($3-$2)<0{print$1"\t"$3"\t"$2"\t"$4}' RAL_sorted_unfixed.bed >> RAL_unsorted.bed
sort -k1,1 -k2,2n RAL_unsorted.bed > RAL_sorted.bed
bedtools intersect -wo -a TE_chr.bed -b RAL_sorted.bed | head
bedtools intersect -wo -a TE_chr.bed -b RAL_sorted.bed > TE_depletion_full.txt
```

```
#    In awk for next command, 117 is the number of non-repetitive strains.
bedtools groupby -i TE_depletion_full.txt -g 8 -c 4 -o freqasc | tr "," "\n" | tr ":" "\t" | awk '{print $1"\t"$2/177}' > TE_to_DGRP_freq.txt
awk '{print  $1"\t"1-$2}' TE_to_DGRP_freq.txt > TE_to_DGRP_freq_retention.txt
#    However, there were 4 TEs that were present in all strains as found out by
bedtools intersect -wo -v -a TE_chr.bed -b RAL_sorted.bed | head
#    They could not be intersected as they were not present in the RAL files and thus not depleted (still present in the strains).
bedtools intersect -wo -v -a TE_chr.bed -b RAL_sorted.bed | awk '{print$4"\t""1.000"}' >> TE_to_DGRP_freq_retention.txt
```

Appendix 04: FILE data for 'TE_to_DGRP_freq_retention.txt'

TE ID	Frequency in Population	TE ID	Frequency in Population
FBti0060388	1.00000	FBti0020091	0.79661
FBti0019199	1.00000	FBti0019056	0.785311
FBti0062283	1.00000	FBti0020149	0.774011
FBti0019082	1.00000	FBti0019065	0.734463
FBti0059793	0.99435	FBti0019627	0.706215
FBti0063191	0.99435	FBti0020119	0.700565
FBti0019200	0.988701	FBti0019386	0.689266
FBti0020329	0.971751	FBti0019747	0.689266
FBti0019655	0.966102	FBti0019679	0.672316
FBti0018880	0.960452	FBti0019170	0.666667
FBti0020125	0.943503	FBti0020082	0.610169
FBti0019430	0.937853	FBti0019164	0.59322
FBti0019564	0.926554	FBti0019354	0.564972
FBti0060479	0.926554	FBti0019624	0.497175
FBti0061742	0.903955	FBti0019372	0.451977
FBti0019443	0.875706	FBti0020042	0.423729
FBti0020086	0.819209	FBti0020046	0.423729

FBti0019144	0.80226	–	–

Appendix 04. List of TEs and retention frequencies
List of TEs studied with retention frequencies in all DGRP strains from TIDAL-FLY.

Appendix 05: Shell Script for Distance between studied TEs and closest H3K4me1 site.

```
#  TE_start_stop_sorted.bed is a sorted list of the TEs in a .bed format.
#  H3K4me1_sorted.gff is a sorted .gff file of loci where this type of histone modification exerts an influence.
bedtools closest -a TE_start_stop_sorted.bed -b H3K4me1_sorted.gff > TE_to_H3K4me1.txt
cut -f1,4,14 TE_to_H3K4me1.txt > TE_to_H3K4me1.txt
```

Appendix 06: Indexed adapters.

http://marshall-lab.org/wp-content/uploads/2016/05/Marshall_lab_DamID-seq_protocol_v2016-05-03.pdf

Sequencing adaptors and primer sequences
(adaptor barcodes highlighted in red)

Universal	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
Index 1	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**ATCACG**ATCTCGTATGCCGTCTTCTGCTT*G
Index 2	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**CGATGT**ATCTCGTATGCCGTCTTCTGCTT*G
Index 3	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**TTAGGC**ATCTCGTATGCCGTCTTCTGCTT*G
Index 4	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**TGACCA**ATCTCGTATGCCGTCTTCTGCTT*G
Index 5	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**ACAGTG**ATCTCGTATGCCGTCTTCTGCTT*G
Index 6	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**GCCAAT**ATCTCGTATGCCGTCTTCTGCTT*G
Index 7	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**CAGATC**ATCTCGTATGCCGTCTTCTGCTT*G
Index 8	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**ACTTGA**ATCTCGTATGCCGTCTTCTGCTT*G
Index 9	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**GATCAG**ATCTCGTATGCCGTCTTCTGCTT*G
Index 10	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**TAGCTT**ATCTCGTATGCCGTCTTCTGCTT*G
Index 11	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**GGCTAC**ATCTCGTATGCCGTCTTCTGCTT*G
Index 12	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**CTTGTA**ATCTCGTATGCCGTCTTCTGCTT*G
Index 13	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**AGTCAA**ATCTCGTATGCCGTCTTCTGCTT*G
Index 14	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**AGTTCC**ATCTCGTATGCCGTCTTCTGCTT*G
Index 15	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**ATGTCA**ATCTCGTATGCCGTCTTCTGCTT*G
Index 16	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**CCGTCC**ATCTCGTATGCCGTCTTCTGCTT*G
Index 18	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**GTCCGC**ATCTCGTATGCCGTCTTCTGCTT*G
Index 19	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**GTGAAA**ATCTCGTATGCCGTCTTCTGCTT*G
PCR1	AATGATACGGCGACCACCGA*G
PCR2	CAAGCAGAAGACGGCATACGA*G

[Phos] = 5′ Phosphorylation
* = Phosphorothioate linkages

Appendix 06. List of indexed adapters
List of Indexed adapters used for ligation in order to be tagged for sequencing using Illumina.

Appendix 07: TE Presence or Absence in individual DGRP strains (25203 and 25189).

Bloomington Strain →	25203	25189	2 = both	Bloomington Strain →	25203	25189	2 = both
DGRP Strain →	732	379	1 = one	DGRP Strain →	732	379	1 = one
TE Name	–	–	0 = none	TE Name	–	–	0 = none
FBti0018880	1	1	2	FBti0019655	1	1	2
FBti0019056	0	1	1	FBti0019679	0	1	1
FBti0019065	1	1	2	FBti0019747	1	1	2
FBti0019082	1	1	2	FBti0020042	0	1	1
FBti0019144	1	1	2	FBti0020046	0	1	1
FBti0019164	0	1	1	FBti0020082	0	0	0
FBti0019170	1	1	2	FBti0020086	1	1	2
FBti0019199	1	1	2	FBti0020091	1	1	2
FBti0019200	1	1	2	FBti0020119	1	1	2
FBti0019354	1	1	2	FBti0020125	1	1	2
FBti0019372	1	1	2	FBti0020149	1	1	2
FBti0019386	0	1	1	FBti0020329	1	1	2
FBti0019430	1	1	2	FBti0059793	1	1	2
FBti0019443	1	1	2	FBti0060388	1	1	2
FBti0019564	1	1	2	FBti0060479	1	1	2
FBti0019624	0	1	1	FBti0061742	1	1	2
FBti0019627	1	0	1	FBti0062283	1	1	2
–	–	–	–	FBti0063191	1	1	2

Appendix 07. Presence or absence of TE from strain 25203 and 25189
Presence or absence of TEs from RAL-379 and RAL-732. (2) indicates that they were present in both TEs, (1) means that they were present in one and absent in the other, (0) indicates that they were absent in both. The unhighlighted 0s & 1s show absence or presence of TEs from that strain respectively.

Appendix 08: Solutions for Nuclei Preparation

Stop solution (500mL stock):
125 mM glycine
0.1% Triton X-100 (v/v) in PBS (use the 1X PBS rather than diluted 10X)
Filter sterilize

0.1M PMSF (100X):
Dissolve 0.174 g in 10mL 100% isopropanol.
Make 500uL aliquots and store in -20 freezer.

PBT (500mL):
0.1% (vol/vol) Triton X-100 in 1X PBS
Filter sterilize

1X Cell Lysis Buffer (500mL):
85 mM KCl
0.5% IGEPAL CA-630 (v/v) 5 mM HEPES,
pH 8.0 UltraPure H20
Filter sterilize

Appendix 09: Supplementary Files

"TE_to_H3K4me1.txt"
"TE_to_H3K4me1_short.txt"
"TE_to_H3K4me1_ultrashort.txt"
"all_DGRP_sorted.bed"
"TE_chr.bed"

All these files are too large (in terms of lines) in order to put into a .pdf.
All these files are in the folder Supplementary info. with the path –

Ellison – Lab → 01-Spring 2017 → Submissions → RESEARCH PAPER → FULL → Supplementary info.