Dynamique des éléments transposables chez Drosophila suzukii

Vincent Mérel

Numéro National de Thèse : 2021LYSE1070

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON

opérée par

l'Université Claude Bernard - Lyon 1

École doctorale 341 :
Evolution Ecosystèmes, Microbiologie, et Modélisation (E2M2)

Spécialité de doctorat : Biologie

Soutenue publiquement le 6 Avril 2021, par :

Vincent MÉREL

Dynamique des Éléments Transposables chez *Drosophila suzukii*

Devant le jury composé de :

BOISSINOT,	Stéphane	Professeur	NYU Abu Dhabi	Rapporteur
BOURQUE,	Guillaume	Professeur	Université McGill	Rapporteur
FABLET,	Marie	Maître de conférences	UCB-Lyon 1	Co-directrice de thèse
LEFÉBURE,	Tristan	Maître de conférences	UCB-Lyon 1	Examinateur
MOUCHIROUD,	Dominique	Professeur	UCB-Lyon 1	Examinatrice
SABOT,	François	Directeur de recherche	IRD	Examinateur
TENAILLON,	Maud	Directrice de Recherche	CNRS	Rapporteuse
VIEIRA,	Cristina	Professeur	UCB-Lyon 1	Directrice de thèse

Résumé de la thèse

Les Éléments Transposables (ET) sont des éléments génomiques égoïstes qui se multiplient en se copiant/coupant -collant en divers endroits du génome de leur hôte. Cette propriété particulière leur permet de se maintenir dans les populations, et même de proliférer, sans nécessairement apporter d'avantage à leur hôte. À l'exception de quelques exemples de copies présumées adaptatives, il a même été montré que ces séquences ont plutôt un impact négatif, voire neutre. À ce jour, de nombreuses zones d'ombres subsistent encore quant à leur dynamique. L'espèce invasive *Drosophila suzukii* offre une opportunité unique d'étudier la dynamique des ET. Originaire d'Asie elle a débuté une invasion des continents Américain et Européens en 2008, permettant ainsi de tester les effets des changements environnementaux et démographiques sur la dynamique des ET. Au cours de cette thèse un portrait du contenu en ET chez *D. suzukii* a d'abord été brossé dans un génome de référence, révélant la présence de 47% d'ET, principalement concentrés dans les régions pauvres en gènes. L'abondance des ET a ensuite été quantifiée dans 22 populations de *D. suzukii* mais aussi chez des espèces proches. La comparaison du contenu en ET entre populations invasives et natives, indique une accumulation de ces séquences au cours du processus invasif. Cette accumulation est associée à une diminution du proxy de la taille efficace $\widehat{\theta_W}$, suggérant un rôle d'un relâchement de la sélection purificatrice à l'encontre des ET dans leur prolifération. À l'échelle des espèces, la comparaison du contenu en ET entre *D. suzukii* et espèces proches, montre une accumulation d'ET dans la lignée *D. suzukii* au cours des quatre derniers millions d'années. Cette accumulation, concentrée dans les régions pauvres en gènes, est vraisemblablement responsable du fort pourcentage d'ET génomique chez *D. suzukii*. L'étude du dN/dS n'a pas permis d'associer cette augmentation du contenu en ET avec un relâchement de la sélection purificatrice. Au delà d'un potentiel rôle des variations d'intensité de la sélection, l'impact des changements environnementaux et de la variabilité génétique, sur la dynamique des ET a été étudié dans les populations de *D. suzukii*. Les résultats indiquent un rôle complexe de la variabilité génétique, avec une abondance d'ET associée à ~5000 régions génomiques, mais à aucune variable bioclimatique. L'étude du polymorphisme d'insertion dans les populations montre toutefois que les pressions de sélection exercées par l'environnement pourraient entraîner certaines copies d'ET jusqu'à fixation, avec notamment six insertions potentiellement impliquées dans l'adaptation des populations invasives.

PhD thesis summary

Transposable Elements (TEs) are selfish genomic elements that multiply by copying and pasting themselves at various locations in the host genome. This particular property allows them to maintain in populations, and even to proliferate, without necessarily bringing any advantage to their host. With the exception of a few examples of presumed adaptive copies, it has even been shown that these sequences have a negative, or neutral impact. So far, their dynamics remains unclear. The invasive species *Drosophila suzukii* offers a unique opportunity to study TE dynamics. Originally from Asia, it began an invasion of the American and European continents in 2008, allowing to test the effects of environmental and demographic changes on TE dynamics. During this PhD, a portrait of TE content in *D. suzukii* was first drawn in a reference genome, revealing the presence of 47% ET, mainly concentrated in gene-poor regions. TE abundance was then quantified in 22 populations of *D. suzukii* but also in related species. Comparison of TE contents between invasive and native populations indicates an accumulation of these sequences during the invasive process. This accumulation is associated with a decrease of the effective size proxy $\widehat{\theta_W}$, suggesting a role of a relaxation of purifying selection against TEs in their proliferation. At the species level, the comparison of TE contents between *D. suzukii* and related species shows an accumulation of TEs in *D. suzukii* lineage over the last four million years. This accumulation, concentrated in gene-poor regions, is likely responsible for the high percentage of TEs in *D. suzukii* genome. The study of dN/dS suggests that this increase in TE content is not associated with a relaxation of purifying selection. Beyond the potential role of variations in selection intensity, the impact of environmental changes and genetic variability on TE dynamics was studied in *D. suzukii* populations. Results indicate a complex role of genetic variability, with ET abundance associated with ~5,000 genomic regions, but no impact of bioclimatic variables on TE activity. However, the study of insertion polymorphism in the populations shows that environmental selection pressures could increase the frequency of some TE copies, including six insertions potentially involved in the adaptation of invasive populations.

« Faut qu'je marche,
Parce que j'comprends quand je marche.
Faut qu'je marche,
Parce que j'apprends quand je marche.
Faut qu'je marche,
Parce que je pense quand je marche.»

Ben Mazué

REMERCIEMENTS

— Concis, toujours concis ... —

Merci,

À Cristina, Marie et Matthieu pour votre bienveillance, votre implication, l'apport scientifique et humain à ma thèse, ainsi que la confiance que vous m'avez accordée.

À Patricia, pour avoir permis thèse et liberté.

À Annabelle, pour m'avoir aidé à enrichir mes travaux.

À Nelly, qui en Biologie moléculaire m'a pour ainsi dire tout appris.

Aux princesses de la pipette et à leur coach, qui iront plus loin que je ne le pourrai jamais.

Aux membres du jury.

À Inessa & Valentina, pour avoir réussi à rendre la moustiquaire accueillante.

À Cecilia, Justine et Marlène, parce qu'en ch*** c'est toujours mieux à plusieurs.

À Élise, Corentin, et tous ceux qui ont partagé avec moi cette expérience singulière qu'est la thèse.

À Alexis, Benjamin, Camille, Chloé, Sylvain, Pierre et Zaïnab pour les rires et la bonne humeur, ...

À Juliette et Paloma, parce que vous remercier pour ma thèse est aussi nécessaire qu'insuffisant.

À tous ceux qui comptent encore ...

À Jean-Pierre, Hélène, Cécile et Delphine,
Si il est vrai que "la famille, la vraie on se la choisit",
Vous êtes mon premier et mon second choix.

Table des matières

Liste des figures

Liste des tables

0 Introduction

Sommaire

0.1 Les Éléments Transposables

0.1.1 Deux définitions

En 1950, Barbara McClintock met en évidence chez le maïs des séquences génétiques mobiles qui impactent la coloration des grains (McClintock, 1950). Par la suite, plusieurs séquences, issues de diverses espèces, seront assimilées à celles découvertes par Barbara McClintock. Toutes ces séquences seront désignées comme « Éléments Transposables (ET) ». Aujourd'hui ce terme regroupe une variété de séquences qui ne partagent pas d'ancêtre commun exclusif. Par conséquent, ces séquences présentent des caractéristiques extrêmement variées et la question de leur définition n'est pas triviale. Dans les publications scientifiques sur les ET, ceux ci sont souvent définis, soit comme des séquences mobiles et/ou répétées (Jakšić et collab., 2017; Muñoz-López et García-Pérez, 2010; Ou et collab., 2019), soit comme des éléments génétiques égoïstes ou parasites (Weilguny et Kofler, 2019).

0.1.1.1 Des séquences mobiles et/ou répétées

Après la découverte des deux premiers ET par Barbara McClintock (McClintock, 1950), c'est dans un premier temps le caractère mobile qui va prévaloir pour définir un ET. C'est ainsi que Calos et Miller évoquent le « phénomène central de la transposition, c'est-à-dire l'apparition d'une portion d'ADN (l'ET) au milieu de séquences où il n'avait pas été détecté auparavant »[1] en décrivant les ET en 1980. Toutefois ces éléments possèdent une autre caractéristique remarquable, ils sont généralement répétés. Un même ET est souvent présent en plusieurs copies dans un même génome, de quelques copies à 1.2 millions de copies pour les éléments de type Alu chez l'humain (Deininger, 2011). Le terme ET étant parfois utilisé pour désigner une copie particulière, afin d'éviter toute ambiguïté on parle souvent de famille d'ET pour désigner l'ensemble des copies[2]. La création d'une nouvelle copie passant par une insertion de l'ET en un point précis du génome (voir Section 0.1.2), on parle aussi d'insertion pour désigner une copie. Avec le développement de technologies de séquençage et la quantité exponentielle de génomes à annoter, le caractère répété des ET va devenir primordial pour définir un ET. En effet, le caractère répété des séquences, à la différence de leur caractère mobile, est facile à évaluer à partir des données de séquençage. Seront alors considérées comme ET les séquences répétées dans un génome, à l'exclusion des répétitions simples [3] ou des gènes dupliqués. Il est important de noter que, si les ET sont souvent définis comme des séquences mobiles et/ou répétées, certains ET peuvent perdre leur caractère mobile et certains ne se retrouvent qu'en un seul endroit dans le génome.

[1] Traduit de l'anglais: *« the central phenomen of transposition, that is, the appearance of a defined length of DNA (the transposable element) in the midst of sequences where it had not previously been detected »*, (Calos et Miller, 1980)

[2] Dans ce manuscrit une copie sera explicitement désignée comme telle et toute mention de « l'élément P » par exemple désignera la famille

[3] Les répétitions simples sont des suites de nucléotides répétées à l'identique plusieurs fois.

0.1.1.2 Des éléments génétiques égoïstes/parasites

La deuxième définition fréquemment retrouvée dans les publications scientifiques décrit les ET comme des éléments génétiques égoïstes (WEILGUNY et KOFLER, 2019), voire comme des parasites (ORGEL et CRICK, 1980). Cette définition résulte de deux observations. La première est l'absence de fonction spécifique dans le génome de leur hôte. En effet, si les deux premiers ET ont été identifiés pour leur effet sur la coloration des grains de maïs (MCCLINTOCK, 1950), les ET n'ont pas nécessairement d'effet sur le phénotype de leur hôte (DOOLITTLE et SAPIENZA, 1980; ORGEL et CRICK, 1980). La deuxième observation est que ces éléments possèdent une capacité de multiplication, un événement de transposition pouvant s'accompagner de la formation d'une ou plusieurs copies de l'ET. Ceci leur permet de subsister malgré leur absence de fonction, voire un effet négatif sur la valeur sélective de leur hôte (CHARLESWORTH et CHARLESWORTH, 1983). L'utilisation du terme « parasite », plus lourd d'implications que l'appellation « éléments génétiques égoïstes » vient de ce que les auteurs considèrent un impact global négatif ou pas.

0.1.2 Deux mécanismes de transposition

Le regroupement sous le terme d'ET de diverses séquences, du fait de leur mobilité ou répétitivité, a abouti à la formation d'un groupe de séquences ne partageant pas d'ancêtre commun exclusif et possédant des caractéristiques extrêmement variées. Parmi ces caractéristiques variables se trouve notamment le mécanisme de transposition. Les deux principales variantes définissent les deux classes d'ET. La classification des ET comprend des niveaux hiérarchiques en dessous du niveau classe. Toutefois cette classification étant d'une complexité notoire, controversée (KAPITONOV et JURKA, 2008; SEBERG et PETERSEN, 2009; WICKER et collab., 2007, 2009), et décrite pour les principaux ET présents chez la Drosophile (notre objet d'étude par la suite) en section « TE diversity » du Chapitre 1, nous n'entrerons pas ici dans le détail.

0.1.2.1 Via un intermédiaire ARN : ET de classe 1

Les ET de classe 1, ou rétrotransposons, correspondent aux ET qui transposent via un intermédiaire ARN. La séquence génomique de l'ET est dans un premier temps transcrite en ARN, avant d'être rétrotranscrite en ADN. Le fragment d'ADN ainsi créé est alors inséré à une autre position du génome par le biais d'une intégrase. On parle de rétrotranscription ou d'un mécanisme de transposition par « copier-coller ». Cette dernière appellation souligne ainsi que ce mécanisme génère une copie de l'ET tout en conservant l'ET initial. Un rétrotransposon autonome code pour ses propres protéines lui permettant d'être mobilisé. Des copies ne possédant pas les séquences codantes pour ces protéines (copies dites « non-autonomes ») peuvent être mobilisées grâce à d'autres copies de la même famille, ou d'une famille apparentée, qui produisent à leur place les protéines nécessaires.

0.1.2.2 Via un intermédiaire ADN : ET de classe 2

Les ET de classe 2, ou transposons, correspondent aux ET qui transposent via un intermédiaire ADN. Plus précisément, après cassure de l'ADN, l'un ou les deux brins de l'ET, serviront d'intermédiaires. Dans les cas où les deux brins servent d'intermédiaire, on parle de mécanisme de transposition par « couper-coller ». Ici l'ET se déplace et il n'y a pas nécessairement de génération d'une nouvelle copie. Attention toutefois, une nouvelle copie peut être générée si la transposition se produit lors de la réplication de l'ADN. En effet, si l'ET transpose d'une région dupliquée à une région non encore dupliquée, alors l'ET ne sera plus présent que sur l'une des deux chromatides en région déjà dupliquée mais sera aussi présent sur la chromatide de la région non dupliquée. Les ET de classe 2 qui utilisent un seul brin comme intermédiaire peuvent générer plusieurs copies lors d'un événement de transposition car l'intermédiaire peut être répliqué à l'état libre.

0.1.3 Impact sur l'hôte

Parce que les ET sont relativement grands, possèdent des séquences régulatrices et des marques épigénétiques dont l'effet peut s'étendre aux gènes voisins (Lee et Karpen, 2017; Rebollo et collab., 2012), ils ont un impact considérable sur leur hôte.

0.1.3.1 Un impact global négatif

Plusieurs éléments de preuves suggèrent un impact global négatif des ET sur leur hôte. Chez la Drosophile par exemple, des études montrent que l'activité des ET diminue la valeur sélective (ou *fitness*) de l'hôte. Ainsi l'activité des éléments P et I chez *D. melanogaster* peut entraîner la stérilité (Kidwell et collab., 1977; Picard, 1976). Chez *D. simulans* l'activité somatique de l'élément mariner est associée à une réduction de la longévité (Nikitin et Woodruff, 1995). En accord avec un impact global négatif, une forte contre sélection à l'encontre des ET a été retrouvée chez une variété d'organismes (Boissinot et collab., 2006; Bourgeois et collab., 2020; Petrov et collab., 2003) (voir section 1.2). Enfin, l'évolution dans le génome de l'hôte de systèmes de défense complexes contre les ET suggère très fortement un impact global négatif (voir Section 0.1.4). Cet impact négatif est très vraisemblablement lié à deux effets principaux des ET au niveau génomique: une altération de la fonction des gènes (ou *gene-disruption hypothesis*) (Finnegan, 1992), et une favorisation de la recombinaison ectopique (ou *ectopic recombination hypothesis*) (Montgomery et collab., 1987). Il est à noter que l'hypothèse que la transcription et la traduction des ET soit coûteuse et que les protéines produites puissent affecter les processus cellulaires à été émise (*deleterious TE-product expression hypothesis*) (Nuzhdin, 1999). Il y a toutefois peu de support empirique à cette hypothèse pour le moment (Petrov et collab., 2011), elle ne sera donc pas détaillée ici.

0.1.3.2 ... par altération de la fonction des gènes

Une partie de l'impact global des ET sur leur hôte résulte de leur capacité à altérer la fonction des gènes (Finnegan, 1992). Chez la Drosophile par exemple l'élément P a été massivement

utilisé pour générer des mutations altérant la fonction des gènes et ainsi déterminer leur fonction (SPRADLING et collab., 1999). Cette altération de fonction peut être liée soit à une protéine rendue non effective, par exemple suite à une insertion dans un exon, soit par un changement du niveau d'expression du gène. Le changement de niveau d'expression du gène lui-même peut avoir deux origines. Une séquence régulatrice peut être apportée par l'ET au gène (REBOLLO et collab., 2012), ou alors les marques épigénétiques portées par l'ET peuvent affecter l'expression du gène (LEE et KARPEN, 2017). Le terme épigénétique désignant des modifications du niveau d'expression sans changement de la séquence d'ADN.

0.1.3.3 ... par favorisation de la recombinaison ectopique

Le deuxième impact majeur des ET est lié à leur propension à favoriser la recombinaison ectopique. La recombinaison ectopique correspondant à la recombinaison entre des séquences plus ou moins identiques présentes à différents endroits du génome (MONTGOMERY et collab., 1987). Les réarrangements chromosomiques produits peuvent avoir de très forts effets délétères. Cette capacité des ET à induire la recombinaison ectopique dépend à la fois de leur taille et de leur nombre (PETROV et collab., 2003). Un ET de grande taille et avec un fort nombre de copies sera plus susceptible d'induire des événements de recombinaison ectopique qu'un petit ET avec un faible nombre de copies.

0.1.4 Les défenses de l'hôte

Au cours de l'évolution, des mécanismes de défense contre les ET et leur effets négatifs ont émergé chez les hôtes (SLOTKIN et MARTIENSSEN, 2007). La plupart de ces mécanismes relèvent de l'épigénétique. Il existe une variabilité importante dans les mécanismes de défense entre organismes. Ainsi on trouve trois principaux types de mécanismes de défense non mutuellement exclusifs: 1) la méthylation de l'ADN (e.g. chez les plantes) (SLOTKIN et MARTIENSSEN, 2007), 2) la modification des histones (chez la plupart des animaux) (OZATA et collab., 2019), et 3) l'utilisation de petits ARNs de séquences complémentaires aux ET (chez la plupart des animaux) (OZATA et collab., 2019). Ces mécanismes de défenses de l'hôte, limitent l'expression des ET au niveau transcriptionnel ou post-transcriptionnel. La méthylation de l'ADN, et la modification des histones, vont entraîner un changement de conformation local de la chromatine[1] et affecter le niveau de transcription. Les petits ARN vont entraîner la dégradation des transcrits d'ET de séquences complémentaires, mais ils peuvent aussi promouvoir la modification des histones. Les mécanismes de contrôle des ET chez la Drosophile sont décrits plus en détail dans le Chapitre 1 (voir section 1.8 Host Defenses). Brièvement, la lignée germinale, c'est-à-dire les cellules à l'origine des gamètes, est protégée par la voie dite des ARN interagissants avec PIWI (ou *PIWI-interacting RNA* en anglais) (piRNA). Ces piRNA correspondent à des petits ARN qui vont interagir avec des protéines dites PIWI pour réprimer les ET au niveau transcriptionnel ou post-transcriptionnel. Ces petits ARN sont issus de la transcription de loci particuliers appelés cluster de piRNA qui contiennent de nombreuses copies d'ET plus ou moins dégradées.

[1] La chromatine correspond à la structure dans laquelle l'ADN est empaqueté et compacté

0.2 Dynamique des ET

La question de la dynamique des ET est ancienne et de nombreuses zones d'ombre subsistent encore à ce jour (CHARLESWORTH et CHARLESWORTH, 1983). En mettant en lumière la variabilité de contenu génomique en ET entre organismes, les années 2000 et les premiers projets de séquençage, ont grandement contribué à l'intérêt porté à cette question. En effet, le contenu en ET varie fortement entre espèces éloignées: il est par exemple de 12% chez le nématode contre environ 80% chez le maïs (C. ELEGANS SEQUENCING CONSORTIUM, 1998; SCHNABLE et collab., 2009). Des variations sont aussi observables entre espèces d'un même genre (e.g. *Drosophila* ou *Leptidea*) (SESSEGOLO et collab., 2016; TALLA et collab., 2017), et même entre populations d'une même espèce (e.g. les populations européennes de *D. melanogaster*) (KAPUN et collab., 2018). Jusqu'à aujourd'hui les moteurs de ces variations restent mal compris.

0.2.1 Les principaux moteurs

0.2.1.1 Sélection purificatrice et taille efficace des populations

Du fait de leur effet négatif sur la valeur sélective de l'hôte, la sélection purificatrice limite la prolifération des ET (BOISSINOT et collab., 2006; BOURGEOIS et collab., 2020; PETROV et collab., 2003). La taille efficace des populations, en jouant sur l'efficacité de la sélection, devrait donc avoir un impact fort sur la dynamique des ET. Dans une population « idéale« [1] la taille efficace d'une population (ou *effective population size* ou N_e) correspond au nombre d'individus dans la population. Dans la pratique les populations ne sont pas idéales, e.g. il n'y a pas 50% de mâles et 50% de femelles, la taille efficace de la population est donc plus petite que le nombre d'individus dans la population. Plus une population s'éloignera de la population idéale plus sa taille efficace déviera de la taille de la population. Parce que les ET sont principalement délétères donc, et que de petits N_e entraînent une plus faible efficacité de la sélection contre les effets délétères, de petits N_e devraient être associés à de plus gros contenus en ET (LYNCH et CONERY, 2003). En accord avec cette hypothèse, LYNCH et CONERY (2003) ont trouvé une corrélation négative entre la taille du génome, positivement correlée au contenu en ET, et les variations de N_e entre diverses espèces. Toutefois l'impact de N_e est encore controversé. En effet, en prenant en compte l'inertie phylogénétique dans la corrélation de LYNCH et CONERY (2003) celle-ci devient non significative (WHITNEY et collab., 2010). De manière générale un effet de la taille efficace a été souvent suggéré (GARCÍA GUERREIRO et collab., 2008; GARCÍA GUERREIRO et FONTDEVILA, 2011; LOCKTON et collab., 2008; TALLA et collab., 2017), mais rarement démontré. Chez les isopodes toutefois la colonisation du milieu souterrain s'est accompagnée d'une diminution de la taille efficace des populations et d'une augmentation du contenu en ET (LEFÉBURE et collab., 2017). Chez la Drosophile en revanche, l'étude du contenu en ET dans les génomes de 12 espèces éloignées suggère une corrélation positive entre contenu en ET

[1] Pour une population idéale, ou population idéale au sens de Wright-Fisher:

- La taille de la population est constante;
- Les générations sont non chevauchantes;
- Il n'y a pas d'effet de la sélection;
- Le nombre de descendants suit une loi de Poisson de paramètre 1.

et taille efficace (CASTILLO et collab., 2011).

0.2.1.2 L'environnement

Un autre facteur qui pourrait jouer un rôle important dans la dynamique des ET c'est l'environnement (voir aussi le Chapitre 1 section 1.6 pour ce qui est de la Drosophile). Plusieurs études en laboratoire suggèrent en effet que l'activité des ET peut varier avec les conditions environnementales. Ainsi il a été montré que l'activité de certains ET variait avec la température chez la Drosophile (JAKŠIĆ et collab., 2017; KOFLER et collab., 2018). Dans les cellules somatiques du riz, l'ET mPing est mobilisé en réponse à de fortes pressions hydrostatiques (LIN et collab., 2006). Cette réponse des ET à l'environnement est vraisemblablement liée à la présence de séquences régulatrices qui interagissent avec des facteurs de transcription de l'hôte produits de manière environnement dépendante (JAKŠIĆ et collab., 2017). Toutefois des changements de l'activité des ET en relation avec l'environnement peuvent aussi être liés à une perturbation du système de défense de l'hôte. Chez la Drosophile par exemple, l'infection par le virus Sindbis affecte l'activité des ET en modulant la quantité de petits ARN dirigés contre les ET (ROY et collab., 2020). Il est important de noter que, si le nombre d'études en laboratoire est important, les études *in natura* sont elles plutôt rares. Chez *Arabidopsis thaliana* des corrélations entre le nombre de copies ET des variables géo-climatiques ont été trouvées pour 15 ET (QUADRANA et collab., 2016). Chez *D. melanogaster* des corrélations significatives ont été trouvées entre abondance et variables géographiques/environnementales pour 4 ET (LERAT et collab., 2019). Ceci étant dans ce dernier cas, un effet confondant de l'histoire démographique des populations ne peut être exclu.

0.2.1.3 Variation génétique

Un dernier moteur capable d'influencer la dynamique des ET dans les populations, c'est l'existence de certaines variations génétiques, au niveau du génome de l'hôte ou des séquences d'ET. En effet, chez la Drosophile plusieurs études suggèrent un effet du génotype sur l'activité des ET (BIÉMONT et collab., 1987; PASYUKOVA et NUZHDIN, 1993), avec une variation de cette activité entre lignées isogéniques[1]. Cet effet du génotype peut-être lié à: 1) des variations au niveau des séquences géniques de l'hôte, 2) la présence de copies d'ET particulières appelées « master copies », 3) des variations dans la proportion du génome hôte occupée par des clusters de piRNA.

Variation des séquences géniques de l'hôte

Intuitivement des variations de séquences au niveau de deux types de gènes en particulier sont susceptibles d'affecter l'activité des ET : les facteurs de transcription et les gènes impliqués dans la défense de l'hôte. Pour l'instant, les études à ce sujet sont rares. Une étude chez *A. thaliana*, plus spécifiquement une étude d'association entre les petits variants, de type SNP et de type Insertion/Déletion (InDel), et le contenu en ET, suggère un effet de variants au niveau d'un facteur de transcription ainsi que d'un gène potentiellement impliqué dans les défenses de l'hôte (QUADRANA et collab., 2016).

[1] Une lignée isogénique est constituée d'individus partageant le même patrimoine génétique, elle est souvent assimilée à un individu.

Master copies

Certaines copies d'ET pourraient à elles seules expliquer des différences de dynamique entre individus. Par exemple, chez l'homme, l'activité de l'élément L1 est principalement le résultat de l'activité d'un nombre restreint de copies (BROUHA et collab., 2003). On parle de « master-copies » en anglais. Le rôle de ces copies particulières dans les variations de contenus en ET, entre espèces ou entre populations, reste relativement méconnu. Les mécanismes impliqués, e.g. variation génétique au niveau des sites de fixation des facteurs de transcription, sont eux aussi peu clairs.

Variations de la proportion du génome occupée par des clusters de piRNA

Enfin un dernier facteur qui pourrait expliquer des différences de dynamique d'ET entre individus est la proportion du génome occupée par des clusters de piRNA (chez les animaux) (KOFLER, 2020). Un ET a moins de chance de s'insérer dans un cluster à chaque génération, et ainsi d'entraîner la production de piRNA dirigés contre lui, lorsque les clusters occupent une petite proportion du génome. En conséquence, les ET devraient proliférer chez les individus où la taille cumulée des clusters est faible. Si l'importance de la taille relative des clusters a été démontrée par un travail de simulation (KOFLER, 2020), il est important de noter qu'à ce jour on sait peu de choses sur les clusters de piRNA dans les populations. On ignore notamment s'il existe des clusters présents uniquement chez certains individus ou des clusters dont la taille varie entre individus. Toutefois, l'alignement des petits ARNs séquencés chez quatre lignées isogéniques de *D. melanogaster* et *D. simulans* sur leurs génomes suggère que certains clusters existent uniquement chez certains individus (MOHAMED et collab., 2020)(voir Annexe 1).

0.2.2 ET et sélection positive

La découverte dans les années 1980-1990 que les insertions d'ET étaient rarement fixées, ou même en fréquence importante dans les populations (BIÉMONT et collab., 1994; BROWN et MOSS, 1987; CHARLESWORTH et collab., 1992), a été considérée comme une preuve forte indiquant que la sélection positive ne jouait pas un rôle majeur dans la dynamique des ET. En effet, des insertions évoluant sous sélection positive devraient rapidement augmenter en fréquence, et éventuellement atteindre la fixation. L'observation que la majorité des insertions ségrégeait à faible fréquence, initialement obtenue par Hybridation In Situ chez la Drosophile, a ensuite été retrouvée grâce au séquençage chez divers organismes (BOURGEOIS et collab., 2020; LI et collab., 2018). Malgré tout, au début des années 2000 ont été découvertes quelques insertions supposées évoluer sous sélection positive (parce qu'associées à des signatures de sélection positive et/ou potentiellement adaptatives).

0.2.2.1 Quelques cas bien documentés

Au début des années 2000, DABORN et collab. (2002) démontrent que la résistance à l'insecticide DDT chez *D. melanogaster* est associée à un fort niveau de transcription du gène *Cyp6g1* et à la présence d'une insertion de l'élément Accord à proximité de ce gène (Table 1). En raison de son effet phénotypique cette insertion a été présumée adaptative. Deux ans plus tard, CATANIA et collab. (2004) confirment le rôle adaptatif de cette insertion en montrant qu'un fort balayage sélectif lui

est associé[1]. Depuis un nombre restreint d'insertions potentiellement adaptatives ont vu leur impact fonctionnel faire l'objet d'une validation expérimentale (Table 1). Dans chacun de ces cas l'insertion joue un rôle sur l'expression d'un gène proche. Les mécanismes derrière cet effet ne sont pas toujours élucidés mais pointent soit vers un rôle de séquences régulatrices portées par les ET (GUIO et collab., 2014; SCHLENKE et BEGUN, 2004), soit vers une altération de la stabilité de l'ARN par les ET (NIU et collab., 2019). Le niveau de preuve de l'impact fonctionnel de ces insertions est variable, de la comparaison du niveau d'expression du gène voisin entre individus avec ou sans insertion, à la transgénèse. Il est important de noter qu'il n'y a pas toujours de signature de sélection associée à ces insertions présumées adaptatives. Par exemple, dans leur étude chez *Capsella rubella* NIU et collab. (2019) considèrent l'insertion d'un Helitron à proximité du gène *FLC* comme potentiellement adaptative car le phénotype qu'elle entraîne est vraisemblablement adaptatif dans la région géographique où elle ségrège. On note que sur les sept cas d'insertions potentiellement adaptatives répertoriés Table 1, seuls trois gènes sont impliqués. Ceci pourrait être lié à un biais d'insertion des ET (SULTANA et collab., 2017), ou bien à des a priori lors de la recherche d'insertions potentiellement adaptatives.

0.2.2.2 Un nombre restreint d'études à l'échelle du génome

La découverte au début des années 2000 d'insertions potentiellement adaptatives a amené la question de la proportion d'insertions adaptatives par rapport au nombre total d'insertions ségrégeant dans les populations. Il est important de noter, que cette proportion ne reflète vraisemblablement pas le nombre d'événements d'insertions adaptatives par rapport au nombre d'événements d'insertion total. En effet, on s'attend à ce que la majorité des insertions disparaissent rapidement sous l'effet de la sélection purificatrice (BURT et TRIVERS, 2006). Afin d'apporter un début de réponse à la question de la proportion d'insertions adaptatives plusieurs scans génomiques à la recherche de telles insertions ont été menés (Table 2). Jusqu'à présent ces scans sont majoritairement limités aux espèces modèles. Certains scans sont réalisés sans a priori, alors que certaines études recherchent des insertions adaptatives dans des populations particulières. Ainsi, chez *D. melanogaster* les études se focalisent sur des insertions adaptatives avant ou pendant la sortie d'Afrique. Plusieurs méthodes sont utilisées pour discriminer les insertions adaptatives. Sont généralement utilisées, la recherche d'une forte différenciation génétique (ou simplement « différenciation » dans ce manuscrit) ainsi que l'association avec un balayage sélectif local. La différenciation génétique correspond à la variation des fréquences alléliques et elle peut être évaluée de différentes manières. Dans les études qui nous intérèssent celle-ci est évaluée 1) par simple soustraction de fréquences alléliques; 2) à partir de l'indice de fixation de Wright (F_ST)[2] 3) à partir de statistiques prenant en compte l'histoire démographique des populations (e.g. XtX ou eBp_{is}) (BOURGEOIS et collab., 2020; GAUTIER, 2015). De la même manière l'évaluation d'un balayage sélectif se fera selon différentes méthodes (VILLANUEVA-CAÑAS et collab., 2017). On note que certaines études prennent en compte le fait que l'insertion soit présente ou non dans une région de forte recombinaison. En effet, dans de telles régions, du fait de la recombinaison ectopique une insertion a moins de chance d'atteindre

[1] On parle de balayage sélectif lorsqu'un allèle favorable apparaît dans une population et que les variants génétiques à proximité sont mécaniquement emportés avec lui provoquant une baisse locale de la diversité génétique.

[2] L'indice de fixation de Wright, le F_ST correspond à la réduction d'hétérozygotie par rapport l'attendu, du fait de la structuration de la population totale en sous-populations (ST pour Sous-population/Total).

ET (Organisme)	Gène(s) impacté(s) (fonction)	Méthode de validation - Impact de l'insertion	Signature de sélection	Réf(s)
Accord (*D. melanogaster*)	*Cyp6g1* (Résistance aux insecticides)	Comparaison de lignées - Sur-expression du gène et résistance aux insecticides	Balayage sélectif	[1] [2]
Doc (*D. simulans*)	*Cyp6g1* (Résistance aux insecticides)	Comparaison de lignées - Sur-expression du gène	Balayage sélectif	[3]
Bari1 (*D. melanogaster*)	*Jheh2 - Jheh3* (Résistance aux insecticides)	EAS[1] et Introgression[2] - Sous-expression des gènes et résistance aux insecticides	Balayage sélectif	[4]
carb-TE (*B. betularia*)	*cortex* (Contrôle de la mélanisation)	Comparaison de génotypes - Sur-expression du gène	Balayage sélectif	[5]
Copia(s) (*A. thaliana*)	*FLC* (Contrôle de la floraison)	Comparaison d'accessions - Sous-expression du gène et floraison précoce	Aucune	[6]
Helitron (*C. rubella*)	*FLC* (Contrôle de la floraison)	Transgénèse - Sous-expression du gène	Aucune	[7]
Copia (*A. arenosa*)	*FLC* (Contrôle de la floraison)	Comparaison de populations - Sous-expression du gène et floraison précoce	Aucune	[8]

Table 1: **Description des insertions d'ET potentiellement adaptatives et dont l'impact fonctionnel a fait l'objet d'une validation expérimentale.** Pour chaque insertion sont mentionnés le type d'ET, l'espèce concernée, le gène impacté et sa fonction. La méthode de validation de l'impact fonctionnel de l'insertion, ainsi que l'impact constaté, sont aussi indiqués. Enfin la présence d'une signature de sélection associée à l'insertion est mentionnée.
Note.— [1] Expression Allèle Spécifique (ou *Allele Specific Expression* en anglais) (EAS) désigne une expérience consistant à tester une différence d'expression entre deux allèles d'un même gène. Ici il s'agit d'un allèle « avec insertion » et d'un allèle « sans insertion » chez des hybrides entre une lignée avec insertion et une lignée sans insertion. Un tel test contrôle pour un effet du fond génétique.
[2] L'introgression consiste à transférer par croisements une portion du génome, ici l'insertion étudiée, d'une lignée, accession, etc. a une autre.

une forte fréquence si elle n'est pas positivement sélectionnée (voir section 0.1.3.3). De plus, dans les régions de forte recombinaison une insertion a moins de chance d'avoir été entraînée par un variant adaptatif proche. Certains auteurs ne vont considérer dans leur scans que les ET qui sont généralement soumis à une sélection purificatrice forte, afin de limiter les risques d'une augmentation par dérive. La variabilité dans les méthodologies employées rend difficile la comparaison de résultats entre les études.

Organisme	Cible	Signatures de sélection	N Ins. (/Total)	Réf(s)
D. melanogaster	Adaptation pendant ou après la sortie d'Afrique	• Différenciation (freq.) + En région recombinante + D'une famille contre sel. (**13/13**) • Balayage sélectif (**5/13**) • Plus fréquente en climat temperé que tropical (**8/13**)	**13** (/902)	[9] (voir aussi [10])
D. melanogaster	Adaptation pendant ou après la sortie d'Afrique	• Fréquence plus forte qu'attendu selon son âge	**8** (/190)	[11]
A. thaliana	Adaptation parmi 201 accessions	• Balayage sélectif	**2** (/2311)	[12]
H. sapiens	Adaptation parmi 15 populations	• Fréquence plus forte qu'attendu (simulation évolution neutre)	**7** (/14384)	[13]
D. melanogaster	Adaptation parmi 60 populations	• Différenciation (fréq.) + En région recombinante (**300/300**) • Différenciation (F_ST) (**8/300**) • Balayage sélectif (**36/300**)	**300** (/1223)	[14]
A. carolinensis	Adaptation pendant la colonisation du Sud-Est des États-Unis depuis la Floride	Différenciation (XtX et eBp_{is}) + Balayage sélectif	**4** (/339149)	[15]

Table 2: **Description des scans génomiques à la recherche d'insertions d'ET adaptatives.** Pour chaque scan sont mentionnés: l'organisme étudié, la cible du scan, les signatures de sélection utilisées pour discriminer les insertions adaptatives et le nombre d'insertions adaptatives trouvées (en gras) par rapport au nombre total d'insertions étudiées.

0.2.3 Variation intra-génomique

Parce que conditionnée par des variables dont les valeurs peuvent varier au sein d'un même génome, il existe une variabilité intra-génomique de la dynamique des ET.

0.2.3.1 Taux de recombinaison

Un premier facteur qui va affecter la dynamique des ET localement est le taux de recombinaison. La propension des ET à promouvoir la recombinaison ectopique est attendue d'être plus forte dans les régions fortement recombinantes. Du fait, de l'impact négatif de la recombinaison ectopique sur la valeur sélective de l'individu, les ET devraient être plus fortement contre sélectionnés dans les régions fortement recombinantes. En accord avec cela, une corrélation négative existe entre le contenu en ET et le taux de recombinaison chez *D. melanogaster* et l'homme (Bartolomé et collab., 2002; Medstrand et collab., 2002). Cette corrélation n'a toutefois pas été retrouvée chez *A. thaliana* (Wright et collab., 2003). Chez le nématode *Caenorhabditis elegans* une corrélation positive a été

trouvée pour les ET de classe II mais pas pour les ET de classe I (DURET et collab., 2000).

0.2.3.2 Densité en gènes

Parce que les insertions proches des, ou dans les, gènes peuvent altérer leur fonction, une forte sélection purificatrice est attendue à l'encontre de ces insertions. En effet, chez une variété d'espèces, les ET sont moins présents dans les régions riches en gènes (BARTOLOMÉ et collab., 2002; MEDSTRAND et collab., 2002; WRIGHT et collab., 2003).

0.2.3.3 X vs Autosomes

Enfin, dans le cas des espèces où le chromosome X est hémizygote, les insertions récessives sont attendues être plus fortement contre sélectionnées sur le chromosome X, où le caractère récessif s'exprime, que sur les autosomes. Cependant, les études chez *D. melanogaster*, *C. elegans* et l'homme ne suggèrent pas un contenu global en ET inférieur sur le X (BOISSINOT et collab., 2006; CRIDLAND et collab., 2013; DURET et collab., 2000; KOFLER et collab., 2012). On note toutefois, des différences entre familles.

0.3 Les ET et l'outil bioinformatique

L'obtention des séquences d'ADN et d'ARN est primordiale en biologie. Dans les années 2000, l'avènement des techniques de séquençage dites NGS a permis une croissance exponentielle du nombre de séquences disponibles pour les biologistes. Cette explosion de la quantité de données a bouleversé le monde de la biologie, y compris celui des ET. De nouvelles techniques sont apparues encore récemment offrant de nouvelles perspectives.

0.3.1 De plus en plus de données et de nouvelles techniques

0.3.1.1 NGS & Explosion de la quantité de donnée

Le séquençage de première génération, le séquençage de type Sanger, a été élaboré dans les années 1970. Cette technique permet de séquencer des morceaux d'ADN d'environ 1000 nucléotides avec une grande fiabilité. En séquençant des morceaux d'ADN chevauchants il est alors possible de les assembler pour reconstituer la séquence génomique d'une espèce. C'est avec cette méthode que le génome humain a été initialement séquencé [1] (LANDER et collab., 2001). Ce projet a duré environ 10 ans. Cette méthode a permis l'assemblage du génome d'un petit nombre d'espèces modèles (ADAMS et collab., 2000; C. ELEGANS SEQUENCING CONSORTIUM, 1998), mais toujours à un rythme relativement lent et par le biais d'investissements financiers et humains très importants. Dans les années 2000 est arrivée une nouvelle technologie : le séquençage de

[1] Dans le cas du génome humain, du fait notamment de sa taille importante, la stratégie utilisée est en réalité un peu plus complexe. Le génome est d'abord découpé en grands segments, dont on connait la position relative grâce à une carte génétique, c'est-à-dire la représentation de la disposition de marqueurs sur les chromosomes. On séquence ensuite individuellement chacun de ses fragments avant de les assembler.

nouvelle génération (ou NGS). Il permet aujourd'hui de séquencer un génome pour moins de 1000 euros en quelques jours/semaines. L'émergence de cette technologie a été à l'origine d'une croissance exponentielle du nombre de génomes séquencés. À ce jour la banque de données de séquences du Centre Américain pour les Informations Biotechnologiques (ou *National Center For Biotechnology Information* en anglais) (NCBI) recense 14641 génomes Eucaryotes, 287040 génomes Procaryotes et 41523 génomes de virus (noa, c). Malgré sa rapidité et son faible coût, le séquençage de nouvelle génération présente des limitations. Les NGS séquencent de courts fragments d'ADN, de l'ordre de 100 pb. Ces séquences, appelées lectures (ou *reads* en anglais), sont relativement difficiles à assembler en segments plus grands. La présence d'ET dans les génomes notamment, constitue une limitation majeure à l'assemblage de génomes complets. En effet, parce que les ET font souvent plusieurs centaines voir plusieurs milliers de paires de bases et sont répétés, il est impossible d'établir correctement les jonctions de chaque insertion avec le génome de l'hôte (fig.A.1A). Même la dernière variante de NGS, le séquençage de lectures pairées (ou *paired-end reads sequencing* en anglais)(fig.A.1B), c'est-à-dire que chaque lecture est pairée avec une lecture distante de quelques centaines de paires de bases, n'a pas parfaitement résolu ce problème (RIUS et collab., 2016). L'apparition récente du séquençage longues lectures (ou *long-reads sequencing* en anglais) constitue toutefois un grand pas vers l'assemblage complet des génomes.

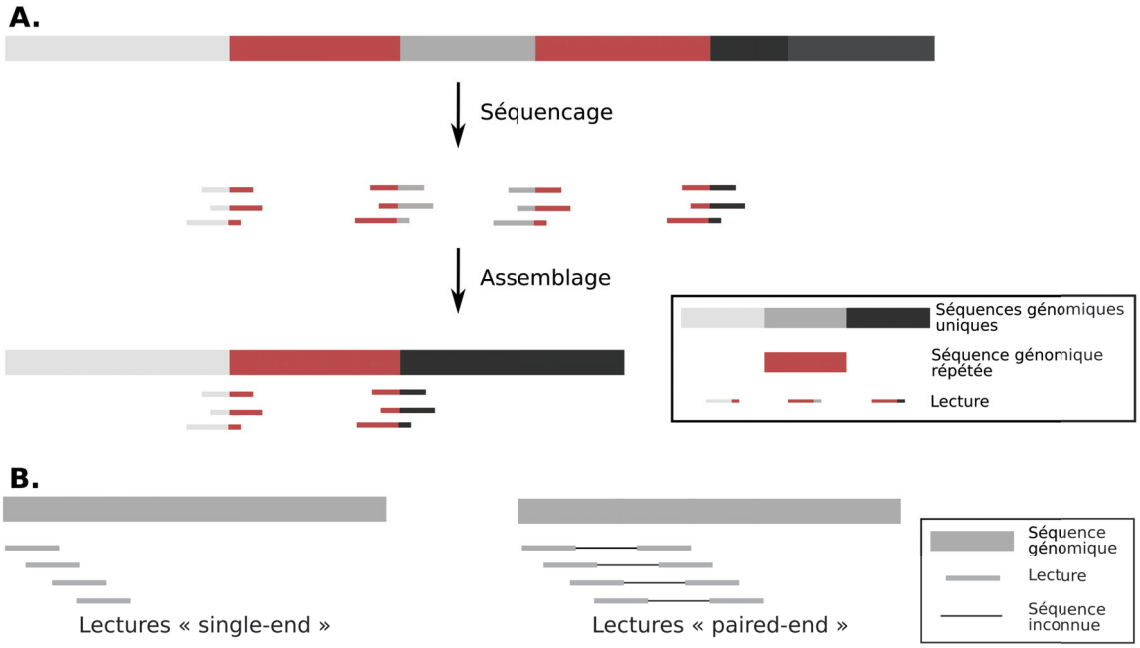


Figure 1: **Next Generation Sequencing (NGS): assemblage des séquences répétées et « single-end » vs « paired-end ». A.** Exemple de mésassemblage du à la présence de séquences répétées pour un génome séquencé avec les technologies NGS. **B.** Illustration des différences entre séquençage « single-end » et « paired-end »

0.3.1.2 Le séquençage longues lectures

Le séquençage longues lectures, développé dans les années 2010, permet de séquencer de plus grands fragments d'ADN et ainsi d'obtenir de plus grandes lectures (de l'ordre de 20 000 nucléotides) (RHOADS et AU, 2015). La plupart des ET étant de taille inférieure à 20 kb l'assemblage des séquences répétées est ainsi beaucoup moins problématique. Cette technique permet l'obtention

d'assemblages beaucoup moins fragmentés. Le taux d'erreur, c'est-à-dire la probabilité qu'un nucléotide soit mal lu, est cependant plus élevé qu'avec les NGS. De plus le coût du séquençage est plus important.

0.3.1.3 Le séquençage de « pools » ou Pool-Sequencing (PoolSeq)

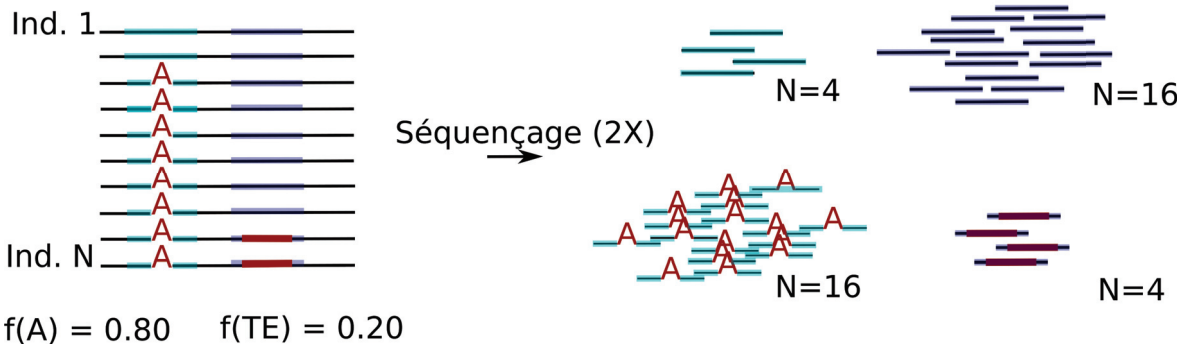


Figure 2: **Illustration du principe du séquençage de « pools » ou Pool-Sequencing (PoolSeq).** L'ADN d'un « pool » d'individus d'une population donnée est extrait puis séquencé. La proportion de lectures supportant tel ou tel variant est égale à la fréquence de ce variant dans la population

La nécessité d'évaluer le polymorphisme génétique dans les populations, c'est à dire les variations de séquences entre individus, à vu apparaitre le séquençage de « pools » ou Pool-Sequencing (PoolSeq) (SCHLÖTTERER et collab., 2014). Cette méthode permet d'évaluer à relativement faible coût les fréquences auxquelles ségrègent différents variants génétiques, SNP ou insertions d'ET par exemple, dans une population. Plutôt que de séquencer séparément un certain nombre d'individus de la population, ce qui est relativement onéreux, il s'agit d'extraire l'ADN d'un pool d'individus directement et de le séquencer (fig.2). La proportion de lectures supportant tel variant ou tel variant est alors supposée représenter la fréquence de ce variant dans une population (sous réserve d'un nombre d'individus séquencés et d'une profondeur de séquençage[1] suffisante notamment).

0.3.2 L'annotation des ET dans un génome

L'augmentation considérable du nombre de génomes séquencés a rapidement amené la question: « comment annoter les ET dans les génomes ? ». Une telle annotation est par exemple nécessaire pour estimer la quantité d'ET dans les génomes ou encore leur distribution intra-génomique.

0.3.2.1 La reconstruction de séquences consensus

La première étape dans l'annotation des ET d'un génome est la reconstruction de leurs séquences consensus. En effet, parce qu'après leur insertion les différentes copies d'un même ET voient leurs séquences diverger au gré des mutations, l'information de séquence de chaque ET est souvent

[1] La profondeur de séquençage, ou couverture d'un échantillon de séquençage, correspond au nombre de fois où chaque nucléotide est séquencé. Ainsi une profondeur de séquençage égale à un, indique que chaque base a été séquencée une fois. Une couverture égale à un est notée 1X, une couverture égale à 5 est notée 5X, etc.

résumée dans une unique séquence consensus. Ainsi la séquence consensus correspond à la séquence des positions les plus fréquentes à chaque position de l'alignement des copies. Il est à noter que cette reconstruction de séquences consensus n'est pas toujours nécessaire lors de l'étude d'espèces modèles telles que *D. melanogaster* ou *A. thaliana*. En effet, pour ces espèces les séquences consensus sont connues depuis longtemps et présentes dans des bases de données telles que RepBase ou Dfam. En faisant l'hypothèse que le contenu en ET est similaire entre espèces proches, on peut aussi utiliser l'information contenue dans ces bases de données pour l'étude d'espèces proches des espèces modèles. Pour les autres cas, il est possible d'obtenir des séquences consensus, soit directement à partir de lectures, soit à partir d'un assemblage.

À partir de lectures

Il existe trois outils principaux pour reconstruire des séquences consensus d'ET à partir de lectures de séquençage: RepeatExplorer, dnaPipeTE et RepARK. Ces trois outils utilisent le caractère répété des ET. Les deux premiers outils, RepeatExplorer et dnaPipeTE, ont un même principe de base: dans un échantillon de lectures tel que la couverture est inférieure à 1X, c'est-à-dire que chaque base du génome est présente moins d'une fois, seules les lectures de séquences répétées telles que les ET devraient être en quantité suffisante pour permettre un assemblage (GOUBERT et collab., 2015; NOVÁK et collab., 2010). Ces deux outils diffèrent légèrement, notamment par la méthode d'assemblage utilisée, mais semblent donner des résultats similaires (GOUBERT et collab., 2015). RepARK, le troisième outil, « découpe » les lectures en séquences plus courtes, de k nucléotides, appelées k-mer (KOCH et collab., 2014). Les k-mers les plus abondants sont alors assemblés en séquences consensus. Il est à noter qu'avec de telles méthodes il est difficile de reconstruire les consensus d'ET présents en faible nombre de copies.

Avec un génome assemblé

Plusieurs programmes bioinformatiques permettent l'obtention de séquences consensus à partir de génomes assemblés. Les deux principaux sont RepeatModeler et TEdenovo (FLUTRE et collab., 2011; FLYNN et collab., 2020). Ce dernier est inclu dans une suite appelée REPET qui permet aussi l'annotation des copies dans les génomes assemblés. RepeatModeler et TEdenovo sont en réalité des pipelines d'analyse c'est-à-dire un groupe de logiciels interdépendants. La première étape de TEdenovo, consiste en un alignement du génome avec lui-même après un découpage de ce dernier en séquences courtes. Les différentes homologies trouvées sont ensuite regroupées à l'aide des programmes RECON et/ou PILER et/ou GROUPER, et des séquences consensus sont formées (BAO et EDDY, 2002; EDGAR et MYERS, 2005; QUESNEVILLE et collab., 2003). Les auteurs du package REPET ont incorporé la possibilité d'ajouter aux séquences ainsi construites des séquences consensus générées à partir de RepeatScout et LTRharvest (ELLINGHAUS et collab., 2008; PRICE et collab., 2005). RepeatScout recherche des k-mers particulièrement abondants dans le génome assemblé et les étend en une séquence consensus plus longue. LTRharvest recherche des ET de classe I particuliers, appelés éléments à Longues Terminaisons Répétées (ou *Long Terminal Repeats* en anglais) (LTR). Pour cela il exploite leurs caractéristiques structurales et notamment la présence de séquences répétées inversées à leurs extrémités. Le fonctionnement général de RepeatModeler est similaire à celui de TEdenovo (FLYNN et collab., 2020). RepeatModeler utilise en combinaison RepeatScout et RECON ainsi que des programmes qui recherchent les éléments à LTR sur la base de caractéristiques structurales.

0.3.2.2 La classification des séquences consensus

La classification des séquences d'ET obtenues se fait en général par homologie ou via une étude de la structure de la séquence. L'annotation par homologie consiste à rechercher une homologie entre la séquence à annoter et les séquences déjà annotées présentes dans les bases de données. Cette recherche d'homologie peut se faire à l'aide d'algorithmes tel que Blast (ALTSCHUL et collab., 1990). Il est aussi possible d'utiliser des critères structurels pour annoter les séquences comme implémenté dans la suite REPET (FLUTRE et collab., 2011; HOEDE et collab., 2014).

0.3.2.3 L'annotation des copies dans les génomes assemblés

L'approche la plus classiquement utilisée pour annoter les copies d'ET dans un génome assemblé consiste à utiliser le programme RepeatMasker. À partir d'une banque de consensus d'ET, RepeatMasker annote chaque copie du génome, i.e. taille et position, par homologie de séquences. La suite REPET dispose de son propre programme d'annotation des copies appelé TEannot (FLUTRE et collab., 2011). Avec des contrôles supplémentaires, TEannot réduit le nombre de faux positifs et regroupe les fragments d'une même copie qui peuvent être séparés dans le génome à cause d'insertions ultérieures dans l'élément. Il est aussi possible de regrouper les fragments d'une même copie à partir d'une annotation de RepeatMasker en utilisant le programme OneCodeToFindThemAll (BAILLY-BECHET et collab., 2014).

0.3.3 L'étude du polymorphisme d'insertion

L'évaluation de la présence des différentes insertions dans les populations, et surtout l'estimation de leurs fréquences, est capitale afin de mieux comprendre la dynamique des ET. Il est relativement aisé, à partir de lectures issues des NGS, d'évaluer la présence de petits variants de types SNP/InDel. En effet, l'alignement des lectures sur un génome de référence puis l'étude des positions variables dans l'alignement permet la détection de tels variants. Si il s'agit d'un échantillon de type PoolSeq (voir section 0.3.1.3), alors la fréquence d'un variant peut-être estimée comme le simple ratio du nombre de lectures avec ce variant et du nombre total de lectures (fig.2). Dans le cas des ET la situation est légèrement plus complexe. Comme pour l'assemblage des génomes, le problème réside encore un fois dans le fait que la taille des ET est supérieure à la taille des lectures obtenues par les NGS. De ce fait, il est difficile à partir du simple alignement des lectures courtes sur un génome de référence d'évaluer la présence et la position des insertions. Dans les années à venir ce problème sera potentiellement contourné par l'utilisation systématique des technologies de séquençage longues lectures, y compris pour le séquençage de type PoolSeq. Pour l'instant toutefois, du fait notamment du coût de ces technologies, de tels jeux de données restent rares. Des outils sont donc encore développés pour estimer malgré tout le polymorphisme d'insertions à partir de séquences courtes (VENDRELL-MIR et collab., 2019). Il existe de nombreuses différences dans le fonctionnement de ces divers logiciels, mais ils exploitent généralement la propension des insertions à générer des alignements particuliers (fig.3). Ces logiciels se basent notamment sur la présence de lectures « coupées » (ou *split reads* en anglais), ou de lectures discordantes (ou *discordant reads* en anglais), après alignement sur un génome de référence et/ou une banque de données

d'ET. On parle de lecture coupée quand deux portions d'une même lecture s'alignent à des endroits différents. Ainsi le logiciel TIDAL (RAHMAN et collab., 2015) considère comme une signature de la présence d'une insertion absente dans le génome de référence une lecture coupée dont une portion s'aligne sur le génome et l'autre s'aligne sur une banque de données d'ET (fig.3A). Ce même logiciel considère comme une signature de l'absence d'une insertion présente dans le génome de référence une lecture coupée dont une portion s'aligne d'un côté de l'insertion et l'autre portion s'aligne de l'autre côté de l'insertion (fig.3B). On parle de lecture discordante quand les deux lectures d'une paire s'alignent à des endroits différents. Le logiciel TEMP (ZHUANG et collab., 2014) considère comme une signature de la présence d'une insertion absente dans le génome de référence, une lecture discordante telle qu'une des lectures s'aligne sur un ET et l'autre sur le génome (fig.3C). A l'inverse une lecture discordante telle qu'une des lectures s'aligne d'un côté de l'insertion ET l'autre de l'autre côté est considérée comme une signature de l'absence d'une insertion présente dans le génome de référence (fig.3B). Le logiciel PoPoolationTE2 (KOFLER et collab., 2016) utilise aussi les lectures discordantes, mais contrairement aux autres logiciels, il utilise un même pipeline pour la détection des insertions absentes ou présentes dans le génome de référence. Il est à noter que certains programmes ne permettent pas d'évaluer le polymorphisme des insertions absentes dans le génome de référence (e.g. T-LEX3) (BOGAERTS-MÁRQUEZ et collab., 2020). Seuls quelques logiciels (e.g. TEMP, PoPoolationTE2, T-LEX3) permettent l'estimation des fréquences des différentes insertions. Pour ce faire ces logiciels divisent le nombre de signatures de présence par l'ensemble des signatures. Toutes ces méthodes sont connues pour être imparfaites (VENDRELL-MIR et collab., 2019). Il est notamment quasiment impossible d'évaluer le polymorphisme d'insertion dans la régions où la densité d'insertions est grande. De plus, ces méthodes sont connues pour donner des résultats extrêmement différents sur des jeux de données identiques (LERAT et collab., 2019). Des outils ont donc été développés pour tester ces méthodes, notamment des outils permettant de simuler des jeux de données de PoolSeq avec un polymorphisme d'insertions connu (KOFLER, 2018).

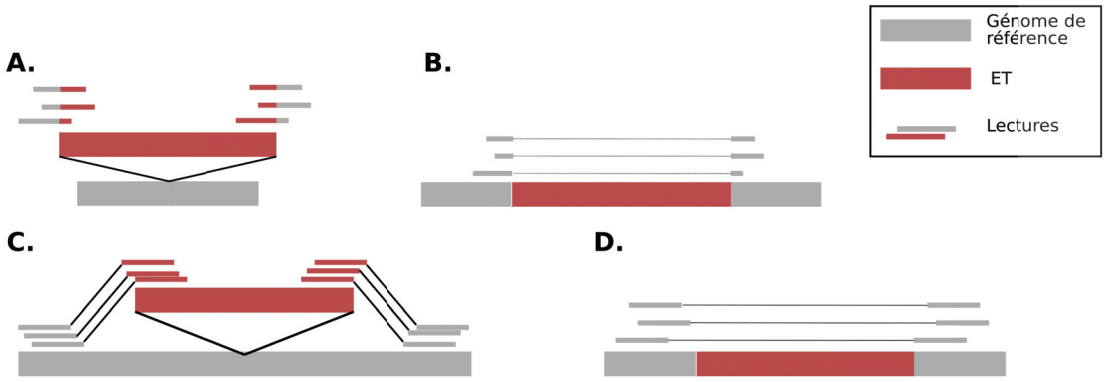


Figure 3: **Singularités d'alignements liées aux insertions d'ET. A.** Lectures « coupées » générées par la présence dans l'échantillon séquencé d'une insertion absente dans le génome de référence. **B.** Lectures « coupées » générées par l'absence dans l'échantillon séquencé d'une insertion présente dans le génome de référence. **C.** Lectures discordantes générées par la présence dans l'échantillon séquencé d'une insertion absente dans le génome de référence. **D.** Lectures discordantes générées par l'absence dans l'échantillon séquencé d'une insertion présente dans le génome de référence.

0.3.4 L'étude de la quantité d'ET dans un génome

Il est possible de dériver des estimations de la quantité d'ET dans un génome, à la fois de l'étude du polymorphisme d'insertions (voir section 0.3.1.3), et de l'annotation des copies dans un génome assemblé (voir section 0.3.2). Dans le premier cas, la somme des fréquences d'insertions permet d'obtenir le nombre d'insertions par génome haploïde. Dans le deuxième cas, il est possible d'obtenir un nombre de fragments/copies d'ET, et à partir de leur taille un pourcentage d'occupation de l'assemblage [1]. Ces deux méthodes ont des limites. En effet, l'assemblage génomique est souvent incomplet, notamment en ce qui concerne les copies d'ET, et ce même avec un séquençage longues lectures (PARIS et collab., 2020). De plus, les outils dédiés à l'étude du polymorphisme d'insertions présentent certains biais, tels que la difficulté d'estimer le polymorphisme dans les régions trop denses en ET. On utilise donc souvent une autre méthode pour estimer le contenu en ET à partir d'un échantillon de séquençage lectures courtes. Cette méthode ne nécessite pas de génome assemblé. Il s'agit dans un premier temps d'aligner les lectures sur une banque de séquences d'ET (ou banque d'ET)(fig.5). La quantité d'ET dans l'échantillon peut ensuite être estimée de deux façons différentes. Dans le premier cas, on assimile la proportion de lectures qui s'alignent sur chaque ET de la banque au pourcentage génomique qu'il occupe. Cette méthode est implementée dans les programmes dnaPipeTE et deviaTE (GOUBERT et collab., 2015; WEILGUNY et KOFLER, 2019). Elle a aussi été utilisée par LERAT et collab. (2019) après un alignement des lectures sur une banque d'ET à l'aide de l'outil RepeatMasker. La deuxième façon d'estimer le contenu en ET à partir d'un alignement de lectures courtes sur une banque d'ET nécessite que les lectures aient aussi été alignées sur des régions non répétées du génomes (telles que des gènes en copie unique). Alors, à partir de la couverture de chaque ET, et des régions non répétées, et en faisant le ratio de ces deux valeurs, il est possible d'obtenir le nombre d'insertions par génome haploïde. Cette technique est implementée dans deviaTE (WEILGUNY et KOFLER, 2019).

[1] Par abus de langage, on parle souvent de pourcentage d'occupation du génome ou de pourcentage génomique

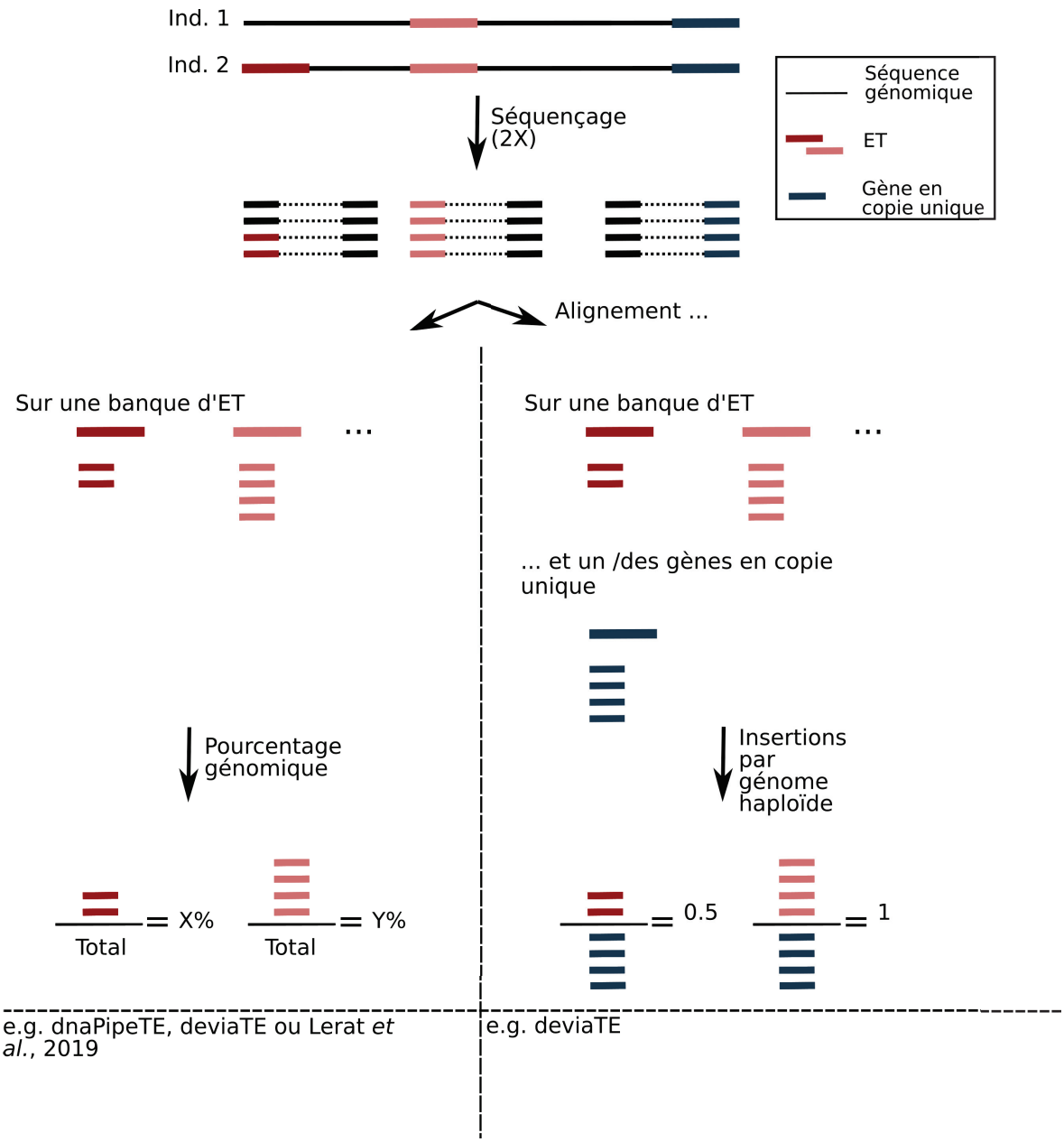


Figure 4: **Estimation du contenu en ET génomique avec des lectures courtes.** Les lectures sont alignées soit sur une banque d'ET uniquement soit sur une banque d'ET et un/des gènes en copie unique. Dans le premier cas, le pourcentage génomique d'un ET est estimé à partir de la proportion de lectures qui s'alignent sur cet ET. Dans le deuxième cas le nombre de copies par génome haploïde est obtenu en faisant le ratio de la couverture de l'ET et de la couverture de régions non répétées (e.g. gènes en copie unique.)

0.4 La drosophile à ailes tachetées *D. suzukii*

À ce jour, une des rares espèces qui dispose d'un assemblage obtenu à partir de lectures longues, et de données de séquençage de type PoolSeq pour plusieures populations de part le monde, est la Drosophile à ailes tachetées. La Drosophile à ailes tachetées *Drosophila suzukii* (MATSUMURA, 1931), ou *the spotted-wing fly* en anglais, est décrite pour la première fois en 1931 dans l'ouvrage *6000 insectes de l'Empire Japonais illustrés* par Shōnen Matsumura. C'est un Diptère (insecte) de la famille des Drosophilidae. L'analyse de divergence nucléotidique suggère que son espèce soeur est *Drosophila subpulchrella* (HAMM et collab., 2014), et leur plus proche parent *Drosophila biarmipes*

avec une date de divergence estimée à 4 millions d'années (ROTA-STABELLI et collab., 2020). *D. suzukii* est relativement proche de l'espèce modèle *D. melanogaster* dont elle aurait divergé il y a environ 8 millions d'années (ROTA-STABELLI et collab., 2020).

0.4.1 Biologie descriptive

0.4.1.1 Morphologie

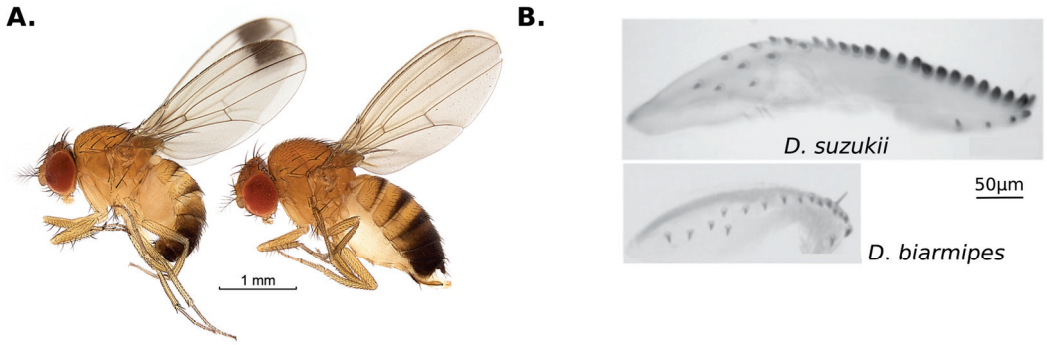


Figure 5: **Morphologie de *D. suzukii*. A.** Cliché d'un individu mâle (à gauche) ET d'un individu femelle (à droite) de *D. suzukii* ©Shane F. McEvey CC BY 4.0. Cette image montre notamment la présence d'une tache alaire chez le mâle. **B.** Morphologie comparée de l'ovipositeur de *D. suzukii* et *D. biarmipes*, image issue de ATALLAH et collab. (2014). On remarque la longueur de l'ovipositeur de *D. suzukii* ainsi que la présence de dents sur la moitié de sa longueur.

Comme son nom vernaculaire l'indique la Drosophile à ailes tachetées se distingue par la présence de tâches noires sur les ailes de ses mâles (fig.5A). *D. suzukii* partage en réalité cette caractéristique avec ses proches parentes *D. biarmipes* et *D. subpulchrella*. Des espèces plus éloignées présentent elles aussi des tâches alaires, une caractéristique co-évoluant avec la parade nuptiale et qui aurait été acquise indépendamment et perdue plusieures fois au sein du genre *Drosophila* (KOPP et TRUE, 2002; PRUD'HOMME et collab., 2006). *D. suzukii* possède une seconde caractéristique morphologique remarquable, qu'elle partagerait avec *D. subpulchrella* seulement, la présence d'un long ovipositeur en dents de scie (fig.5B). C'est cet ovipositeur particulier qui lui permet de pondre dans des fruits avant leur pourrissement. Il est à noter que, avec la diminution des températures saisonnières, la cuticule se mélanise et les ailes s'allongent (SHEARER et collab., 2016). Le phénotype engendré est parfois qualifié de morphe d'hiver (à l'opposé du morphe d'été).

0.4.1.2 Cycle de vie

Le cycle de vie de *D. suzukii* dure environ deux semaines à 22°C (EMILJANOWICZ et collab., 2014; TOCHEN et collab., 2014), avec toutefois des variations importantes selon la température. La femelle pond ses œufs dans les fruits mûrs ou pourrissant. Ceux-ci éclosent après ~1.4 jours à 22°C. La larve se nourrit du fruit, et atteint le stade pupal après ~6 jours. Le stade pupal dure environ ~6 jours. Les femelles ne pondent qu'un à cinq jours après émergence de la pupe (HAMBY et collab., 2016). À d'abord lieu un accouplement, lui-même précédé par une parade nuptiale executée par le mâle (REVADI et collab., 2015). Un adulte peut vivre de deux à 25 semaines en laboratoire (SHEARER et collab., 2016). *D. suzukii* est supposée passer l'hiver comme adulte, notamment

dans les forêts où elle survivrait sous les feuilles (STOCKTON et collab., 2019). Pendant l'hiver, *D. suzukii* entre en diapause, un arrêt développemental associé à une forte résistance au froid en lien avec le développement du morphe d'hiver (TOXOPEUS et collab., 2016). Les femelles ont notamment des ovaires atrophiés et un niveau d'expression changé de plusieurs gènes impliqués dans la reproduction et la réponse au stress.

0.4.1.3 Ecologie

Habitat & hôtes

Les larves de *D. suzukii* émergeant quasi-exclusivement de fruits (MITSUI et collab., 2010), la présence de *D. suzukii* est conditionnée à la présence de plantes hôtes, notamment au sein de cultures. Parmi ces plantes on trouve notamment une large variété de fruits rouges (framboises, cerises, myrtilles, canneberge, etc.) https://www.cabi.org/...tohostsOrSpeciesAffected, mais aussi les pêches, les prunes ou le raisin (WALSH et collab., 2011). La présence de *D. suzukii* semble liée à celle des forêts (GRASSI et collab., 2011), celles-ci apportant potentiellement les conditions nécessaires pour survivre la saison hivernale. *D. suzukii* tolère une large variété de climat.

Dispersion naturelle

Une étude de capture-marquage-recapture suggère une faible capacité de dispersion chez *D. suzukii*, avec une distance journalière parcourue inférieure à 100m en l'absence de vent dominant (VACAS et collab., 2019).

Ennemis naturels

Le répertoire des espèces invasives recense 15 ennemis naturels de *D. suzukii* : huit parasites, cinq prédateurs, et deux champignons pathogènes (https://www.cabi.org...tonaturalEnemies). Parmi les espèces parasites on trouve notamment les guêpes parasitoïdes des genres Ganaspis et Leptopilina qui pondent leurs œufs dans les larves de Drosophile (CINI et collab., 2012). Ces espèces, ainsi que les deux champignons pathogènes, sont actuellement envisagés comme potentiels agents de contrôle des populations de *D. suzukii* (NARANJO-LÁZARO et collab., 2014).

0.4.1.4 Génétique

Le caryotype de la Drosophile à ailes tachetées comprend 3 paires d'autosomes, dont 2 métacentriques (chromosome 2 et 3) et un petit chromosome (chromosome 4) (DENG et collab., 2007). A cela s'ajoutent un chromosome X acrocentrique et un chromosome Y télocentrique. Ce caryotype est proche de celui de l'espèce modèle *D. melanogaster*, le chromosome 4 est toutefois plus grand chez *D. suzukii* et le chromosome Y est télocentrique (et non pas métacentrique comme chez *D. melanogaster*). La taille du génome de *D. suzukii*, estimée par cytométrie en flux pour deux lignées issues de populations invasives, est d'environ 330 Mb (HJELMEN et collab., 2019; SESSEGOLO et collab., 2016). Cette valeur, la deuxième plus haute du genre après celle de *D. virilis*, est associée à une importante proportion d'ET dans le génome (33%) (SESSEGOLO et collab., 2016).

0.4.2 Une espèce invasive

La Drosophile à ailes tachetées est aujourd'hui considérée comme à l'origine d'une invasion biologique parmi les plus importantes, tant par sa rapidité et son succès, que par la menace qu'elle représente pour l'économie. En effet, en quelques années la présence de *D. suzukii*, d'abord limitée à l'Asie, s'est étendue à l'Amérique du Nord, l'Amérique du Sud, l'Afrique et l'Europe https://www.cabi.org/...todistribution. Son invasion représente une menace prise très au sérieux pour la culture et le commerce de fruits comestibles.

0.4.2.1 L'invasion

Les routes d'invasion de *D. suzukii* ont été retracées au moyen de méthodes bayésiennes, et à partir de données de polymorphisme des microsatellites (25 loci pour 23 populations), par FRAIMOUT et collab. (2017)(fig. 6). L'aire de répartition « pré-invasion » inclut l'Asie, de l'Inde à l'ouest au Japon à l'est, et du sud de la Thaïlande au nord de la Chine. L'espèce aurait d'abord colonisé l'île d'Hawaï depuis le Japon dans les années 1980. *D. suzukii* aurait ensuite été introduite aux alentours de 2008 au Etats-Unis d'Amérique depuis le Japon et le Sud de la Chine, et à la même date en Europe depuis le Nord de la Chine. La colonisation du Brésil depuis les Etats-Unis et de l'île de la Réunion depuis l'Europe ce seraient faites plus tardivement (~2013). Les populations invasives étudiées ici auraient toutes subies des goulots d'étranglement (ou *bottlenecks* en anglais), c'est-à-dire des réductions de taille des populations, au moment de l'invasion. On note une intensité variable pour ces goulots d'étranglement, le plus sévère ayant eu lieu à Hawaï. La carte des routes d'invasions reste aujourd'hui incomplète et l'origine des populations de *D. suzukii* observées en Russie ou au Maroc par exemple reste inconnue https://www.cabi.org/...todistribution. En raison des faibles capacités de dispersion de *D. suzukii*, la piste d'une dispersion passive liée à l'homme est privilégiée pour cette invasion.

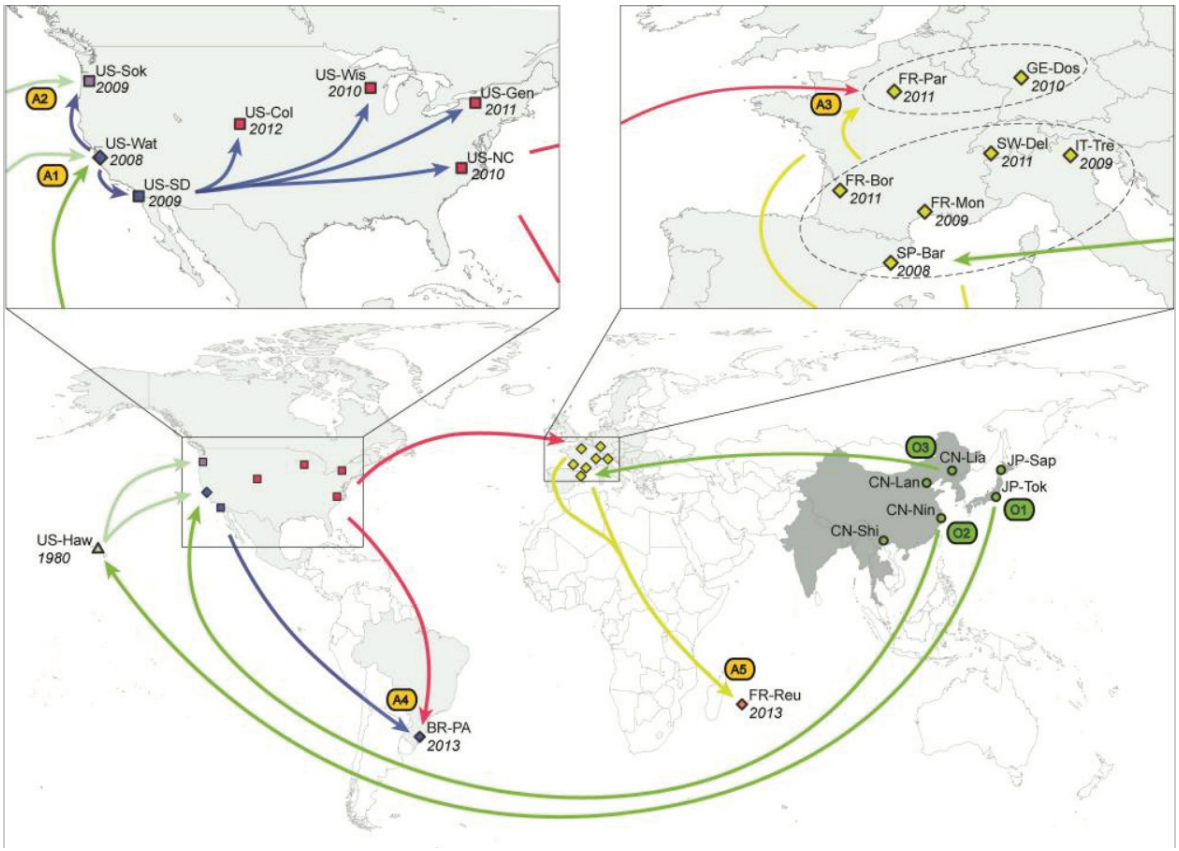


Figure 6: **Scénario d'invasion de *D. suzukii* inféré à partir de données de microsatellites** (FRAIMOUT et collab., 2017). La partie en gris foncé correspond à l'aire d'origine et la partie en gris claire correspond à l'aire envahie. Chaque symbole coloré indique un des sites où des individus ont été échantillonnés et génotypés. Les cercles indiquent des échantillons de l'aire d'origine. Les carrés, losanges et triangles correspondent à des échantillons de l'aire envahie. Les losanges indiquent un goulot d'étranglement faible, les carrés un goulot d'étranglement modéré et les triangles un goulot d'étranglement sévère. La couleur des formes correspond aux groupes génétiques inférés. La date indique l'année de première observation. Les flèches indiquent les routes d'invasion probables. A1-A5: les évenements de mélange génétique. O1–O3: les sources les plus probables pour les premières introductions.

0.4.2.2 Menaces pour l'économie

Les dommages causés par les larves de la drosophile à ailes tachetées sur les fruits cultivés les rendent impropres à la vente et à la consommation (BOLDA et collab., 2010). Les coûts économiques associés sont encore relativement peu connus, les rares études font toutefois état de pertes importantes. L'une d'entre elles fait état de pertes maximales, par rapport au rendement maximal, de 40% pour les cultures de myrtilles, 50% pour les cultures de mûres et de framboises et de 33% pour les cerises (BOLDA et collab., 2010). En extrapolant ces pertes à l'ensemble de la production des États de Californie, de l'Oregon et du Wisconsin les auteurs estiment une perte économique pouvant atteindre 388 millions de dollars [1]. Un sondage mené auprès de 82 cultivateurs de framboises dans l'état du Minnesota estime une perte de rendement médian de 20% en 2017 (DIGIACOMO et collab., 2019). Après application de ce chiffre aux rendements des années précédentes, les auteurs concluent à une perte totale potentielle de 2.36 millions de dollars pour les cultivateurs de framboises du

[1] On soulignera une totale opacité quand à la méthode d'obtention des pourcentages de perte évoqués, ou encore l'absence de référence dans cet article massivement cité. Je citerai donc encore une fois D. Raoult: « Vous savez je vous suggère de vous méfier, c'est seulement quand on aura les comptes qu'on saura, et les comptes on les verra bien ».

Minnesota sur un an.

0.4.2.3 Les clés du succès ?

Le succès invasif de la Drosophile à ailes tachetées réside vraisemblablement dans un certain nombre de traits énoncés précédemment. La ponte à l'intérieur de petits fruits transportés de manière régulière à l'international notamment permet un transport passif sur de longues distances et qui a peu de chance d'être détecté (GIPPET et collab., 2019). La possession d'un long ovipositeur en dents de scie, et la possibilité associée de pondre dans des fruits mûrs, offre à *D. suzukii* une niche écologique libre de toute compétition avec les autres Drosophiles, qui pondent elles dans des fruits pourrissants uniquement (ATALLAH et collab., 2014).

La contribution du processus adaptatif au succès invasif a notamment été étudiée par OLAZCUAGA et collab. (2020). L'étude, à l'échelle du génome, de l'association des fréquences des variants nucléotidiques avec le statut d'invasion des populations a permis de mettre en évidence un petit nombre de régions génomiques potentiellement impliquées dans le processus invasif (environ 200 régions). Plus particulièrement deux gènes, *RhoGEF64C* et *cpo*, contenaient des polymorphismes nucléotidiques significativement associés avec le statut invasif sur les deux principales routes d'invasion suivies par *D. suzukii*. Si la fonction de *RhoGEF64C* reste peu étudiée, des variations génétiques de *cpo* ont été associées à des variations du phénotype de diapause chez des populations de *D. melanogaster* (SCHMIDT et collab., 2008).

La plasticité phénotypique, qui correspond au développement d'un individu vers un phénotype en réponse aux facteurs de l'environnement, mais sans modification de son génotype, a été également proposée comme une des clés du succès de *D. suzukii* (LITTLE et collab., 2020). Selon cette hypothèse, on peut attendre 1) une plasticité plus importante chez les espèces invasives que chez les espèces non invasives, 2) au sein des espèces invasives une plasticité plus importante des populations invasives que des populations natives. Si plusieurs études évaluent la plasticité de divers traits chez *D. suzukii* (HAMBY et collab., 2016; SHEARER et collab., 2016; STOCKTON et collab., 2018), les comparaisons avec des espèces proches et non invasives, ou les comparaisons entre populations invasives et natives, sont rares. La morphologie des ailes (liée à la capacité de dispersion de l'animal)(FRAIMOUT et collab., 2018), la taille de la tache alaire (potentiellement liée au succès reproducteur)(VARÓN-GONZÁLEZ et collab., 2020), et la taille et la forme de l'ovipositeur (VARÓN-GONZÁLEZ et collab., 2020), ont été comparés entre populations invasives et natives. L'ensemble de ces traits s'est avéré être plastique vis à vis de la température de développement, mais les résultats ne suggèrent pas une plus grande plasticité chez les populations invasives.

0.5 Objectifs de la thèse

Le projet de ma thèse était d'étudier la dynamique des ET chez une espèce invasive, *Drosophila suzukii*, et de tester les hypothèses portant sur l'écologie des espèces invasives et le contenu en ET. Cela m'a amené à étudier de façon approfondie le contenu et la dynamique des ET dans cette espèce à deux échelles de temps, entre populations et entre espèces.

Drosophila suzukii est une excellente espèce pour étudier la dynamique des ET. Première-

ment, *D. suzukii* est une proche parente de la très étudiée *D. melanogaster*. Ceci permet d'utiliser l'homologie de séquence avec *D. melanogaster* pour l'annotation à la fois des gènes et des ET. Cette proximité avec une espèce modèle permet aussi de bénéficier d'outils bioinformatiques développés et testés chez une espèce avec des caractéristiques similaires (FLUTRE et collab., 2011; KOFLER et collab., 2016). Deuxièmement, depuis environ 1980, *D. suzukii* est impliquée dans un processus d'invasion biologique qui offre une remarquable opportunité d'étudier les effets combinés d'un changement d'environnement et des changements démographiques sur la dynamique des ET (FRAIMOUT et collab., 2017). Troisièmement, la comparaison du contenu en ET entre *D. suzukii* et espèces proches, montre une augmentation rapide du contenu en ET chez *D. suzukii* (SESSEGOLO et collab., 2016), et suggère qu'un, ou des événements majeurs, aient affecté la dynamique des ET chez *D. suzukii* au cours des derniers millions d'années.

Le premier objectif était d'esquisser un portrait général des ET chez la Drosophile. Pour ce faire une revue bibliographique a été réalisée et publiée dans la revue Mobile DNA. Un grand nombre de sujets sont abordés, de la description des principaux ET du genre *Drosophila*, à la génomique des populations en passant par les défenses de l'hôte. Une attention particulière a été apportée aux récentes avancées mais aussi aux aspects historiques. Les apports du modèle Drosophile à la recherche sur les ET sont mis en avant. À ce titre les travaux sur l'espèce modèle *D. melanogaster* sont évidemment cités, mais aussi les travaux chez les autres espèces afin de souligner de potentielles variations interspécifiques.

Le deuxième objectif était l'étude de la dynamique des ET dans les populations de *D. suzukii*. Ce travail a été réalisé à partir d'un génome récemment assemblé en lectures longues et de données de séquençage « poolé » de 22 populations. Au préalable un portrait des ET chez *D. suzukii*, du contenu en ET à leur activité, en passant par leur distribution génomique, a été réalisé. Le caractère envahissant de *D. suzukii* a ensuite permis d'évaluer un potentiel impact de l'environnement et de la démographie sur le contenu en ET. Enfin, le potentiel adaptatif des ET a été estimé.

Le troisième objectif était l'étude de la dynamique des ET chez l'espèce *D. suzukii* au cours des derniers millions d'années, à la recherche notamment d'une explication à l'important contenu en ET observé dans son génome. Les données de séquençage génomique de *D. suzukii* et sept espèces proches de *D. suzukii* ont permis de mettre en évidence et caractériser une importante prolifération d'ET commencée il y a environ 4 millions d'années. La temporalité de la prolifération a été étudiée, mais aussi de potentielles variations de dynamique entre ET. Les différences de prolifération entre les différentes régions ont été étudiées. Enfin, ce jeu de données à permis de tester si un relâchement de la sélection était associé à la prolifération d'ET observée.

Enfin, au cours de ma de thèse j'ai eu l'occasion de collaborer à trois projets. Le premier projet portait sur l'apport du séquençage grandes lectures à l'étude du contenu en ET chez la Drosophile et à l'étude de la dynamique d'invasion des ET (ce travail a été publié dans la revue Cells, voir Annexe 1). Ma contribution à ce travail consistait en l'étude du contenu en ET à l'aide de lectures courtes de séquençage de 8 lignées de Drosophile afin de valider l'utilisation d'assemblages dérivés de lectures longues pour l'étude du contenu en ET. Le deuxième projet s'inscrivait dans le cadre du projet de thèse de P. Marin (coencadré par C. Vieira et P. Gibert (LBBE)). Il s'agissait d'étudier les réponses phénotypiques et transcriptomiques de 3 lignées de *D. suzukii* au stress oxydatif. J'ai notamment étudié le polymorphisme d'insertion dans trois lignées de *D. suzukii* à l'aide de

données de séquençage. Les données obtenues suggèrent notamment une dépletion des ET au niveau des gènes répondant au stress oxydatif (ce travail a été soumis à la revue Genome Biology and Evolution, voir Annexe 2). Le troisième projet s'inscrivait lui aussi dans le cadre d'un projet de thèse, celui de Cecilia A. Banho (coencadré par C. Vieira et C Carareto (UNESP, Brésil)). Cette étude testait l'hypothèse que le phénotype d'un individu hybride soit associé au niveau de dérégulation de l'expression de ses gènes. Les résultats obtenus montrent notamment que le degré de dérégulation de l'expression des gènes est plus fort pour des hybrides sans motilité du sperme. Pour ce projet j'ai participé à la conception des scripts bioinformatiques nécessaires à l'analyse de données RNAseq (ce projet a été soumis à la revue Scientific Report, voir Annexe 3).

1 Transposable Elements in Drosophila

Vincent Mérel[1], Matthieu Boulesteix[1], Marie Fablet[1], Cristina Vieira[1].

Corresponding author: Cristina Vieira - cristina.vieira@univ-lyon1.fr

Affiliations:

1: Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, F-69622 Villeurbanne, France

Sommaire

1.1 Avant-propos

Ce travail a été réalisé suite à une invitation par la revue Mobile DNA a participé à la série "Transposable Elements in Model Organisms". Cette revue offre une vision globale des ET chez la Drosophile, avec une considération pour les aspect historiques mais aussi les récentes avancées. La version définitive et revue par les pairs de cet article est aussi disponible ici: https://mobilednajournal...0213-z.

Mérel et al. Mobile DNA (2020) 11:23
https://doi.org/10.1186/s13100-020-00213-z

Mobile DNA

REVIEW Open Access

Transposable elements in *Drosophila*

Vincent Mérel, Matthieu Boulesteix, Marie Fablet and Cristina Vieira*

Abstract

Drosophila has been studied as a biological model for many years and many discoveries in biology rely on this species. Research on transposable elements (TEs) is not an exception. *Drosophila* has contributed significantly to our knowledge on the mechanisms of transposition and their regulation, but above all, it was one of the first organisms on which genetic and genomic studies of populations were done. In this review article, in a very broad way, we will approach the TEs of *Drosophila* with a historical hindsight as well as recent discoveries in the field.

Keywords: Population genomics, *Drosophila*, intra and interspecific TE diversity, epigenetics

Background

A few words about Transposable Elements

Transposable elements (TEs) are selfish genetic elements that are able to multiply in a genome by copying themselves to other locations. This particular property allows them to persist and multiply in populations without the need of providing any advantage to the host [1–3]. Discovered in maize in the late 1940's by Barbara McClintock, they were understudied for decades [4, 5]. With the advent of molecular biology, notably their use for genetic engineering, an enormous amount of work has been done on TEs. The first sequencing projects stimulated the interest in these sequences, as they underscored their ubiquitous character. Indeed, TEs are found in virtually all eukaryotic species investigated so far [6–9]. They may represent up to 80% of a genome, as in Maize [10]. Additionally, one may expect these large elements, up to 20 kb, possessing coding sequences, regulatory sequences, and a unique epigenetic profile, to produce large-effect mutations [11, 12]. Actually, TEs have been shown to profoundly impact not only genomes, from chromosomal rearrangements to genome size, but also individuals, from deleterious to adaptive effects. Like many other research topics in biology, research on TEs owes much to *Drosophila*.

A few words about *Drosophila*

The *Drosophila* genus is estimated to include several thousand species [13] sharing their most recent common ancestor ~25-40 My ago [14]. So far, ~1500 drosophilid species have been described. The most extensively studied *Drosophila* species is, by far, *Drosophila melanogaster*. Originating from Sub-Saharan Africa, it has colonized all continents, except for Antarctica, as a human commensal [15, 16]. During the last 15,000-20,000 year, it expanded its range to Europe and Asia and was only recently introduced to Australia and the Americas (~200 years ago) [17]. *D. melanogaster* is raised in the lab since the beginning of the XXth century [16, 18]. Easy to maintain and having a short generation time, this species has been extensively studied since then. Nowadays, a search for the terms "*Drosophila*" and "*melanogaster*" on *pubmed* returns approximately 55,000 references, with more than 2000 published in 2018.

A great number of genetic tools, such as genetic transformation vectors using TEs, and the P-element in particular [19], the GAL4/UAS system to study gene expression, or more recently, the CRISPR/Cas9 system for site-specific genome engineering, are available for *Drosophila* species (see [18] for review). In addition to genetic tools, genome sequencing is relatively easy in this genus. Due to their relatively small size, *Drosophila* genomes can be sequenced at relatively low cost [20]. *D. melanogaster* genome was

* Correspondence: cristina.vieira@univ-lyon1.fr
Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, F-69622 Villeurbanne, France

among the first eukaryotic genomes sequenced, and is arguably the best annotated genome so far. A lot of sequencing data are available in the *Drosophila* genus. The genome of at least 46 species were sequenced and assembled [21]. In addition, in *D. melanogaster*, several studies aimed at sequencing either individuals or populations (PoolSeq) [22–29]. This sequencing effort benefited largely from diverse consortia. One of the first, and probably one of the best-known, the *Drosophila melanogaster Genetic Reference Panel* (DGRP) consortium made available the genomic sequence of more than two hundred inbred lines from an American population [22, 24]. At a broader geographical scale, the global diversity lines consortium sequenced a panel of 84 worldwide strains [29]. We also should mention the European *Drosophila* Population Genomics Consortium (*DrosEU*) which recently produced PoolSeq data fom 48 European population samples [28]. Nowadays, more than 1,121 individual Drosophila genomes are available [30], as well as pooled genomes from 30 localities in Europe and 23 in North America. For some individual genomes of the DGRP, data about gene expression and various phenotypic traits are also available [22, 31–34]. DGRP lines and a large variety of mutants and natural strains of *D. melanogaster,* collected from all over the world at different times, are currently maintained and available for researchers [35]. In addition, more than 250 species are accessible [36]. From an ecological/genomics perspective, *Drosophila* species offer a unique opportunity to perform comparative studies. For instance, the pair *D. melanogaster*/*D. simulans*, with a short time of divergence (around 1.5 My), share a common geographical range, as both are cosmopolitan species, but have very different ecologies, the former being close to human habitats and the second being found only in forest environments [14, 37]. Other *Drosophila species*, such as *D. suzukii*, are classified as invasive species, and represent an opportunity to study the genomic determinants of the invasive process. A last example that we can cite is the use of *Drosophila* species as models for speciation studies. This has been done extensively using the species close to *D. melanogaster* (*D. simulans*, *D. sechellia* and *D. mauritiana*) [38–40] and species from the *repleta* group (*D. mojavensis* and *D. arizonae* [41–44]; *D. buzzatii* and *D. koepferae* [45–47].

A few words about Transposable Elements & *Drosophila*

Drosophila has been used as a model to study TEs for more than forty years now. The activity of the then-called "mobile dispersed genes" was already studied at the beginning of the 80's [48, 49]. Even before, they were studied as the uncharacterized inducers of the hybrid dysgenesis phenomenon [50, 51], in which the transmission of some genetic factor by the male but not the female resulted in a sterile progeny. Since then, research on TE in *Drosophila* heavily benefited from the advantages provided by this model, from genetic engineering to sequencing techniques. Not only the molecular mechanisms beyond the hybrid dysgenesis are now much better understood, but the study of this phenomenon also led to major discoveries in TE regulation, such as regulation by small RNAs. In this review, we aimed at giving an overview of the accumulated knowledge on Transposable Elements from molecular aspects to populations genomics in *Drosophila*, comparing the *D. melanogaster* to other *Drosophila* species where relevant.

TE diversity

About the classification

The abundance and ubiquity of TEs rapidly brought the necessity of a unified classification system for these sequences. The question of TE classification has been, and continues to be, a subject of debate [11, 52–54], especially the necessity for such system to reflect the phylogeny of TEs. From an evolutionary perspective, a purely phylogenetic classification seems ideal, however this may be hard to achieve. Beyond the polyphyletic nature of TEs, there are several other difficulties. One is that TE phylogeny does not necessarily reflect the organism phylogeny. Another is that the phylogenetic analysis of TE protein sequences may be arduous, because some TEs do not possess any coding sequence, some TEs possess several coding sequences with different phylogenetic signals due to recombination events, and some TEs are present in thousands of copies in the genome. In the sequencing era, when genome annotation is fundamental, Wicker *et al.* (2007) proposed a set of rules to rapidly classify TEs [11]. This widely used classification relies on transposition mechanisms, sequence similarities and structural relationships. In decreasing hierarchical order, we find the following classification levels: class, sometimes subclass, order, superfamily and family (and sometimes subfamily). The highest-level category, *i.e.* class, divides TE sequences into those with or without an RNA transposition intermediate. Next, the order category distinguishes sequences according to the insertion mechanism. Orders are further divided into superfamilies. The superfamily category discriminates sequences on the basis of particular features, for instance protein or non-coding domain structure, presence and length of direct repeats generated on both sides of a TE upon insertion (Target Site Duplication, TSD). The lowest-level category, *i.e.* family, includes sequences with a high rate of identity at the DNA level (at least 80% of identity over at least 80% of their internal or coding domain, or within their terminal repeat regions, or in both). Note that a

distinction also exists between autonomous TEs, *i.e.* TEs able to move by themselves, and non-autonomous TEs, *i.e.* TEs relying on other TEs to move, usually because they lack a certain protein.

Class I TEs: retrotransposons

Class I TEs are also called retrotransposons. They transpose via an RNA intermediate. The RNA intermediate is transcribed from a genomic copy, then reverse-transcribed into DNA by a TE-encoded reverse transcriptase. Each complete replication cycle produces one new copy. Retrotransposons can be divided into five orders: long terminal repeat (*LTR*) *retrotransposons*, Dictyostelium intermediate repeat sequence (DIRS)-like elements, Penelope-like elements (PLEs), long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). All of them are present in *Drosophila,* but LTR retrotransposons and LINEs are by far the most abundant [20, 55].

In *Drosophila*, LTR retrotransposons usually range from 5 to 7 kb (Fig. 1) [11, 57–59]. They owe their names to the direct Long Terminal Repeats (~300-400 bp) flanking them. They typically display two genes: *gag* and *pol*. *gag* encodes the capsid, and *pol* encodes a protease (Prot), an integrase (Int) and a reverse transcriptase (RT) with an RNase domain. After the transcription step, some transcripts will be translated while the others may end up transposed (Fig. 1) (see [60] for more details on transposition mechanisms). The protease of *pol* cleaves Pol into a protease, an integrase and a reverse transcriptase [61]. The Gag protein assemble into a capsid that makes a particle around untranslated transcripts, the integrase, reverse transcriptase and a tRNA [62]. Because the formed ribonucleoprotein (RNP) does not comprise the transcript from which proteins were translated, we typically refer to a trans-preference mechanism of RNP assembly. Using the tRNA as a primer for synthesis, the reverse transcriptase initiates the production of double stranded DNA from the TE transcript [63]. After reverse transcription, the particle falls apart, the integrase recognizes the two ends of the cDNA and inserts them into the host genome. Upon integration, LTR retrotransposons produce a TSD of 4-6 bp [11]. Note that the LTR order is further divided into five superfamilies: *Copia* (*e.g. Copia* and *1731* families), *Gypsy* (*e.g. HMSBEAGLE* and *412* families), *Bel-Pao* (*e.g. BEL, Roo* and *Max*

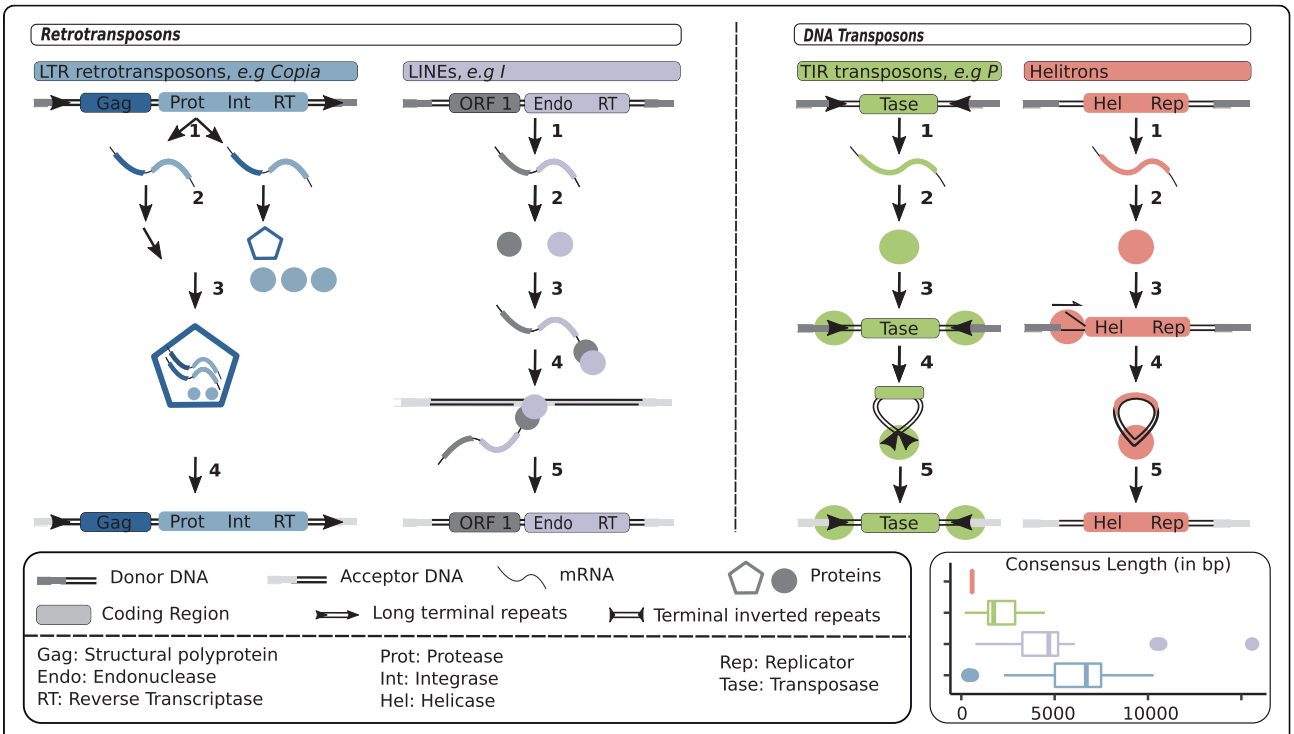


Fig. 1 TE structure and transposition mechanisms. LTR retrotransposons: 1. Transcription. 2. Translation of one part of the transcripts. The protease (Prot) cleaves *pol* polyprotein. 3. *gag* proteins assemble around untranslated transcripts, the integrase (Int), reverse transcriptase (RT) and a tRNA. 4. Reverse transcription and integration. LINE retrotransposons: 1. Transcription. 2. Translation. 3. Protein(s) bind to the transcript. 4. A strand of donor DNA is cut, target-primed reverse transcription starts at the exposed 3' extremity. 5. The TE is integrated. TIR DNA transposons: 1. Transcription. 2. Translation. 3. Two transposases bind to the TIRs. 4. Transposases dimerize and cut TIR extremities forming a free complex. 5. The complex binds to donor DNA and is integrated. Helitron DNA transposons: 1. Transcription. 2. Translation. 3. At the donor site, the plus strand is cut. A replication fork is formed. 4. Replication results in a double stranded transposon circle. 5. Integration. The bottom right panel represents the distribution of the lengths of *D. melanogaster* consensus sequences (RepBase [56]), using the same color code as above.

families), *Retrovirus* and *Endogenous RetroViruses (ERV)*. According to Wicker and colleagues classification, *Retroviruses* and *ERVs* also have an envelope gene (*env*). The corresponding protein allows *Retroviruses* to infect other cells. In *Drosophila*, few families have been shown to possess an *env* coding ORF, for example *Idefix*, *Gypsy*, *Tirant* and *ZAM* families [58, 64, 65]. Note that the insect endogenous retroviruses belong to the *Gypsy* superfamily, and that their origin is distinct from that of vertebrate ERVs [66]. Infectious properties have been demonstrated for *Gypsy* and *ZAM* families [67, 68].

LINEs are 3 to 5 kb-long, and generally contain two ORFs (Fig. 1) [11, 59, 69–71]. The first ORF encodes a protein with both RNA binding and nucleic acid chaperone properties [72, 73]. The second ORF encodes a protein that displays two domains: an endonuclease (Endo) and a Reverse Transcriptase [74, 75]. Contrary to LTR retrotransposons, LINEs exhibit a cis-preference mechanism of RNP assembly. After translation, the protein(s) bind to the mRNA molecule from which they originate, and form an RNP in the cytoplasm [76] (see [77] for more details on transposition mechanisms). The ribonucleoprotein particle moves back to the nucleus, and the protein cuts a single strand of the host genome at the point of insertion. The exposed 3' end allows the initiation of reverse transcription (target-primed reverse transcription). Subsequent events remains unclear, however the following has been proposed. During or after reverse transcription, the second strand of the host genome is cleaved. The newly reverse transcribed single-stranded DNA binds to the generated 3' extremity, and this extremity acts as a primer for the synthesis of the second strand of DNA. LINEs generate TSDs of various sizes upon insertion. Note that, probably as a consequence of early termination of reverse transcription, transposition may result in creation of 5' - truncated copies [78].

As mentioned above, besides LTR retrotransposons and LINEs that are abundant in *Drosophila* genomes, Class I comprises three other orders: DIRS, PLEs and SINEs. To our knowledge, DIRS and SINEs have not been found in Drosophila so far [20, 79]. PLEs were initially discovered in *D. virilis* and are involved in the hybrid dysgenesis phenomenon (Table 1). These TEs are present at least in the *virilis* group and in *D. willistoni* [89]. PLEs resemble LINEs, in a sense that they encode an endonuclease and a reverse transcriptase. However, they possess terminal repeats that can be in a direct or an inverse orientation.

Class II TEs: DNA transposons

Class II TEs are DNA transposons. They do not transpose via an RNA intermediate but via a DNA intermediate. There are four orders: terminal inverted repeat (TIR) transposons, Crypton, Helitron and Maverick. TIRs and Helitrons are the most abundant in *Drosophila*.

TIR Transposable Elements are typically ranging from 1.5 to 3 kb in *D. melanogaster*, and are characterized by their TIRs of variable lengths (Fig. 1) [11, 59, 90, 91]. TIRs encode one unique protein called transposase (Tase). The transposition mechanism begins with two transposases recognizing and binding to the TIRs [92]. Transposases dimerize and cleave the ends of TIRs forming a free complex containing the TE [93]. The formed entity binds to the target DNA locus, where the transposon is integrated. The TSD size and the sequences of TIRs are highly variable across the nine known superfamilies [11]. Although the transposition mechanism in itself is not replicative, such TEs can increase their copy numbers in two ways. First, by transposing during chromosomal replication from a position that has already been replicated to a position ahead of the replication fork [94]. Second, they can exploit gap repair following excision to create an extra copy at the donor site [95].

The Helitron order, which is represented by the unique Helitron superfamily, gave rise to rather small TEs in *D. melanogaster* (< 1 kb, Fig. 1) [11, 96, 97]. Helitrons encode one unique protein with both a DNA helicase (Hel) and a replicator (Rep) domain. Because Helitrons were discovered only in 2001, and the lack of active Helitron examples limits experimental work, Helitron transposition mechanisms remain murky. However, using an artificially reconstructed active Helitron, Grabundzija and colleagues provided new insights and suggested the model synthesized hereafter [98]. First, the plus strand, the original donor strand, is nicked at the 5'-extremity of the TE and a replication fork is created. DNA replication results in a reconstituted double stranded donor site and a double stranded TE circle. This step may be repeated several times, producing several TE circles. Moreover, on the TE circles, a second DNA cleavage may occur on the original donor strand, a new replication fork established, and two double stranded transposon circles obtained from one. Finally, the double stranded TE may be integrated at the acceptor site. Note that the small sizes of Helitrons in *D. melanogaster* are explained by their non-autonomous character.

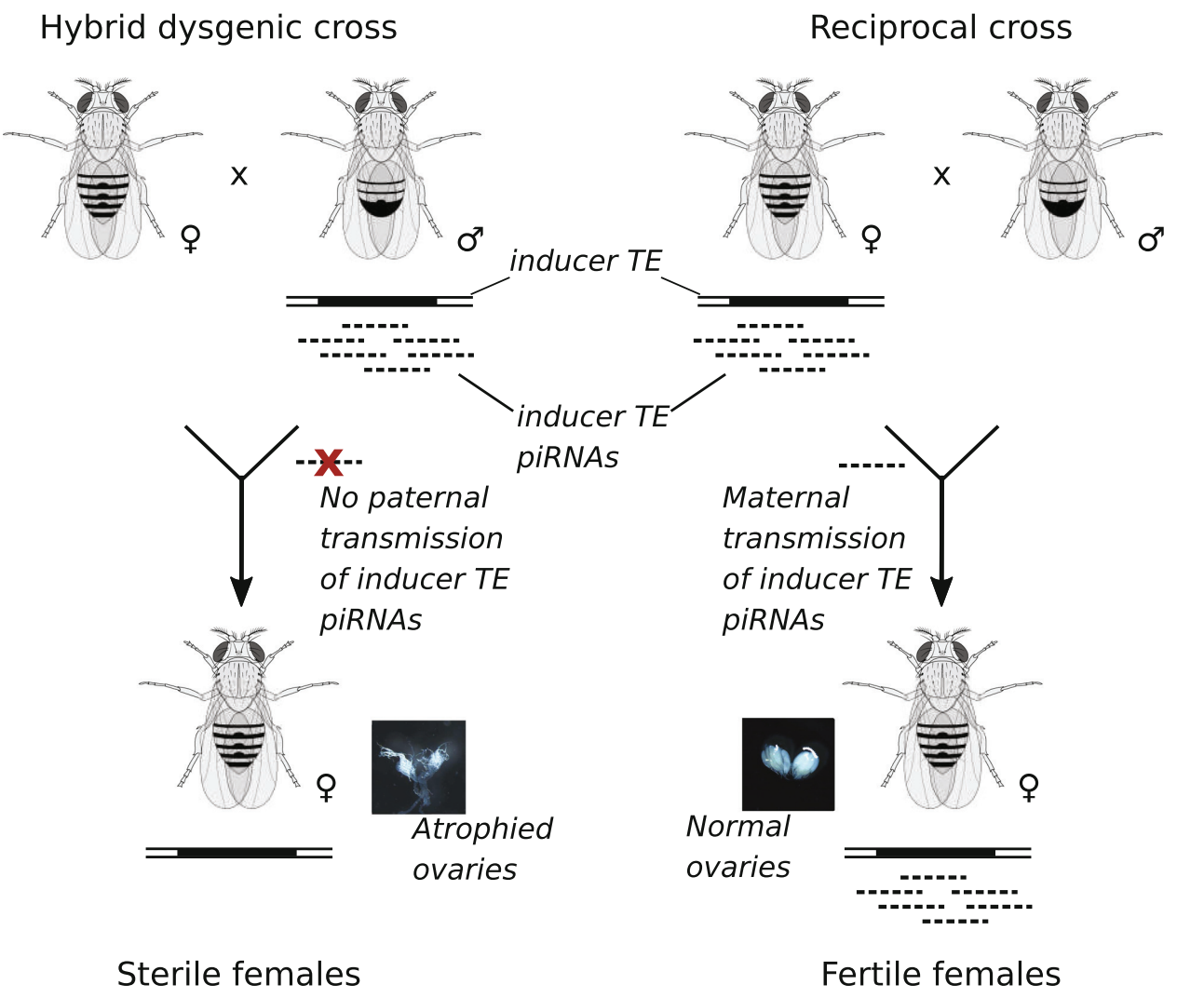
TE abundance

The Drosophila melanogaster reference genome

To obtain a picture of TE content in *D. melanogaster* genome, we investigated TE copy numbers and TE sequence occupancy in the last release of the reference genome assembly (Fig. 2). We used a combination of RepeatMasker, to identify genomic fragments

Table 1 Hybrid dysgenesis

In *Drosophila*, some intraspecific crosses were observed to produce sterile females [50, 51, 80–85]. This phenomenon is called hybrid dysgenesis (Ovaries pictures from ref 81). It happens when males possessing a particular TE, hereafter referred as the inducer TE, are crossed with females whose genome is devoid of this TE. On the contrary, the reciprocal cross leads to viable and fertile individuals. The explanation is related to piwi-interacting RNAs (piRNAs), small RNAs repressing TEs with sequence complementarity (see the piRNA section). Because piRNAs are maternally transmitted, in dysgenic crosses the inducer TE insertion is transmitted to the progeny without the piRNAs directed against it [86]. In the reciprocal cross, both the inducer TE insertion and its piRNAs are transmitted, allowing the control of the TE family in the progeny, and the hybrid to be fertile.
Hybrid dysgenesis was documented in three systems in *D. melanogaster:* associated with *P-element*, *I-element* or *Hobo* [50, 51, 80]. *P-element* also appears to induce hybrid dysgenesis in *D. simulans* [81]. In addition, in *D. virilis,* a hybrid dysgenic cross potentially implying several TEs was reported [82–85]. From a historical perspective, the hybrid dysgenesis phenomenon played an important role not only in the discovery of horizontal transfers of TEs, but also in the study of host defenses against TEs [86–88].



homologous to a library of *Drosophila* TE consensus sequences available in the RepBase database, and the bioinformatic tool OneCodeToFindThemAll to reconstitute TE copies [56, 100, 101]. As previously reported, *D. melanogaster* genome contains ~20% of TEs [55, 102]. Note that a significant variation exists regarding these estimates [103–105]. These differences are likely to be at least partly explained by the genome assembly, or the part of the genome assembly that is analyzed, or both. For example, the *Drosophila 12 genomes* consortium considered only the best-assembled part of the genome, likely representative of the euchromatic portion of the genome, and found the TE content ranging from 2 % to 8 % (see Population Genomics section for details about TE density in different genomic regions). On the contrary, even if far from reporting the entire sequence of heterochromatic regions, the assembly used in Fig. 2 comprises at least 20 Mb of heterochromatic sequences, *i.e.* ~15% of the 140 Mb assembly [106]. Nevertheless,

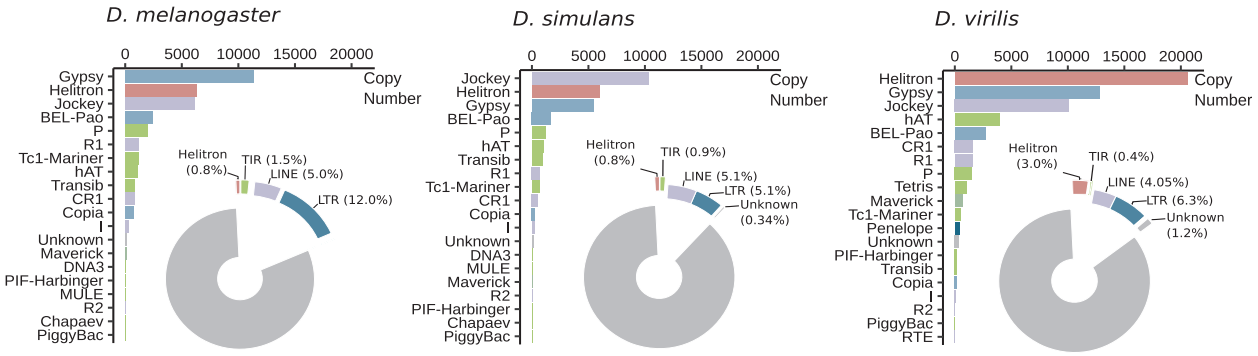


Fig. 2 TE contents in *D. melanogaster*, *D. simulans* and *D. virilis* (from left to right). Barplots represent TE copy numbers for the top 20 TE superfamilies. Piecharts illustrate genomic sequence occupancy of each TE order (in percentages of the assemblies). These results were obtained using the *D. melanogaster* reference genome assembly (r6.29), and recently produced long-reads assemblies of *D. simulans* and *D. virilis* [99]. RepeatMasker was used to recover TE fragments and TE genomic sequence occupancy (RepeatMasker v1.332, -nolow, -norna, -species drosophila; Repbase-derived RepeatMasker libraries 20181026 [100],). TE fragments were assembled into TE copies using OneCodeToFindThemAll [101].

the relative abundance of the different TE orders is globally conserved across studies and similar to what is represented in Fig. 2 [55, 102, 103, 105]. Retrotransposons, and essentially LTRs and LINEs (respectively 12% and 5% of the genome in our analysis), contribute substantially to *D. melanogaster* TE content. DNA transposons correspond to a smaller proportion of the genome: we found that they represent less than 2%, including 0.9% for Helitrons and 0.7% for TIR elements. This ten-fold difference in terms of genomic sequence occupancy between retrotransposons and DNA transposons is mostly due to the larger size of retrotransposons (Fig. 2). Indeed, in terms of insertion numbers we found 11,657 DNA transposons (6,284 Helitrons and 5,373 TIR elements) and 23,148 retrotransposons (14,540 LTR retrotransposons and 8,608 LINEs) (see also [103] and [101]). For each of the four major orders, one superfamily is often over-represented: *Gypsy* for LTR elements, *Jockey* for LINEs, *P* for TIR elements, *Helitron* for Helitrons. According to our analysis, the different TE orders exhibit different numbers of families: indeed, we found insertions belonging to 721 LTR families, 331 LINE families, 213 TIR families and 63 Helitron families. The mean copy number per family is 26, but large variations exist. The family having the highest number of insertions is DNAREP1_DM, for which we found 1,746 copies. This sequence is annotated as a non-autonomous Helitron [107] (but see [97, 108] concerning classification).

Interspecific variation

When it comes to TE contents across *Drosophila* species, a direct comparison of studies may be difficult. Indeed, authors are free to choose among a large number of programs and methods dedicated to identifying TEs, which leads to widely different results [105, 109]. For example, using the same TE sequence library but two different tools to annotate the *D. willistoni* genome, the 12 genomes consortium estimated TE content to be either 9 % or 16 %. The library used may also greatly affect results. In the same study, using the same tool, but a *D. melanogaster* TE sequence library or a *de novo* library, the authors found either 12 or 20 % TEs in the *D. ananassae* genome. Overall, in this study seven combinations of library-detection tools were used, leading to a TE content ranging from less than 10 % to up to 30 % in *D. ananassae*. The direct comparison of studies may thus be risky. A further layer of complexity comes from the sequencing technology, which impacts the quality of genome assemblies. Short paired-end read based assemblies lead to underestimation of TE contents compared to Sanger and long read based assemblies [110–112]. For all these reasons, to describe variation of TE contents in the *Drosophila* genus, here we focus on studies directly aiming at comparing TE amounts across species, and we remain cautious when linking them. For illustrative purposes, in addition to the annotation of TE contents in *D. melanogaster*, we estimated TE genomic sequence occupancy and copy numbers in two species: *D. simulans* and *D. virilis* (Fig. 2). We used the exact same methods as for *D. melanogaster,* and we do not expect the TE library to strongly bias the results, as it contains sequences constructed from the three species, which are among the most - studied with regard to TEs [113, 114]. Beyond that, we chose these two species because of their different positions relatively to *D. melanogaster* in the *Drosophila* phylogeny. On one hand, *D. simulans* is a close relative to *D. melanogaster;* they diverged approximately 1.5 Mya. Both species belong to the *melanogaster* subgroup within the *melanogaster* group, itself in the *sophophora* subgenus [14]. On the other hand, *D.*

melanogaster and *D. virilis* diverged about 25 Mya and *D. virilis* belongs to a different subgenus, the *drosophila* subgenus.

The first study intending to compare global TE contents across a significant number of *Drosophila* species was performed by the *Drosophila 12 genomes* consortium. This consortium investigated TE genomic sequence occupancy in eight species from the *sophophora* subgenus, mostly from the melanogaster subgroup, and four species from the *drosophila* subgenus. As stated above, the researchers focused on genomic parts likely to be euchromatic, and they used different methods. Using the method giving the lowest estimates, they found a global range of variation going from 1% to 9% of TEs in the genome. The method leading to the highest estimates resulted in genome containing from 3% to 30% of TEs. Invariably, *D. ananassae* was the species with the highest proportion of TEs. The authors chose the most unbiased and conservative method to compare the relative abundance of LTR retrotransposons, LINEs, TIR elements and so-called OTHERs among species. They found that the pattern LTRs>LINEs>TIRs>OTHERs is globally conserved across the phylogeny, with LTR retrotransposons usually constituting more than 50% of the repeatome. The two exceptions are *D. mojavensis* and *D. pseudoobscura.* In *D. mojavensis*, LTR elements represent only 45% of the repeatome, and in *D. pseudoobscura*, LTR retrotransposons and LINEs each contribute to roughly 33% of the repeatome. Our analysis shows a slightly different pattern, with equivalent genomic sequence occupancy for LTR elements and LINEs in *D. simulans*, and more Helitrons than TIR elements in *D. virilis* (Fig. 2). Recently, Hill and colleagues investigated both the proportion of TEs and their number of insertions in the genomes of five species. Four of these species were already in the set analyzed by the *Drosophila 12 genomes* consortium, except for *D. innubila*. The LTRs>LINEs>TIRs>OTHERs pattern for TE genomic proportions was not respected by any of the considered species. The dominant category differed: the most abundant elements are LTR retrotransposons in *D. ananassae,* while they are LINEs in *D. pseudoobscura*, and DNA transposons in *D. innubila*. *D. ananassae* was also the species with the highest TE content, with approximately 35% of TEs in the genome. Considering TE copy numbers, the authors found a total ranging from 2,000 to 14,000 depending on the species. Once again, the difference with the previous results may probably be explained by data/method differences. Relative abundances of the different TE categories were found to differ across genomes. For example, DNA transposons were the most abundant in *D. willistoni,* whereas in *D. ananassae* they were as numerous as LINEs or LTR elements. The study with the largest dataset of species compared in terms of TE content was published by Sessegolo and collaborators [20]. These authors investigated the TE contents of 26 *Drosophila* species. Once again, the LTRs>LINEs>TIRs>OTHERs pattern did not hold for many species. The genomic content of repeats ranged from 4.65% in *D. busckii* to 30.80% in *D. suzukii*. The authors found a significant effect of phylogenetic inertia on TE content, but because of uneven sampling across the phylogeny, it was difficult to extract a pattern for each subgroup, many being represented by only one species. Overall, the data suggest large variations in the abundance of TEs across the *Drosophila* genus.

Intraspecific variation

At the intraspecific level, genome size, which is correlated to TE abundance in *Drosophila*, is variable within populations of both *D. simulans* and *D. melanogaster*. This suggests that TE contents may change between populations, at least quantitatively [20, 55, 115]. In addition, the discovery of hybrid dysgenesis, *i.e.* the generation of a sterile hybrid by crossing particular parental strains differing by TE families, has highlighted qualitative differences in TE content at the intraspecific level (Table 1) [50, 51, 69]. TE contents in populations were extensively studied by *in situ* hybridization on polytene chromosomes, restricting the results to a few families. Quantitative differences related to the hybrid dysgenesis phenomenon have been observed for *I-Element*, *P-Element* and *Hobo* in *D. melanogaster* [69, 80, 116]. It has been demonstrated that the P-element has recently been acquired by horizontal transfer, likely from *D. willistoni*, and then spread step by step in worldwide populations between 1950 and 1990 [87, 117–119]. The history seems to repeat itself with the current invasion of *D. simulans* by the *P-element* after a horizontal transfer event from *D. melanogaster* [81, 120]. Horizontal transfers of TEs have now been extensively described in eukaryotes [121] and the study of TEs in the genomes of *D. melanogaster*, *D. simulans* and *D. yakuba* suggests that one-third of TE families has originated by recent horizontal transfers between these species [122]. In addition to hybrid dysgenesis, the study of 34 TE families from various populations of *D. simulans* by Vieira and colleagues showed fairly large qualitative differences between populations. Indeed, they found at least 14 families of TEs that were present only in certain populations [123, 124]. Quantitatively, and as an example, a study of the *412* element in *D. simulans* showed a gradient in copy numbers ranging from 1–10 in South Africa to 23 in Europe [125]. Genome size and TE content variations parallel the worldwide colonization of *D. melanogaster* but not that of *D. simulans* [115]. In *D. subobscura*, *Bilbo* and *Gypsy* families show slightly more copies in colonizing than original populations [126]. Similar

results were obtained when contrasting copy numbers of *Bilbo* and *Osvaldo* between colonizing and original populations of *D. buzzatii* [127]. In both cases, the study of insertion frequencies suggested that genetic drift associated with a founder effect that accompanied the colonization was responsible for the observed variation of copy numbers. Recently, genomic analyses of European *D. melanogaster* populations from *DrosEU* confirmed that intraspecific variation of TE contents may be substantial, and reveals TE proportions ranging from 16% to 21% of genomes [28].

TE activity

Spontaneous rate of transposition

A recent study by Adrion and colleagues [128] provided the first genome-wide estimate of TE movement rate in *D. melanogaster*. These authors used NGS data to compare TE contents across laboratory lines before and after ~150 generations of mutation accumulation. They found that the TE movement rate is slightly lower than the point mutation rate: 2.45 × 10(-9) per site per generation against 2.8 × 10(-9) per site per generation, respectively [129]. The rate of insertions is higher than the rate of deletions: 2.11 × 10(-9) per site per generation against 1.37 × 10(-10) per site per generation, respectively. Considering that there are 270 millions sites in the genome assembly, these numbers correspond to approximately 0.57 insertions and 0.037 deletions per generation. Those estimates were obtained across all TE superfamilies and are consistent with previous reports using *in situ* hybridization to determine transposition events for one or a few families [130–132]. Adrion and colleagues found superfamily-specific insertion and deletion rates to range between 0 and 5.13 × 10(-3) per copy per generation, and between 0 and 1.29 × 10(-4) per generation, respectively. They also found a significant effect of the genetic background, as previously reported [133–135].

Transposition bursts

Beyond the spontaneous rate of transposition, a significant number of studies have shown that transposition bursts could occur in *Drosophila* (see [136] for a review). A burst is characterized by movement of large numbers of TE sequences through the genome during a short evolutionary time [137]. Although these bursts can happen without any apparent reason, they are commonly associated with stressful conditions such as extreme temperatures, irradiation, chemical exposure, or viral infection [138–142]. For example, Vasil'eva and colleagues showed that gamma radiation could increase the *412* transposition rate up to 5.6 events per genome per generation. Note that the attempts to induce TE mobilization with thermal shocks led to contradictory results in *Drosophila*, potentially due to the differences between tested genetic backgrounds, or tested TEs, or both, but also to methodological considerations (see [136]). Furthermore, although to our knowledge it has not been observed in *Drosophila* so far, stress may also lead to repression of TE activity [143]. Another stress widely studied in *Drosophila* for its effect on transposition is the genomic stress occurring when two somehow divergent genomes are united after hybridization (Table 1). In several biological systems it increases TE activity with potentially dramatic consequences on the phenotype, including sterility [144, 145]. It was observed when crossing individuals from different species, but also when crossing particular strains from the same species which corresponds to the hybrid dysgenesis phenomenon mentioned above [47, 50, 51, 80]. The causes of the TE bursts are not completely elucidated yet. Concerning hybridization, it has been shown that a failure of the host defense against TEs could be at stake (see below and Table 1). Regarding TE activation in response to stressful conditions, it has long been suggested that it could be due to TEs displaying binding sites for stress specific transcription activators, such as transcription factors [146]. In agreement with this idea, the temperature responding *Mariner* and *Copia* elements were shown to display sequences homologous to the promoter of heat shock proteins [147, 148]. More recently, a transcriptomic study demonstrated that temperature dependent TE expression is TE family specific and dependent on the genetic background. The authors proposed that TE transcription is indeed regulated by an interaction between TE family-specific regulatory sequences and host *trans*-acting factors [149]. Note, however, that this study was done on a range of temperatures that are not necessarily stressful (13–29°C). It is also important to consider that all the reports mentioned above concern laboratory experiments in conditions that are potentially unlikely *in natura*. The mechanisms at play in natural populations still remain poorly understood. One study demonstrated a burst of transposition for *DINE-1* in *D. yakuba* [150], and its causes are still unknown. In *D. simulans,* the copy numbers of the *412* element increase with latitude following the minimum temperature, and in *D. melanogaster*, significant correlations were found between TE abundance and different geographical and environmental variables for four families [125, 151]. However, in both cases, a possible confounding effect of demographic history cannot be excluded. Only one study established a direct link between TE activity and a geo-climatic variable: in *D. simulans,* the *Mariner* element somatic activity varies along a latitudinal cline between tropical Africa and Europe [152].

Interspecific variation

So far, few studies tried to compare TE activity across *Drosophila* species. In 2011, Lerat and colleagues compared the TE contents of four *Drosophila species* from the *melanogaster* subgroup: *D. melanogaster*, *D. simulans*, *D. sechellia* and *D. yakuba* [153]. They found that *D. simulans, D. sechellia* and *D. yakuba* genomes contained a large fraction of degraded copies compared to *D. melanogaster*. The authors suggested a recent TE activity in *D. melanogaster*, compared to the three other species. This can partially be observed when comparing the so-called TE landscapes of *D. melanogaster* and *D. simulans* (Fig. 3). These landscapes constitute an easy way to visualize TE activity through time. The X axis corresponds to the divergence of the TE sequences from the consensus, and it can be seen as a proxy of the time passed since the last wave of transposition. In Fig. 3, we can see a recent peak of activity of LTR elements, especially in *D. melanogaster*. In *D. simulans*, the peak of activity is also recent but much smaller. Another study was aimed at comparing TE activity between *D. melanogaster* and *D. simulans* using NGS population data [155]. Based on TE insertion frequency data, the authors determined that more than 58 families are probably highly active in both species. Half of the TE families show evidence of variation of activity through time, and are not the same depending on the species. Finally, they found that retrotransposons were the most active TEs in *D. melanogaster,* while DNA transposons were the most active TEs in *D. simulans*. A recent study compared TE frequencies in five distant points of the *Drosophila* phylogeny [55]. These species shared a common ancestor around 30 Mya [14]. The authors found evidence that an excess of low frequency insertions is prevailing in the phylogeny and is observed for most TE families. This suggests that an active repeatome is frequent, at least in the *Drosophila* genus.

Impacts of TEs

On the genome

TEs play an important role in the structural evolution of genomes through the generation of various types of mutations: chromosomal rearrangements, gene disruption and changes in gene expression. The simplest mechanism by which TEs can cause chromosomal rearrangements is through participation in an ectopic recombination event [156]. Ectopic recombination corresponds to recombination between more-or-less identical sequences inserted at different locations in the genome, such as TEs [157]. Depending on their relative positions and orientations, their recombination can result in different kinds of chromosomal rearrangements: duplication, deletion, inversion, or translocation. TEs were associated with chromosomal rearrangements *in natura* in various species of *Drosophila*, and mainly with inversions [158–161]. In several cases, ectopic recombination was identified as the cause of these rearrangements [159, 160]. When they insert into genes or their regulatory sequences, TEs can disrupt gene function. A perfect example is the use of the *P-element* in the Berkeley *Drosophila* Genome Project [162–164]. The Berkeley *Drosophila* Genome Project aimed at disrupting each *D. melanogaster* gene using the *P-element* in order to decipher gene functions. More than 5,000 genes were disrupted in that way. TEs can affect gene expression in two principal ways. First, they may bring regulatory sequences (see [165] for a review). For example, *Bari-Jheh* adds extra antioxidant response elements upstream of the *Jheh1* and *Jheh2* genes and is associated with

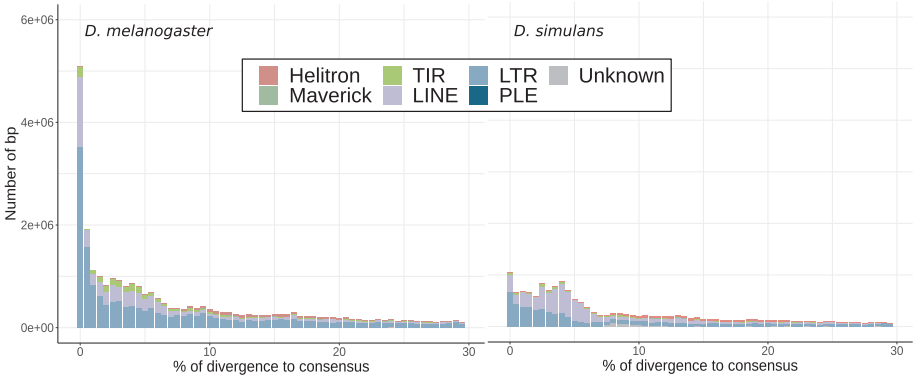


Fig. 3 TE landscapes in *D. melanogaster* and *D. simulans*. For each TE fragment the divergence to consensus was estimated. For each TE order the total amount of DNA (in bp) is shown as a function of the percentage of divergence. The percentage of divergence to the consensus sequences is a proxy for age: old TEs have accumulated mutations, young TEs are similar to consensus sequences. RepeatMasker was used to recover TE fragments in genomic assemblies (same method as Figure 2 [100],). Percentages of divergence to consensus were evaluated from RepeatMasker output .align file using A. Kapusta script [154]

upregulation of *Jheh1* and *Jheh2* [166]. Second, the spread of repressive epigenetic marks targeting TEs can reduce the expression of nearby genes (see below, host defenses against TEs), as it was also demonstrated in the *Jheh* cluster [167]. Lee and Karpen demonstrated recently that the spread of repressive epigenetic marks to nearby DNA occurs for more than half of euchromatic TEs, and can extend up to 20 kb [12]. This effect is TE dependent, copy number dependent, but also species dependent, with stronger epigenetic effect in *D. simulans* compared to *D. melanogaster*.

On the individual

While some of the aforementioned genomic changes might remain phenotypicaly silent, others may have dramatic repercussions at the individual level. TEs are responsible for up to 80% of the phenotypic spontaneous mutations observed in *D. melanogaster* [168] and many observations suggest deleterious effects of TEs in *Drosophila*. Five to 10 % insertions of active *P-elements* are estimated to cause recessive lethal mutations in *D. melanogaster* [169]. In *D. simulans,* somatic transposition of *Mariner* decreases lifespan [170]. In 2004, a study used two *D. melanogaster* lines with the same genetic background, but different TE copy numbers, to evaluate the impact of TE number on fitness. The authors found differences in fitness and egg hatchability between the two lines, the line with more TEs performing worse than the other. Both homozygous and heterozygous TE insertions were shown to have deleterious effects on fitness and its components [134]. Overall, TE insertions are expected to be generally neutral or deleterious to the host genome [171]. Considering that adaptive mutations are supposed to quickly reach fixation in populations, the low numbers of fixed insertions in *D. melanogaster* and *D. simulans* support this theory. In 2006, Burt and Trivers calculated the number of insertions since the divergence between the two species and concluded that, given both genome size and number of fixed insertions, the occurrence and fixation of a beneficial insertion is a really rare event [156]. However, they also underscored the difficulty to detect fixed insertions using *in situ* hybridization, and suggested it would have been interesting to estimate the rate of fixation from sequencing data. In 2015, using population sequencing data, Kofler and colleagues estimated the number of fixed insertions in *D. melanogaster* since its divergence from *D. simulans* to be approximately 200 [155]. Considering a 1.4 Mya divergence [14], we computed a fixation rate of 1.4 fixed insertions every 10,000 years, *i.e.* maximum 1.4 beneficial fixed insertions every 10,000 years. If we update the Burt and Trivers calculation and compare the number of fixed insertions to the total number of insertions over this period: Population size × Insertion rate per genome per generation × Divergence time between *D. melanogaster* and *D. simulans* × Number of generations per year = 10 (6) × 0.57 × 1.4 × 10 (6) × 24 = 1.9 × 10 (13) insertions, that is to say 200/(1.9 × 10 (13)) = 1.0 × 10(-11) insertions reaching fixation. Finally, we estimated maximum 1.4 beneficial fixed insertions every 10,000 years, or maximum 1 out of 1e11 insertions, being beneficial and fixed. These numbers are upper bounds because all fixed insertions are unlikely to be beneficial. Indeed, most of the fixed insertions are present in regions where the effect of selection is weak, and are essentially old. Therefore, they are more likely to have reached fixation slowly by drift than quickly by positive selection [172, 173]. So far, 21 fixed insertions have been identified within or near genomic regions showing low Tajima's D values, and 12 fixed insertions are relatively young. Considering the above, one could expect to find very few putatively adaptive insertions among unfixed insertions. Surprisingly, there are at least 57 of such insertions in the reference genome [173], suggesting a high rate of TE mediated adaptation recently or even ongoing. The discrepancy between the number of candidates for recent adaptation and the fixation rate was discussed considering the three following points: 1. The migration of *D. melanogaster* out of Africa may have caused a significant augmentation of the adaptation rate. 2. TE derived adaptations might be ephemeral. 3. Adaptive TE sequences may evolve quicker than neutral insertions, resulting in an underestimation of the number of fixed insertions [174]. One may also add that the TE mutation rate has potentially increased recently [175]. It is worth noting that few insertions were clearly associated with an adaptive phenotype so far [166, 176–178]. Interestingly, candidate adaptive insertions are often close to, or within genes associated with stress response, behavior and development. Moreover, two of the historical examples of adaptation associated with TEs correspond to two different insertions in the same gene implicated in the response to oxidative stress, *cyp6g1,* in two different species: *D. melanogaster* and *D. simulans* [176, 178, 179].

The case of telomeric elements

A few TEs appear to have evolved a new function in *Drosophila* genomes. Because of the DNA replication mechanism, a *Drosophila* chromosome end loses 70-80 bp each generation [180]. This gradual reduction of

chromosome ends is threatening internal regions containing essential genes and may contribute to ageing [181]. Organisms have evolved different mechanisms that protect their chromosomes. Usually in eukaryotic genomes a ribonucleoprotein enzyme, the telomerase, mediates the RNA dependent synthesis of tandemly repeated simple sequences at chromosome ends [182]. In *D. melanogaster*, the three families, *HeT-A*, *TART* and *TAHRE,* transpose to chromosome extremities, and protect them from shortening [180, 183–186]. Many phylogenetically distinct telomeric retrotransposons have been found in more distant species [187]. All these telomeric elements belong to a single monophyletic clade inside the *Jockey* superfamily. The telomeric element phylogeny and species phylogeny are congruent, suggesting vertical transmission from a common ancestor and a conserved host-element relationship [187]. Furthermore, the clade presents evidence of specialization to transpose at chromosome ends [188]. Because of this, the relationship between TEs and their host in this case was referred to as genomic "symbiosis" [188]. However, Saint-Leandre and colleagues investigated more species of the *melanogaster* group [189]. They suggest that these *Jockey* telomeric elements may have evolved to selfishly over-replicate. In agreement with this hypothesis, they found recurrent gains, losses, and replacements of *Jockey* telomeric elements. Moreover, in *D. biarmipes*, the telomere-specialized elements have disappeared completely.

Host defenses

Because of the above-mentioned deleterious effect of TE insertions, several mechanisms of TE control have evolved. Among these, epigenetic modifications play an important role [190]. For example, in mammals and plants, TE insertions are usually associated with DNA methylation and histone modifications. Both are related to repressive chromatin states. In *Drosophila*, DNA methylation has been shown to be almost completely absent, and small RNAs are central to TE regulation [191, 192]. They may also trigger histone tail modifications and chromatin conformation modifications. There are two small RNA pathways controlling TEs in *Drosophila*: the piRNA and the siRNA pathways. Our purpose here is to give a brief overview of these pathways and their role in shaping TE dynamics. In particular, we refer the reader to [193, 194] for comprehensive reviews on the mechanistic aspects of the piRNA pathway.

The piRNA pathway

The piRNA pathway produces small, single stranded RNAs that were first called rasiRNAs (repeat associated small interfering RNAs); however, contrary to regular small interfering RNAs, they are 23-30 nt long, and are associated with the Piwi-subfamily Argonaute proteins, which led to their new designation as piRNAs (piwi-interacting RNAs). These piRNAs silence TEs in germ cells, where maintaining the integrity of the genome is of primary importance, as new mutations are passed on to future generations. This pathway is also active in the ovarian somatic follicle cells, which support oogenesis. It prevents endogenous retroviruses, such as *Gypsy,* from infecting the adjacent oocyte [195]. Research studies in *Drosophila* were seminal in the piRNA field. Much of what we know today was discovered using this model. In fact, piRNAs were identified for the first time in 2001 in fly testis [196]. They were found to silence *Stellate,* a gene involved in male sterility. Some of them were even found to be homologous to TEs and assumed to be involved in transposon regulation. Moreover, a long-term study of the *Gypsy* family activity led to the discovery of *flamenco*, a non protein-coding locus producing piRNAs, which was subsequently shown to be involved in the control of other TE families, essentially LTR retrotransposons [197, 198] (Table 2).

piRNAs originate from discrete genomic loci called piRNAs clusters. These loci contain mainly defective TEs and are transcribed into long piRNA precursors (Fig. 4 [202]). Approximately 150 clusters have been identified in the genome of *D. melanogaster*, representing 3.5% of the assembled genome [208]. The vast majority of them appear to be heterochromatic. The size of piRNA clusters varies substantially, with the largest being 240 kb. Overall, the largest 15 clusters produce a

Table 2 the *flamenco* story

Much of what we know today on the piRNA pathway was discovered using the *Drosophila* model. Especially, a considerable effort started 40 years ago in *D. melanogaster* led to the early discovery of a gene producing piRNAs and silencing *Gypsy* in a piwi dependent manner (see [199] for a review). This gene named *flamenco* was the first piRNA cluster identified.

In the 1980s, the *ovoD* dominant mutation was identified in *D. melanogaster* and associated with female sterility [200]. Interestingly, crosses between *ovoD* males and females from a particular strain led to the reversion of the phenotype and the recovery of fertility of the daughters, in addition to numerous mutations at other loci. Further work revealed that the particular strain used for the mothers actually displayed high copy numbers of uncontrolled *Gypsy*, whose transposition into *ovo* led to a null allele and reversion of sterility [201]. And then, it was demonstrated that the locus controlling *Gypsy* activity was the *flamenco* locus, located on the X chromosome and containing a lot of TE sequences [202, 203]. In 2004, Sarot and collaborators found out that *Gypsy* transposition was sensitive to a mutation in *piwi*, a gene known to affect RNA-mediated silencing [204]. They also demonstrated that small RNAs homologous to *Gypsy* where present in silenced tissues. For the first time a gene producing small RNAs was associated with TE silencing, and Piwi was implicated in this process.

Despite the subsequent discovery of many other piRNA clusters in *D. melanogaster*, flamenco is still widely studied as a model and produces most of the piRNAs in ovarian somatic cells [197, 205–207].

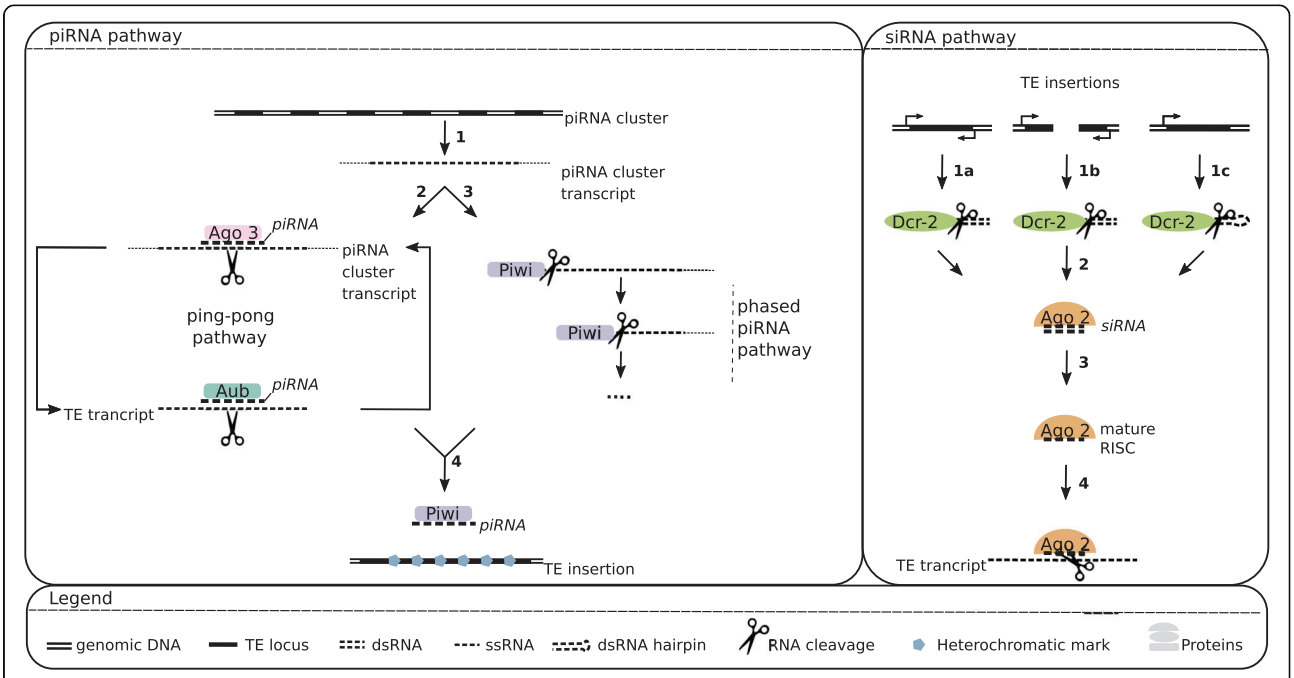


Fig. 4 small RNA pathways controlling TEs. piRNA pathway: 1. RNA PolII transcribes a genomic piRNA cluster into a long single stranded RNA. 2. The transcript thus formed enters the ping-pong pathway, which is ensured by Aub and Ago3, generates sense and antisense piRNAs, and ensure post-transcriptional silencing by transcript slicing. 3. Piwi directs the cleavage of the piRNA cluster transcript and generates a piRNA. This step may be repeated. 4. Transcriptional silencing: in the nucleus, a piRNA guides Piwi and promotes H3K9 methylation of TE DNA sequences. siRNA pathway: 1. Generation of a long dsRNA by: a. bi-directional transcription of a unique TE locus, b. interaction of two complementary transcripts from distinct TE loci. c. hairpin formation, due for example to inverted repeats binding 2. Dcr-2 processes long dsRNAs into siRNAs, which are loaded on Ago2. 3. The passenger strand of the siRNA is sliced by Ago2, only the guide strand remains. 4. The RISC binds to a TE transcript with sequence complementarity to the guide strand and Ago2 cleaves it.

large proportion of the total amount of piRNAs: 70% of the piRNAs uniquely mapped to the genome originate from these clusters.

The beginning of piRNA biogenesis is similar in germline and somatic cells (see [193, 194] for detailed reviews). PiRNA cluster transcription is ensured by RNA Pol II and leads to a single stranded long RNA (Fig. 4). Then, piRNA cluster transcripts may enter either the ping-pong pathway or the phased piRNA pathway [208–213]. The ping-pong pathway occurs in germline cells. In this case, guided by a sense piRNA, Argonaute3 (Ago3) binds to a complementary piRNA cluster transcript and cleaves it. Then, Aubergine (Aub) attaches to the newly formed 5' extremity, slices the transcript and forms an antisense piRNA. Finally, guided by an antisense piRNA, Aub operates a cut in a TE transcript, Ago3 recognizes the resulting 5' extremity, cleaves the transcript and forms a sense piRNA. This is the ping-pong pathway or ping-pong loop. The phased piRNA pathway is not specific to germline cells and may also occur in ovarian somatic follicle cells. Piwi is loaded at the 5' extremity of the piRNA precursor and Zucchini (Zuc) performs cleavage, generating the piRNA. Piwi is then loaded again at the 5' extremity of the precursor piRNA, and the process is repeated in a step-by-step cleavage generating multiple piRNAs. Note that, for clarity, piRNA maturation steps such as trimming are not mentioned here.

After synthesis, piRNAs mediate silencing both at the transcriptional and post-transcriptional levels [214]. The post-transcriptional silencing occurs in the cytoplasm of germline cells only, and corresponds to the ping-pong pathway (Fig. 4). At the transcriptional level, a piRNA guides the Piwi protein to a TE insertion, probably due to sequence complementarity with nascent TE transcripts, and mediates local heterochromatin formation by addition of the repressive mark H3K9me3 to histone tails [215–221]. Note that, despite the fact that as early as 2001 piRNAs were detected in testes, so far most of the work on TE regulation by piRNAs has been done on ovaries [196]. Regulation in testes seems to be quite similar to what happens in female germline, with both ping-pong and phased piRNA pathways being active [222–224]. However, contrary to ovaries, the data suggest an Ago-3 independent amplification loop in spermatogenesis.

The siRNA pathway

In addition to piRNAs, sequencing of small RNAs revealed the existence of another class of interfering RNAs targeting TEs: endogenous small interfering RNAs, or endo-siRNAs [225–227]. These small RNAs are present in both somatic and germline cells. endo-siRNA precursors are double strand RNAs (dsRNAs). These precursors may be produced through three distinct mechanisms (Fig. 4) [228]. 1. Transcription of the same genomic region in both sense and antisense directions (convergent transcription), then base pairing of the overlapping region between sense and antisense transcripts. 2. Transcription of complementary sense and antisense transcripts from different genomic regions and base pairing. 3. Base pairing of inverted repetitive elements of one transcript to form a hairpin RNA. The resulting long dsRNA is loaded on Dicer-2 (Dcr-2) and its cofactor Loquacious-PD (Loqs-PD) and then processed into 21 nt small double stranded RNAs. They are then loaded on the *RNA-induced silencing complex (*RISC) including the Ago2 protein. One strand is held and guides the complex to target transcripts that are then cleaved by the RNase domain of Ago2.

Evolution

Several studies demonstrated rapid evolution of anti-TE RNAi genes in *Drosophila* [47, 229–232]. Indeed, these genes often present signatures of recurrent positive selection. By analogy to the signatures of positive selection observed for genes involved in host-parasite interactions, the rapid evolution of anti-TE RNAi genes is often interpreted as a consequence of an arms race occurring between TEs and TE immunity effectors. Focusing on the piRNA pathway, Blumenstiel and colleagues propose that selection for sensitivity to TE content but also selection for specificity to TE content may drive the rapid evolution of host defense mechanisms [233]. More precisely, concerning the specificity aspect, the authors propose that a too efficient piRNA pathway may induce a too efficient silencing of TE copies that could spread to neighboring genes, which would constitute a cost. They designated this form of off-target gene silencing as “genomic autoimmunity”, an analogous to classic forms of autoimmunity which are caused by an immune response that incorrectly targets self. Despite the rapid evolution of anti-TE RNAi genes in *Drosophila*, suggesting that host defense mechanisms may vary a lot across the genus, most of the literature on this subject concerns *D. melanogaster*. A recent study of 20 arthopod species suggests that somatic piRNAs were probably produced in the ancestral arthropod more than 500 Mya and demonstrated that, in contrast to *D. melanogaster*, *D. virilis* presents somatic piRNAs [234]. This suggests a loss of the piRNA pathway in the soma of *D. melanogaster*.

Population genomics

The *Drosophila* model has been of outstanding importance in the field of population genomics of TEs. The ease to get and maintain wild type strains was obviously a key factor, but so was the development of the *in situ* hybridization method on *Drosophila* polytene chromosomes more than 40 years ago [235, 236]. *In situ* hybridization allows to detect and localize genomic DNA sequences using a labeled sequence (probe) homologous to the targeted sequence. The giant polytene chromosomes are found only in some species and tissues, and offer to the researcher a high degree of resolution [237, 238]. Using TE probes on salivary gland polytene chromosomes of *Drosophila* third instar larvae, researchers were able to detect and localize TE insertions in individuals and thus to accurately estimate TE insertion frequencies in natural populations [239–241].

About the nature of selection acting on TEs

The first *in situ* hybridization studies evaluating TE insertion frequencies in natural populations of *D. melanogaster* demonstrated a predominance of insertions segregating at low frequencies [239–241]. This result obtained for specific families was later confirmed at a broader scale. Population sequencing data showed that, in *D. melanogaster* and *D. simulans*, more than 80% of TE copies have insertion frequencies lower than 0.2 [155]. This observation is often interpreted as the result of purifying selection acting on TEs. So far, three main hypotheses have been formulated concerning the nature of selection against TEs: 1) the gene-disruption hypothesis [3, 242], 2) the ectopic recombination hypothesis [243, 244], 3) the deleterious TE-product expression hypothesis [245].

The gene disruption hypothesis assumes that insertions inside genes or regulatory regions are under strong purifying selection because of their negative effect on the host fitness [242]. A large amount of work supports this hypothesis, demonstrating a depletion of TE insertions in exons and untranslated regions [172, 246–248]. Moreover, Lee and Karpen demonstrated that repressive histone marks affecting euchromatic TEs can spread up to 20 kb both in *D. melanogaster* and *D. simulans*, and that this phenomenon is associated with selection against TEs [12]. Therefore, we may extend this hypothesis beyond insertions inside genes or regulatory regions to include insertions close to genes.

The ectopic recombination hypothesis states that purifying selection acts against chromosomal rearrangements resulting from recombination events between TE sequences showing sequence identity and located at distinct loci [243, 244]. According to this hypothesis, TE size, TE family copy number, and

meiotic recombination rate, expected to be positively correlated with ectopic recombination rate, should be associated with the strength of purifying selection [137]. First, since long insertions provide longer targets for recombination, one can indeed expect a stronger effect of purifying selection against long TEs in the ectopic recombination hypothesis. The negative correlation between TE size and population frequencies suggests that it is actually the case [172, 249]. Second, because ectopic recombination is more likely to occur when TEs are heterozygous, ectopic recombination should happen more frequently for TE families with a high copy number of polymorphic TEs. Therefore, the negative correlation between TE insertion frequencies and copy numbers also supports the ectopic recombination hypothesis [172, 249]. Finally, because ectopic recombination is intrinsically related to the local recombination rate, the fact that low-recombining regions are highly enriched in TEs, and that a negative correlation exists between insertion frequencies and recombination rate [172, 246, 249, 250], constitute one more argument in favor of the ectopic recombination hypothesis. However, this last point may be explained by the Hill-Robertson effect, or the lower density of genes in low-recombining regions, or both. The Hill-Robertson effect corresponds to a reduction in the efficiency of selection on a locus due to selection on related loci. If slightly deleterious insertions are close to adaptive mutations, they will be less efficiently removed in low-recombining regions than in high-recombining regions. The lower density of genes in low-recombining regions may explain the higher TE density in these regions because one may expect that TE insertions are strongly counter-selected close to genes (gene disruption hypothesis). However, one paradox exists when considering the ectopic recombination hypothesis. Indeed, considering the higher rate of recombination on the X chromosome, and the ectopic recombination hypothesis, TE density should be lower on the X chromosome [251]. However, recent studies of *D. melanogaster* natural populations show different results. TE density was found to be either higher on the X chromosome [246], or similar between the X chromosome and autosomes when taking into account differences in the amount of low recombining regions [172]. A higher transposition rate in the X chromosome relatively to autosomes has been proposed as a plausible explanation to the observed paradox [137]. Mutation accumulation data recently showed such tendency with a 1.86 fold change for insertion rate on the X chromosome relatively to autosomes [128].

One last hypothesis remains concerning the nature of the purifying selection affecting TEs: the deleterious TE-product expression hypothesis [245]. Under this model, transcription and translation of TEs may be resource consuming for the host and TE proteins could disrupt cellular processes. According to this hypothesis, and assuming that full length TEs are more transcribed than nearly complete copies, one may expect complete copies to be under more intense purifying selection than nearly complete copies. However, Petrov and colleagues did not find such effect investigating TE frequencies genome wide [249].

Models of TE dynamics

So far, two main models have been formulated to conceptualize TE dynamics in *Drosophila* populations. The historical model is the transposition-selection balance model: it assumes that TE abundance is regulated by a balance between transposition and selection against TEs [3, 252]. According to this model, insertions with low frequency in populations are expected to be mainly insertions subjected to strong purifying selection. However, because transposition rates are not constant over time, another model has been proposed: the transposition burst model [175]. This model proposes that TE dynamics in populations is explained by transposition bursts. Under this hypothesis, a large proportion of low frequency insertions may result from recent TE activity rather than strong selection against TEs. Data, especially on TE genomic distribution (see above), suggest a preeminent role of purifying selection in TE dynamics, and thus support the transposition-selection balance model. Furthermore, an excess of rare TEs compared to the standard neutral model is found, as expected if selection acts against TEs [246]. However, confronting population data with simulation, Kofler and colleagues showed that both in *D. melanogaster* and *D. simulans*, 50% of families have temporally heterogeneous transposition rates and that a correlation exists between insertion frequencies and their age [155, 172]. So far, it is clear that both purifying selection and variation in transposition rate act on TE population dynamics. Until now, TE regulation has been poorly integrated in the models of TE dynamics. In 2010, Lu and colleagues incorporated piRNAs in a population genetics framework [253]. They used simulations to investigate the dynamics of TEs. They focused on retrotransposons, studying the retrotransposons that are targeted by piRNAs but also the retrotransposons generating piRNAs. The results indicate that: piRNAs may reduce TE fitness cost; TEs generating piRNAs may easily reach fixation because they confer a selective advantage; and TEs targeted by piRNAs may also reach fixation because host defenses reduce their deleterious effect. In 2013, the observation that a TE insertion inside a piRNA cluster was able to silence the corresponding TE family led to the formulation of the trap model

[197]. In this model, after invasion of a host genome, a TE family proliferates until it is trapped, *i.e.* one insertion occurs into a piRNA cluster, then the subsequent production of piRNAs silences the invading family. This model was validated and enriched with populational considerations by Kofler and colleagues [88]. Monitoring the *P-Element* invasion, in connection with the piRNA pathway, in experimentally evolving populations of *D. simulans*, they suggested the following three-step model for a TE invasion: 1) TE copies colonize the genome, 2) the first TE insertions in piRNA clusters occur but are not yet sufficient to stop TE proliferation and 3) the TE family is inactivated by the fixation of an insertion within a piRNA cluster. Using simulated data, they were able to demonstrate that this "trap model" accurately describes TE abundance in *D. melanogaster* germline. They also showed that the suppression of TE activity by segregating cluster insertions is reversible. Importantly, they demonstrated that transposition rates and population sizes affected mostly the duration of the invasion steps but not the amounts of accumulating TEs. In fact, the major factor capable of affecting the number of accumulating TEs was the piRNA cluster size.

Conclusions

In today's biology research, increasing weight is given to the study of non-model species. This is clearly justified by the diversity of the living world, and even more so for the study of genetic elements as diverse and dynamic as TEs. However, we should not overlook model organisms, because the vast amount of techniques, data collected and knowledge will help us develop and test new hypotheses. Furthermore, the dissection of conserved pathways in these organisms, such as the piRNA pathway, should provide results valid for a broad range of species. Despite the fact that *Drosophila* is an old biological model, it still presents many opportunities for TE research. In general, studies of TEs could benefit from unified approaches to identifying and quantifying TEs. As we demonstrated above, the ultimate model *D. melanogaster* appears slightly different from its sister species regarding TEs —maybe related to the fact that it ended up as the ultimate model species— however, it is clear that the research community greatly benefits from comparative genomics in the *Drosophila* genus, and a great deal of work remains to be done in *Drosophila* and the species in the group in order to do proper comparative genomics. It is clear that the development of long-read technologies will greatly facilitate this work. Another challenge is to understand the activity of TEs and how, *in natura*, this activity is triggered and controlled. Once again, *Drosophila* is a model of excellence with the possibility of doing experimental evolution with a follow-up of TE dynamics. At the same time, this will allow a better understanding of the fine regulation systems of TE activity. Finally, it seems to us that one of the most exciting challenges is to understand the true impact of TEs in adaptive processes, even more so now, with all the gross changes in our environment. Experimental evolution, with different species and different environmental factors, are a real opportunity to move forward in this field.

Abbreviations

Ago2: Argonaute2; Ago3: Argonaute3; Aub: Aubergine; Dcr-2: Dicer-2; DIRS: Dictyostelium Intermediate Repeat Sequence; Endo: Endonuclease; Env: Envelope gene; ERV: Endogenous RetroViruses; Int: Integrase; Hel: DNA helicase; LINE: Long Interspersed Nuclear Element; Loqs-PD: Loquacious-PD; LTR: Long Terminal Repeat; piRNA: piwi-interacting RNA; PLE: Penelope-Like Element; Prot: Protease; rasiRNA: repeat associated small interfering RNA; Rep: Replicator; RISC: *RNA-induced silencing complex*; *RNP*: *RiboNucleoProtein*; RT: Reverse Transcriptase; Tase: Transposase; TSD: Target Site Duplication; TE: Transposable Element; TIR: Terminal Inverted Repeat; SINE: Short Interspersed Nuclear Element; siRNA: small interfering RNA; Zuc: Zucchini

Acknowledgments

The authors sincerely thank the anonymous reviewers.

Authors' contributions

VM has drafted the initial version of the review and designed the figures; MB, MF and CV have contributed to the writing of the manuscript. All authors have approved the final version.

Funding

This work was supported by the ANR Exhyb and ANR SWING (grant overseen by the French National Research Agency) and the CNRS.

Availability of data and materials

The datasets analyzed during the current study are available in the following repositories: https://github.com/danrdanny/Drosophila15GenomesProject/raw/master/assembledGenomes/ [99], ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.29_FB2019_04/

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests

Received: 16 February 2020 Accepted: 14 April 2020
Published online: 03 July 2020

References

1. Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. Nature. 1980;284:601–3.
2. Orgel LE, Crick FHC. Selfish DNA: the ultimate parasite. Nature. 1980;284:604–7.
3. Charlesworth B, Charlesworth D. The population dynamics of transposable elements. Genet Res. 1983;42:1–27.
4. Feschotte C, Jiang N, Wessler SR. Plant transposable elements: where genetics meets genomics. Nat Rev Genet. 2002;3:329–41.
5. Ravindran S. Barbara McClintock and the discovery of jumping genes. Proc Natl Acad Sci. 2012;109:20198–9.
6. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000;408:796.
7. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420:520.
8. C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science. 1998;282:2012–8.

9. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
10. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009;326:1112–5.
11. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.
12. Lee YCG, Karpen GH. Pervasive epigenetic effects of Drosophila euchromatic transposable elements impact their evolution. eLife. 2017;6 https://doi.org/10.7554/eLife.25762.
13. Singh BN. Species and genetic diversity in the genus Drosophila inhabiting the Indian subcontinent. J Genet. 2015;94:351–61.
14. Obbard DJ, Maclennan J, Kim K-W, Rambaut A, O'Grady PM, Jiggins FM. Estimating divergence dates and substitution rates in the Drosophila phylogeny. Mol Biol Evol. 2012;29:3459–73.
15. Keller A. Drosophila melanogaster's history as a human commensal. Curr Biol CB. 2007;17:R77–81.
16. Markow TA. The secret lives of Drosophila flies. eLife. 2015;4 https://doi.org/10.7554/eLife.06793.
17. David JR, Capy P. Genetic variation of Drosophila melanogaster natural populations. Trends Genet TIG. 1988;4:106–11.
18. Hales KG, Korey CA, Larracuente AM, Roberts DM. Genetics on the Fly: A Primer on the Drosophila Model System. Genetics. 2015;201:815–42.
19. Rubin GM, Spradling AC. Genetic transformation of Drosophila with transposable element vectors. Science. 1982;218:348–53.
20. Sessegolo C, Burlet N, Haudry A. Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. Biol Lett. 2016;12:20160407.
21. Genome List - Genome - NCBI. https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/drosophila. Accessed 4 Aug 2019.
22. Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, et al. The Drosophila melanogaster Genetic Reference Panel. Nature. 2012;482:173–8.
23. Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, Pool JE, et al. Genomic Variation in Natural Populations of Drosophila melanogaster. Genetics. 2012;192:533–98.
24. Huang W, Massouras A, Inoue Y, Peiffer J, Ràmia M, Tarone AM, et al. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome Res. 2014;24:1193–208.
25. Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, et al. Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. PLoS Genet. 2012;8 https://doi.org/10.1371/journal.pgen.1003080.
26. Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, et al. The Drosophila Genome Nexus: A Population Genomic Resource of 623 Drosophila melanogaster Genomes, Including 197 from a Single Ancestral Range Population. Genetics. 2015;199:1229–41.
27. Machado HE, Bergland AO, Taylor R, Tilk S, Behrman E, Dyer K, et al. Broad geographic sampling reveals predictable, pervasive, and strong seasonal adaptation in Drosophila. bioRxiv. 2019:337543. https://doi.org/10.1101/337543.
28. Kapun M, Barrón MG, Staubach F, Vieira J, Obbard DJ, Goubert C, et al. Genomic analysis of European Drosophila melanogaster populations on a dense spatial scale reveals longitudinal population structure and continent-wide selection. bioRxiv. 2019:313759. https://doi.org/10.1101/313759.
29. Grenier JK, Arguello JR, Moreira MC, Gottipati S, Mohammed J, Hackett SR, et al. Global diversity lines - a five-continent reference panel of sequenced Drosophila melanogaster strains. G3 Bethesda Md. 2015;5:593–603.
30. Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE. A Thousand Fly Genomes: An Expanded Drosophila Genome Nexus. Mol Biol Evol. 2016;33:3308–13.
31. Ayroles JF, Carbone MA, Stone EA, Jordan KW, Lyman RF, Magwire MM, et al. Systems genetics of complex traits in Drosophila melanogaster. Nat Genet. 2009;41:299–307.
32. Shorter J, Couch C, Huang W, Carbone MA, Peiffer J, Anholt RRH, et al. Genetic architecture of natural variation in Drosophila melanogaster aggressive behavior. Proc Natl Acad Sci U S A. 2015;112:E3555–63.
33. Durham MF, Magwire MM, Stone EA, Leips J. Genome-wide analysis in Drosophila reveals age-specific effects of SNPs on fitness traits. Nat Commun. 2014;5:4338.
34. Weber AL, Khan GF, Magwire MM, Tabor CL, Mackay TFC, Anholt RRH. Genome-wide association analysis of oxidative stress resistance in Drosophila melanogaster. PLoS One. 2012;7:e34745.
35. Bloomington Drosophila Stock Center. Bloomingt. Drosoph. Stock Cent. https://bdsc.indiana.edu/stocks/stockdata.html. Accessed 27 Oct 2019.
36. The National Drosophila Species Stock Center | College of Agriculture and Life Science. http://blogs.cornell.edu/drosophila/. Accessed 4 Aug 2019.
37. Lachaise D, Cariou M-L, David JR, Lemeunier F, Tsacas L, Ashburner M. Historical biogeography of the Drosophila melanogaster species subgroup. Hist Biogeogr Drosoph Melanogaster Species Subgr. 1988;22:159–225.
38. McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. Regulatory divergence in Drosophila revealed by mRNA-seq. Genome Res. 2010;20:816–25.
39. Coolon JD, McManus CJ, Stevenson KR, Graveley BR, Wittkopp PJ. Tempo and mode of regulatory evolution in Drosophila. Genome Res. 2014;24:797–808.
40. Meiklejohn CD, Coolon JD, Hartl DL, Wittkopp PJ. The roles of cis- and trans-regulation in the evolution of regulatory incompatibilities and sexually dimorphic gene expression. Genome Res. 2014;24:84–95.
41. Bono JM, Markow TA. Post-zygotic isolation in cactophilic Drosophila: larval viability and adult life-history traits of D. mojavensis/D. arizonae hybrids. J Evol Biol. 2009;22:1387–95.
42. Lohse K, Clarke M, Ritchie MG, Etges WJ. Genome-wide tests for introgression between cactophilic Drosophila implicate a role of inversions during speciation. Evol Int J Org Evol. 2015;69:1178–90.
43. Sanchez-Flores A, Peñaloza F, Carpinteyro-Ponce J, Nazario-Yepiz N, Abreu-Goodger C, Machado CA, et al. Genome Evolution in Three Species of Cactophilic Drosophila. G3 GenesGenomesGenetics. 2016;6:3097–105.
44. Lopez-Maestre H, Carnelossi EAG, Lacroix V, Burlet N, Mugat B, Chambeyron S, et al. Identification of misexpressed genetic elements in hybrids between Drosophila-related species. Sci Rep. 2017;7:40618.
45. Vela D, Fontdevila A, Vieira C, García Guerreiro MP. A genome-wide survey of genetic instability by transposition in Drosophila hybrids. PLoS One. 2014;9:e88992.
46. García Guerreiro MP. Changes of Osvaldo expression patterns in germline of male hybrids between the species Drosophila buzzatii and Drosophila koepferae. Mol Genet Genomics MGG. 2015;290:1471–83.
47. Romero-Soriano V, Modolo L, Lopez-Maestre H, Mugat B, Pessia E, Chambeyron S, et al. Transposable Element Misregulation Is Linked to the Divergence between Parental piRNA Pathways in Drosophila Hybrids. Genome Biol Evol. 2017;9:1450–70.
48. Green MM. Transposable Elements in Drosophila and Other Diptera. Annu Rev Genet. 1980;14:109–20.
49. Ananiev EV, Ilyin YV. A comparative study of the location of mobile dispersed genes in salivary gland and midgut polytene chromosomes of Drosophila melanogaster. Chromosoma. 1981;82:429–35.
50. Picard G. Non-Mendelian Female Sterility in DROSOPHILA MELANOGASTER : Hereditary Transmission of I Factor. Genetics. 1976;83:107–23.
51. Kidwell MG, Kidwell JF, Sved JA. Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: A Syndrome of Aberrant Traits Including Mutation, Sterility and Male Recombination. Genetics. 1977;86:813–33.
52. Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet. 2008;9:411–2 author reply 414.
53. Seberg O, Petersen G. A unified classification system for eukaryotic transposable elements should reflect their phylogeny. Nat Rev Genet. 2009;10:276.
54. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. Reply: A unified classification system for eukaryotic transposable elements should reflect their phylogeny. Nat Rev Genet. 2009;10:276.
55. Hill T. Transposable element dynamics are consistent across the Drosophila phylogeny, despite drastically differing content. bioRxiv. 2019:651059. https://doi.org/10.1101/651059.
56. GIRI. https://www.girinst.org/repbase/. Accessed 16 Feb 2020.
57. Mount SM, Rubin GM. Complete nucleotide sequence of the Drosophila transposable element copia: homology between copia and retroviral proteins. Mol Cell Biol. 1985;5:1630–8.
58. Marlor RL, Parkhurst SM, Corces VG. The Drosophila melanogaster gypsy transposable element encodes putative gene products homologous to retroviral proteins. Mol Cell Biol. 1986;6:1129–34.
59. Lindsley DL, Zimm GG. The Genome of Drosophila Melanogaster. San Diego: Academic Press; 2012.
60. McCullers TJ, Steiniger M. Transposable elements in Drosophila. Mob Genet Elem. 2017;7:1–18.

61. Dunn BM, Goodenow MM, Gustchina A, Wlodawer A. Retroviral proteases. Genome Biol. 2002;3 https://doi.org/10.1186/gb-2002-3-4-reviews3006.
62. Shiba T, Saigo K. Retrovirus-like particles containing RNA homologous to the transposable element copia in Drosophila melanogaster. Nature. 1983;302: 119–24.
63. Arkhipova IR, Mazo AM, Cherkasova VA, Gorelova TV, Schuppe NG, Llyin YV. The steps of reverse transcription of Drosophila mobile dispersed genetic elements and U3-R-U5 structure of their LTRs. Cell. 1986;44:555–63.
64. Desset S, Conte C, Dimitri P, Calco V, Dastugue B, Vaury C. Mobilization of two retroelements, ZAM and Idefix, in a novel unstable line of Drosophila melanogaster. Mol Biol Evol. 1999;16:54–66.
65. Akkouche A, Grentzinger T, Fablet M, Armenise C, Burlet N, Braman V, et al. Maternally deposited germline piRNAs silence the tirant retrotransposon in somatic cells. EMBO Rep. 2013;14:458–64.
66. Terzian C, Pélisson A, Bucheton A. Evolution and phylogeny of insect endogenous retroviruses. BMC Evol Biol. 2001;1:3.
67. Kim A, Terzian C, Santamaria P, Pelisson A, Purd'homme N, Bucheton A. Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of Drosophila melanogaster. Proc Natl Acad Sci. 1994; 91:1285–9.
68. Leblanc P, Desset S, Giorgi F, Taddei AR, Fausto AM, Mazzini M, et al. Life Cycle of an Endogenous Retrovirus, ZAM, in Drosophila melanogaster. J Virol. 2000;74:10658–69.
69. Bucheton A, Paro R, Sang HM, Pelisson A, Finnegan DJ. The molecular basis of I-R hybrid Dysgenesis in Drosophila melanogaster: Identification, cloning, and properties of the I factor. Cell. 1984;38:153–63.
70. Fawcett DH, Lister CK, Kellett E, Finnegan DJ. Transposable elements controlling I-R hybrid dysgenesis in D. melanogaster are similar to mammalian LINEs. Cell. 1986;47:1007–15.
71. Priimägi AF, Mizrokhi LJ, Ilyin YV. The Drosophila mobile element jockey belongs to LINEs and contains coding sequences homologous to some retroviral proteins. Gene. 1988;70:253–62.
72. Dawson A, Hartswood E, Paterson T, Finnegan DJ. A LINE-like transposable element in Drosophila, the I factor, encodes a protein with properties similar to those of retroviral nucleocapsids. EMBO J. 1997;16:4448–55.
73. Martin SL, Bushman FD. Nucleic Acid Chaperone Activity of the ORF1 Protein from the Mouse LINE-1 Retrotransposon. Mol Cell Biol. 2001;21:467–75.
74. Finnegan DJ. Transposable elements: How non-LTR retrotransposons do it. Curr Biol. 1997;7:R245–8.
75. Pélisson A, Finnegan DJ, Bucheton A. Evidence for retrotransposition of the I factor, a LINE element of Drosophila melanogaster. Proc Natl Acad Sci U S A. 1991;88:4907–10.
76. del Carmen SM, Disson O, Robin S, Brun C, Teninges D, Bucheton A. In vivo RNA localization of I factor, a non-LTR retrotransposon, requires a cis-acting signal in ORF2 and ORF1 protein. Nucleic Acids Res. 2005;33:776–85.
77. Han JS. Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. Mob DNA. 2010;1:15.
78. Petrov DA, Hartl DL. High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. Mol Biol Evol. 1998;15:293–302.
79. Kramerov DA, Vassetzky NS. Origin and evolution of SINEs in eukaryotic genomes. Heredity. 2011;107:487–95.
80. Yannopoulos G, Stamatis N, Monastirioti M, Hatzopoulos P, Louis C. hobo is responsible for the induction of hybrid dysgenesis by strains of Drosophila melanogaster bearing the male recombination factor 23.5MRF. Cell. 1987;49: 487–95.
81. Hill T, Schlötterer C, Betancourt AJ. Hybrid Dysgenesis in Drosophila simulans Associated with a Rapid Invasion of the P-Element. PLoS Genet. 2016;12:e1005920.
82. Petrov DA, Schutzman JL, Hartl DL, Lozovskaya ER. Diverse transposable elements are mobilized in hybrid dysgenesis in Drosophila virilis. Proc Natl Acad Sci U S A. 1995;92:8050–4.
83. Evgen'ev MB, Zelentsova H, Shostak N, Kozitsina M, Barskyi V, Lankenau DH, et al. Penelope, a new family of transposable elements and its possible role in hybrid dysgenesis in Drosophila virilis. Proc Natl Acad Sci U S A. 1997;94: 196–201.
84. Vieira J, Vieira CP, Hartl DL, Lozovskaya ER. Factors contributing to the hybrid dysgenesis syndrome in Drosophila virilis. Genet Res. 1998;71:109–17.
85. Blumenstiel JP. Whole genome sequencing in Drosophila virilis identifies Polyphemus, a recently activated Tc1-like transposon with a possible role in hybrid dysgenesis. Mob DNA. 2014;5:6.
86. Chambeyron S, Popkova A, Payen-Groschene G, Brun C, Laouini D, Pelisson A, et al. piRNA-mediated nuclear accumulation of retrotransposon transcripts in the Drosophila female germline. Proc Natl Acad Sci. 2008;105:14964–9.
87. Daniels SB, Peterson KR, Strausbaugh LD, Kidwell MG, Chovnick A. Evidence for Horizontal Transmission of the P Transposable Element between Drosophila Species. Genetics. 1990;124:339–55.
88. Kofler R, Senti K-A, Nolte V, Tobler R, Schlötterer C. Molecular dissection of a natural transposable element invasion. Genome Res. 2018;28:824–35.
89. Evgen'ev MB. What happens when Penelope comes? Mob Genet Elem. 2013;3:e24542.
90. Jacobson JW, Medhora MM, Hartl DL. Molecular structure of a somatically unstable transposable element in Drosophila. Proc Natl Acad Sci U S A. 1986;83:8684–8.
91. O'Hare K, Rubin GM. Structures of P transposable elements and their sites of insertion and excision in the Drosophila melanogaster genome. Cell. 1983; 34:25–35.
92. Zhang L, Dawson A, Finnegan DJ. DNA-binding activity and subunit interaction of the mariner transposase. Nucleic Acids Res. 2001;29:3566–75.
93. Tang M, Cecconi C, Kim H, Bustamante C, Rio DC. Guanosine triphosphate acts as a cofactor to promote assembly of initial P-element transposase–DNA synaptic complexes. Genes Dev. 2005;19:1422.
94. Chen J, Greenblatt IM, Dellaporta SL. Molecular Analysis of Ac Transposition and DNA Replication. Genetics. 1992;130:665–76.
95. Nassif N, Penney J, Pal S, Engels WR, Gloor GB. Efficient copying of nonhomologous sequences from ectopic sites via P-element-induced gap repair. Mol Cell Biol. 1994;14:1613–25.
96. Kapitonov VV, Jurka J. Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci U S A. 2001;98:8714–9.
97. Kapitonov VV, Jurka J. Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet. 2007;23:521–9.
98. Grabundzija I, Hickman AB, Dyda F. Helraiser intermediates provide insight into the mechanism of eukaryotic replicative transposition. Nat Commun. 2018;9 https://doi.org/10.1038/s41467-018-03688-w.
99. Miller DE, Staber C, Zeitlinger J, Hawley RS. Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing. G3 GenesGenomesGenetics. 2018;8:3131–41.
100. Smit A, Hubley R, Green P. RepeatMasker Home Page. 2013. http://www.repeatmasker.org/. Accessed 30 Jan 2020.
101. Bailly-Bechet M, Haudry A, Lerat E. "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. Mob DNA. 2014;5:13.
102. Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M. De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes aegypti). Genome Biol Evol. 2015;7:1192–205.
103. Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, et al. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol. 2002;3:research0084.1–2.
104. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in de novo annotation approaches. PLoS One. 2011; 6:e16526.
105. Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007;450:203–18.
106. Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, et al. The Release 6 reference sequence of the Drosophila melanogaster genome. Genome Res. 2015;25:445–58.
107. Dfam. https://www.dfam.org/family/DF0001586/summary. Accessed 2 Feb 2020.
108. Thomas J, Vadnagara K, Pritham EJ. DINE-1, the highest copy number repeats in Drosophila melanogaster are non-autonomous endonuclease-encoding rolling-circle transposable elements (Helentrons). Mob DNA. 2014;5:18.
109. Lerat E. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity. 2010;104:520–33.
110. Rius N, Guillén Y, Delprat A, Kapusta A, Feschotte C, Ruiz A. Exploration of the Drosophila buzzatii transposable element content suggests underestimation of repeats in Drosophila genomes. BMC Genomics. 2016;17 https://doi.org/10.1186/s12864-016-2648-8.
111. Chiu JC, Jiang X, Zhao L, Hamm CA, Cridland JM, Saelao P, et al. Genome of Drosophila suzukii, the Spotted Wing Drosophila. G3 GenesGenomesGenetics. 2013;3:2257–71.

112. Paris M, Boyer R, Jaenichen R, Wolf J, Karageorgi M, Green J, et al. Near-chromosome level genome assembly of the fruit pest Drosophila suzukii using long-read sequencing. bioRxiv. 2020. https://doi.org/10.1101/2020.01.02.892844.
113. Repbase Reports - 2012, Volume 12, Issue 9. https://www.girinst.org/2012/vol12/issue9/. Accessed 31 Jan 2020.
114. Kidwell MG, Evgen'ev MB. How valuable are model organisms for transposable element studies? Genetica. 1999;107:103.
115. Vieira C, Nardon C, Arpin C, Lepetit D, Biémont C. Evolution of Genome Size in Drosophila. Is the Invader's Genome Being Invaded by Transposable Elements? Mol Biol Evol. 2002;19:1154–61.
116. Bingham PM, Kidwell MG, Rubin GM. The molecular basis of P-M hybrid dysgenesis: The role of the P element, a P-strain-specific transposon family. Cell. 1982;29:995–1004.
117. Kidwell MG. Evolution of hybrid dysgenesis determinants in Drosophila melanogaster. Proc Natl Acad Sci U S A. 1983;80:1655–9.
118. Anxolabéhère D, Kidwell MG, Periquet G. Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of Drosophila melanogaster by mobile P elements. Mol Biol Evol. 1988;5:252–69.
119. Bonnivard H. Stability of European natural populations of Drosophila melanogaster with regard to the P -M system: a buffer zone made up of Q populations. J Evol Biol. 1999;12:633–47.
120. Kofler R, Hill T, Nolte V, Betancourt AJ, Schlötterer C. The recent invasion of natural Drosophila simulans populations by the P-element. Proc Natl Acad Sci. 2015;112:6659–63.
121. Panaud O. Horizontal transfers of transposable elements in eukaryotes: The flying genes. C R Biol. 2016;339:296–9.
122. Bartolomé C, Bello X, Maside X. Widespread evidence for horizontal transfer of transposable elements across Drosophila genomes. Genome Biol. 2009;10:R22.
123. Vieira C, Lepetit D, Dumont S, Biémont C. Wake up of transposable elements following Drosophila simulans worldwide colonization. Mol Biol Evol. 1999;16:1251–5.
124. Biémont C, Vieira C, Borie N, Lepetit D. Transposable elements and genome evolution: the case of Drosophila simulans. Genetica. 1999;107:113–20.
125. Vieira C, Biémont C. Geographical variation in insertion site number of retrotransposon 412 in Drosophila simulans. J Mol Evol. 1996;42:443.
126. García Guerreiro MP, Chávez-Sandoval BE, Balanyà J, Serra L, Fontdevila A. Distribution of the transposable elements bilbo and gypsy in original and colonizing populations of Drosophila subobscura. BMC Evol Biol. 2008;8:234.
127. García Guerreiro MP, Fontdevila A. Osvaldo and Isis retrotransposons as markers of the Drosophila buzzatii colonisation in Australia. BMC Evol Biol. 2011;11 https://doi.org/10.1186/1471-2148-11-111.
128. Adrion JR, Song MJ, Schrider DR, Hahn MW, Schaack S. Genome-Wide Estimates of Transposable Element Insertion and Deletion Rates in Drosophila Melanogaster. Genome Biol Evol. 2017;9:1329–40.
129. Keightley PD, Ness RW, Halligan DL, Haddrill PR. Estimation of the spontaneous mutation rate per nucleotide site in a Drosophila melanogaster full-sib family. Genetics. 2014;196:313–20.
130. Harada K, Yukuhiro K, Mukai T. Transposition rates of movable genetic elements in Drosophila melanogaster. Proc Natl Acad Sci U S A. 1990;87:3248–52.
131. Nuzhdin SV, Mackay TF. The genomic rate of transposable element movement in Drosophila melanogaster. Mol Biol Evol. 1995;12:180–1.
132. Maside X, Bartolomé C, Assimacopoulos S, Charlesworth B. Rates of movement and distribution of transposable elements in Drosophila melanogaster: in situ hybridization vs Southern blotting data. Genet Res. 2001;78:121–36.
133. Biémont C, Aouar A, Arnault C. Genome reshuffling of the copia element in an inbred line of Drosophila melanogaster. Nature. 1987;329:742–4.
134. Pasyukova EG, Nuzhdin SV. Doc and copia instability in an isogenic Drosophila melanogaster stock. Mol Gen Genet MGG. 1993;240:302–6.
135. Díaz-González J, Vázquez JF, Albornoz J, Domínguez A. Long-term evolution of the roo transposable element copy number in mutation accumulation lines of Drosophila melanogaster. Genet Res. 2011;93:181–7.
136. Guerreiro MPG. What makes transposable elements move in the Drosophila genome? Heredity. 2012;108:461–8.
137. Barrón MG, Fiston-Lavier A-S, Petrov DA, González J. Population Genomics of Transposable Elements in Drosophila. Annu Rev Genet. 2014;48:561–81.
138. Biémont C, Arnault C, Heizmann A, Ronsseray S. Massive changes in genomic locations of P elements in an inbred line of Drosophila melanogaster. Naturwissenschaften. 1990;77:485–8.
139. Vasilyeva LA, Bubenshchikova EV, Ratner VA. Heavy heat shock induced retrotransposon transposition in Drosophila. Genet Res. 1999;74:111–9.
140. Zabanov SA, Vasil'eva LA, Ratner VA. Induction of transposition of MGE Dm412 using gamma-irradiation of an isogenic line of Drosophila melanogaster. Genetika. 1995;31:798–803.
141. Vasil'eva LA, Ratner VA, Antonenko OV, Lopukhova ED, Bubenshchikova EV. Induction of MGE 412 transposition in an isogenic strain of Drosophila melanogaster by different doses of ethanol fumes. Genetika. 2003;39:717–20.
142. Nabirochkin SD, Gabitova L, Ossokina MA, Soldatov AV, Gazaryan TG, Gazaryan KG. Oncoviral DNAs induce transposition of endogenous mobile elements in the genome of Drosophila melanogaster. Mutat Res Mol Mech Mutagen. 1998;403:127–36.
143. Horváth V, Merenciano M, González J. Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response. Trends Genet. 2017;33:832–41.
144. Baack EJ, Whitney KD, Rieseberg LH. Hybridization and genome size evolution: timing and magnitude of nuclear DNA content increases in Helianthus homoploid hybrid species. New Phytol. 2005;167:623–30.
145. Metcalfe CJ, Bulazel KV, Ferreri GC, Schroeder-Reiter E, Wanner G, Rens W, et al. Genomic instability within centromeres of interspecific marsupial hybrids. Genetics. 2007;177:2507–17.
146. Capy P, Gasperi G, Biémont C, Bazin C. Stress and transposable elements: co-evolution or useful parasites? Heredity. 2000;85:101–6.
147. Strand DJ, McDonald JF. Copia is transcriptionally responsive to environmental stress. Nucleic Acids Res. 1985;13:4401–10.
148. Chakrani F, Capy P, David J. Developmental temperature and somatic excision rate of mariner transposable element in three natural populations of Drosophila simulans. Genet Sel Evol. 1993;25:121.
149. Jakšić AM, Kofler R, Schlötterer C. Regulation of transposable elements: Interplay between TE-encoded regulatory sequences and host-specific trans-acting factors in Drosophila melanogaster. Mol Ecol. 2017;26:5149–59.
150. Yang H-P, Hung T-L, You T-L, Yang T-H. Genomewide Comparative Analysis of the Highly Abundant Transposable Element DINE-1 Suggests a Recent Transpositional Burst in Drosophila yakuba. Genetics. 2006;173:189–96.
151. Lerat E, Goubert C, Guirao-Rico S, Merenciano M, Dufour A-B, Vieira C, et al. Population-specific dynamics and selection patterns of transposable element insertions in European natural populations. Mol Ecol. 2019;28:1506–22.
152. Giraud T, Capy P. Somatic activity of the mariner transposable element in natural populations of Drosophila simulans. Proc Biol Sci. 1996;263:1481–6.
153. Lerat E, Burlet N, Biémont C, Vieira C. Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. Gene. 2011;473:100–9.
154. Kapusta A, Suh A. Evolution of bird genomes—a transposon's-eye view. Ann N Y Acad Sci. 2017;1389:164–85.
155. Kofler R, Nolte V, Schlötterer C. Tempo and Mode of Transposable Element Activity in Drosophila. PLoS Genet. 2015;11:e1005406.
156. Burt A, Trivers R. Genes in conflict: the biology of selfish genetic elements. Cambridge: Belknap Press of Harvard University Press; 2006.
157. Lim JK, Simmons MJ. Gross chromosome rearrangements mediated by transposable elements in Drosophila melanogaster. BioEssays. 1994;16:269–75.
158. Lyttle TW, Haymer DS. The role of the transposable element hobo in the origin of endemic inversions in wild populations of Drosophila melanogaster. Genetica. 1992;86:113–26.
159. Cáceres M, Ranz JM, Barbadilla A, Long M, Ruiz A. Generation of a Widespread Drosophila Inversion by a Transposable Element. Science. 1999;285:415–8.
160. Delprat A, Negre B, Puig M, Ruiz A. The Transposon Galileo Generates Natural Chromosomal Inversions in Drosophila by Ectopic Recombination. PLoS ONE. 2009;4 https://doi.org/10.1371/journal.pone.0007883.
161. Evgen'ev MB, Zelentsova H, Poluectova H, Lyozin GT, Veleikodvorskaja V, Pyatkov KI, et al. Mobile Elements and Chromosomal Evolution in the Virilis Group of Drosophila. Proc Natl Acad Sci U S A. 2000;97:11337–42.
162. Spradling AC, Stern DM, Kiss I, Roote J, Laverty T, Rubin GM. Gene disruptions using P transposable elements: an integral component of the Drosophila genome project. Proc Natl Acad Sci. 1995;92:10824–30.

163. Spradling AC, Stern D, Beaton A, Rhem EJ, Laverty T, Mozden N, et al. The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. Genetics. 1999; 153:135–77.
164. Bellen HJ, Levis RW, Liao G, He Y, Carlson JW, Tsang G, et al. The BDGP Gene Disruption Project: Single Transposon Insertions Associated With 40% of Drosophila Genes. Genetics. 2004;167:761–81.
165. Rebollo R, Romanish MT, Mager DL. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. Annu Rev Genet. 2012;46:21–42.
166. Guio L, Barrón MG, González J. The transposable element Bari-Jheh mediates oxidative stress response in Drosophila. Mol Ecol. 2014;23:2020–30.
167. Guio L, Vieira C, González J. Stress affects the epigenetic marks added by natural transposable element insertions in Drosophila melanogaster. Sci Rep. 2018;8 https://doi.org/10.1038/s41598-018-30491-w.
168. Green MM. Mobile elements and spontaneous gene mutations (1988) Banbury Rep. 30: Eukaryotic transposable elements as mutagenic agents. NY Cold Spring Harb Lab 1988:41 -50.
169. Cooley L, Kelley R, Spradling A. Insertional mutagenesis of the Drosophila genome with single P elements. Science. 1988;239:1121–8.
170. Nikitin AG, Woodruff RC. Somatic movement of the mariner transposable element and lifespan of Drosophila species. Mutat Res. 1995;338:43–9.
171. Charlesworth B, Langley CH, Sniegowski PD. Transposable element distributions in Drosophila. Genetics. 1997;147:1993–5.
172. Kofler R, Betancourt AJ, Schlötterer C. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in Drosophila melanogaster. PLoS Genet. 2012;8:e1002487.
173. Rech GE, Bogaerts-Márquez M, Barrón MG, Merenciano M, Villanueva-Cañas JL, Horváth V, et al. Stress response, behavior, and development are shaped by transposable element-induced mutations in Drosophila. PLoS Genet. 2019;15:e1007900.
174. González J, Petrov DA. The Adaptive Role of Transposable Elements in the Drosophila Genome. Gene. 2009;448:124–33.
175. Bergman CM, Bensasson D. Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in Drosophila melanogaster. Proc Natl Acad Sci. 2007;104:11340–5.
176. Daborn PJ, Yen JL, Bogwitz MR, Goff GL, Feil E, Jeffers S, et al. A Single P450 Allele Associated with Insecticide Resistance in Drosophila. Science. 2002; 297:2253–6.
177. Chung H, Bogwitz MR, McCart C, Andrianopoulos A, Ffrench-Constant RH, Batterham P, et al. Cis-Regulatory Elements in the Accord Retrotransposon Result in Tissue-Specific Expression of the Drosophila melanogaster Insecticide Resistance Gene Cyp6g1. Genetics. 2007;175:1071–7.
178. Schlenke TA, Begun DJ. Strong selective sweep associated with a transposon insertion in Drosophila simulans. Proc Natl Acad Sci. 2004;101:1626–31.
179. Carareto CMA, Hernandez EH, Vieira C. Genomic regions harboring insecticide resistance-associated Cyp genes are enriched by transposable element fragments carrying putative transcription factor binding sites in two sibling Drosophila species. Gene. 2014;537:93–9.
180. Levis RW, Ganesan R, Houtchens K, Tolar LA, Sheen F. Transposons in place of telomeric repeats at a Drosophila telomere. Cell. 1993;75:1083–93.
181. Whittemore K, Vera E, Martínez-Nevado E, Sanpera C, Blasco MA. Telomere shortening rate predicts species life span. Proc Natl Acad Sci. 2019;116: 15122–7.
182. Blackburn EH. Telomerases. Annu Rev Biochem. 1992;61:113–29.
183. Young BS, Pession A, Traverse KL, French C, Pardue ML. Telomere regions in Drosophila share complex DNA sequences with pericentric heterochromatin. Cell. 1983;34:85–94.
184. Traverse KL, Pardue ML. A spontaneously opened ring chromosome of Drosophila melanogaster has acquired He-T DNA sequences at both new telomeres. Proc Natl Acad Sci U S A. 1988;85:8116–20.
185. Biessmann H, Mason JM, Ferry K, d'Hulst M, Valgeirsdottir K, Traverse KL, et al. Addition of telomere-associated HeT DNA sequences "heals"broken chromosome ends in Drosophila. Cell. 1990;61:663–73.
186. Abad JP, de Pablos B, Osoegawa K, de Jong PJ, Martín-Gallardo A, Villasante A. TAHRE, a Novel Telomeric Retrotransposon from Drosophila melanogaster, Reveals the Origin of Drosophila Telomeres. Mol Biol Evol. 2004;21:1620–4.
187. Villasante A, Abad JP, Planello R, Mendez-Lago M, Celniker SE, de Pablos B. Drosophila telomeric retrotransposons derived from an ancestral element that was recruited to replace telomerase. Genome Res. 2007;17:1909–18.
188. Pardue M-L, DeBaryshe PG. Drosophila telomeres: A variation on the telomerase theme. Fly (Austin). 2008;2:101–10.
189. Saint-Leandre B, Nguyen SC, Levine MT. Diversification and collapse of a telomere elongation mechanism. Genome Res. 2019;29:920–31.
190. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 2007;8:272–85.
191. Bewick AJ, Vogel KJ, Moore AJ, Schmitz RJ. Evolution of DNA Methylation across Insects. Mol Biol Evol. 2017;34:654–65.
192. Provataris P, Meusemann K, Niehuis O, Grath S, Misof B. Signatures of DNA Methylation across Insects Suggest Reduced DNA Methylation Levels in Holometabola. Genome Biol Evol. 2018;10:1185–97.
193. Ozata DM, Gainetdinov I, Zoch A, OCarroll D, Zamore PD. PIWI-interacting RNAs: small RNAs with big functions. Nat Rev Genet. 2019;20:89–108.
194. Czech B, Hannon GJ. One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. Trends Biochem Sci. 2016;41:324–37.
195. Chalvet F, Teysset L, Terzian C, Prud'homme N, Santamaria P, Bucheton A, et al. Proviral amplification of the Gypsy endogenous retrovirus of Drosophila melanogaster involves env-independent invasion of the female germline. EMBO J. 1999;18:2659–69.
196. Aravin AA, Naumova NM, Tulin AV, Vagin VV, Rozovsky YM, Gvozdev VA. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. Curr Biol. 2001;11: 1017–27.
197. Zanni V, Eymery A, Coiffet M, Zytnicki M, Luyten I, Quesneville H, et al. Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. Proc Natl Acad Sci U S A. 2013;110:19842–7.
198. Mevel-Ninio MT, Pelisson A, Kinder J, Campos AR, Bucheton A. The flamenco locus controls the gypsy and ZAM retroviruses and is required for Drosophila oogenesis. Genetics. 2007. https://doi.org/10.1534/genetics.106.068106.
199. Goriaux C, Théron E, Brasset E, Vaury C. History of the discovery of a master locus producing piRNAs: the flamenco/COM locus in Drosophila melanogaster. Front Genet. 2014;5 https://doi.org/10.3389/fgene.2014.00257.
200. Busson D, Gans M, Komitopoulou K, Masson M. Genetic Analysis of Three Dominant Female-Sterile Mutations Located on the X Chromosome of DROSOPHILA MELANOGASTER. Genetics. 1983;105:309–25.
201. Mével-Ninio M, Mariol M-C, Gans M. Mobilization of the gypsy and copia retrotransposons in Drosophila melanogaster induces reversion of the ovoD dominant female-sterile mutations: molecular analysis of revertant alleles. EMBO J. 1989;8:1549–58.
202. Robert V, Prudhomme N, Kim A, Bucheton A, Pélisson A. Characterization of the flamenco Region of the Drosophila melanogaster Genome. Genetics. 2001;158:701–13.
203. Prudhomme N, Gans M, Masson M, Terzian C, Bucheton A. Flamenco, a Gene Controlling the Gypsy Retrovirus of Drosophila Melanogaster. Genetics. 1995;139:697–711.
204. Sarot E, Payen-Groschêne G, Bucheton A, Pélisson A. Evidence for a piwi-dependent RNA silencing of the gypsy endogenous retrovirus by the Drosophila melanogaster flamenco gene. Genetics. 2004;166:1313–21.
205. Goriaux C, Desset S, Renaud Y, Vaury C, Brasset E. Transcriptional properties and splicing of the flamenco piRNA cluster. EMBO Rep. 2014;15:411–8.
206. Dennis C, Brasset E, Sarkar A, Vaury C. Export of piRNA precursors by EJC triggers assembly of cytoplasmic Yb-body in Drosophila. Nat Commun. 2016;7:1–12.
207. Sokolova OA, Ilyin AA, Poltavets AS, Nenasheva VV, Mikhaleva EA, Shevelyov YY, et al. Yb body assembly on the flamenco piRNA precursor transcripts reduces genic piRNA production. Mol Biol Cell. 2019;30:1544–54.
208. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, et al. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell. 2007;128:1089–103.
209. Mohn F, Handler D, Brennecke J. piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. Science. 2015;348:812–7.
210. Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, et al. A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in Drosophila. Science. 2007;315:1587–90.
211. Li C, Vagin VV, Lee S, Xu J, Ma S, Xi H, et al. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. Cell. 2009;137: 509–21.

212. Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, et al. Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. Cell. 2009;137:522–35.
213. Han BW, Wang W, Li C, Weng Z, Zamore PD. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. Science. 2015;348:817–21.
214. Saito K, Nishida KM, Mori T, Kawamura Y, Miyoshi K, Nagami T, et al. Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. Genes Dev. 2006;20:2214–22.
215. Le Thomas A, Rogers AK, Webster A, Marinov GK, Liao SE, Perkins EM, et al. Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. Genes Dev. 2013;27:390–9.
216. Darricarrère N, Liu N, Watanabe T, Lin H. Function of Piwi, a nuclear Piwi/Argonaute protein, is independent of its slicer activity. Proc Natl Acad Sci U S A. 2013;110:1297–302.
217. Iwasaki YW, Murano K, Ishizu H, Shibuya A, Iyoda Y, Siomi MC, et al. Piwi Modulates Chromatin Accessibility by Regulating Multiple Factors Including Histone H1 to Repress Transposons. Mol Cell. 2016;63:408–19.
218. Klenov MS, Lavrov SA, Korbut AP, Stolyarenko AD, Yakushev EY, Reuter M, et al. Impact of nuclear Piwi elimination on chromatin state in Drosophila melanogaster ovaries. Nucleic Acids Res. 2014;42:6208–18.
219. Rozhkov NV, Hammell M, Hannon GJ. Multiple roles for Piwi in silencing Drosophila transposons. Genes Dev. 2013;27:400–12.
220. Sienski G, Dönertas D, Brennecke J. Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. Cell. 2012;151:964–80.
221. Wang SH, Elgin SCR. Drosophila Piwi functions downstream of piRNA production mediating a chromatin-based transposon silencing mechanism in female germ line. Proc Natl Acad Sci U S A. 2011;108:21164–9.
222. Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD. A Distinct Small RNA Pathway Silences Selfish Genetic Elements in the Germline. Science. 2006;313:320–4.
223. Nagao A, Mituyama T, Huang H, Chen D, Siomi MC, Siomi H. Biogenesis pathways of piRNAs loaded onto AGO3 in the Drosophila testis. RNA N Y N. 2010;16:2503–15.
224. Quénerch'du E, Anand A, Kai T. The piRNA pathway is developmentally regulated during spermatogenesis in Drosophila. RNA N Y N. 2016;22:1044–54.
225. Ghildiyal M, Seitz H, Horwich MD, Li C, Du T, Lee S, et al. Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. Science. 2008;320:1077–81.
226. Kawamura Y, Saito K, Kin T, Ono Y, Asai K, Sunohara T, et al. Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells. Nature. 2008;453:793–7.
227. Czech B, Malone CD, Zhou R, Stark A, Schlingeheyde C, Dus M, et al. An endogenous small interfering RNA pathway in Drosophila. Nature. 2008;453:798–802.
228. Claycomb JM. Ancient endo-siRNA pathways reveal new tricks. Curr Biol CB. 2014;24:R703–15.
229. Obbard DJ, Gordon KHJ, Buck AH, Jiggins FM. The evolution of RNAi as a defence against viruses and transposable elements. Philos Trans R Soc B Biol Sci. 2009;364:99–115.
230. Kolaczkowski B, Hupalo DN, Kern AD. Recurrent adaptation in RNA interference genes across the Drosophila phylogeny. Mol Biol Evol. 2011;28:1033–42.
231. Simkin A, Wong A, Poh Y-P, Theurkauf WE, Jensen JD. Recurrent and recent selective sweeps in the piRNA pathway. Evol Int J Org Evol. 2013;67:1081–90.
232. Fablet M, Akkouche A, Braman V, Vieira C. Variable expression levels detected in the Drosophila effectors of piRNA biogenesis. Gene. 2014;537:149–53.
233. Blumenstiel JP, Erwin AA, Hemmer LW. What Drives Positive Selection in the Drosophila piRNA Machinery? The Genomic Autoimmunity Hypothesis. Yale J Biol Med. 2016;89:499–512.
234. Lewis SH, Quarles KA, Yang Y, Tanguy M, Frézal L, Smith SA, et al. Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements. Nat Ecol Evol. 2018;2:174–81.
235. Pardue ML, Gerbi SA, Eckhardt RA, Gall JG. Cytological localization of DNA complementary to ribosomal RNA in polytene chromosomes of Diptera. Chromosoma. 1970;29:268–90.
236. Biémont C, Monti-Dedieu L, Lemeunier F. Detection of transposable elements in Drosophila salivary gland polytene chromosomes by in situ hybridization. Methods Mol Biol Clifton NJ. 2004;260:21–8.
237. Saunders R. In Situ Hybridization to Polytene Chromosomes. Methods Mol Biol Clifton NJ. 2000;123:103–13.
238. Stormo BM, Fox DT. Polyteny: still a giant player in chromosome research. Chromosome Res Int J Mol Supramol Evol Asp Chromosome Biol. 2017;25:201–14.
239. Brown AJL, Moss JE. Transposition of the I element and copia in a natural population of Drosophila melanogaster. Genet Res. 1987;49:121–8.
240. Charlesworth B, Lapid A, Canada D. The distribution of transposable elements within and between chromosomes in a population of Drosophila melanogaster. I. Element frequencies and distribution. Genet Res. 1992;60:103–14.
241. Biémont C, Lemeunier F, Garcia Guerreiro MP, Brookfield JF, Gautier C, Aulard S, et al. Population dynamics of the copia, mdg1, mdg3, gypsy, and P transposable elements in a natural population of Drosophila melanogaster. Genet Res. 1994;63:197–212.
242. Montgomery E, Charlesworth B, Langley CH. A test for the role of natural selection in the stabilization of transposable element copy number in a population of Drosophila melanogaster. Genet Res. 1987;89:435–45.
243. Langley CH, Montgomery E, Hudson R, Kaplan N, Charlesworth B. On the role of unequal exchange in the containment of transposable element copy number. Genet Res. 1988;52:223–35.
244. Finnegan DJ. Transposable elements. Curr Opin Genet Dev. 1992;2:861–7.
245. Nuzhdin SV. Sure facts, speculations, and open questions about the evolution of transposable element copy number. Genetica. 1999;107:129–37.
246. Cridland JM, Macdonald SJ, Long AD, Thornton KR. Abundance and Distribution of Transposable Elements in Two Drosophila QTL Mapping Resources. Mol Biol Evol. 2013;30:2311–27.
247. Deloger M, Cavalli FMG, Lerat E, Biémont C, Sagot M-F, Vieira C. Identification of expressed transposable element insertions in the sequenced genome of Drosophila melanogaster. Gene. 2009;439:55–62.
248. Lipatov M, Lenkov K, Petrov DA, Bergman CM. Paucity of chimeric gene-transposable element transcripts in the Drosophila melanogaster genome. BMC Biol. 2005;3:24.
249. Petrov DA, Fiston-Lavier A-S, Lipatov M, Lenkov K, González J. Population genomics of transposable elements in Drosophila melanogaster. Mol Biol Evol. 2011;28:1633–44.
250. Bartolomé C, Maside X, Charlesworth B. On the Abundance and Distribution of Transposable Elements in the Genome of Drosophila melanogaster. Mol Biol Evol. 2002;19:926–37.
251. Comeron JM, Ratnappan R, Bailin S. The Many Landscapes of Recombination in Drosophila melanogaster. PLoS Genet. 2012;8:e1002905.
252. Charlesworth B, Jarne P, Assimacopoulos S. The distribution of transposable elements within and between chromosomes in a population of Drosophila melanogaster. III. Element abundances in heterochromatin. Genet Res. 1994;64:183–97.
253. Lu J, Clark AG. Population dynamics of PIWI-interacting RNAs (piRNAs) and their targets in Drosophila. Genome Res. 2010;20:212–27.

Publisher's Note

2 The worldwide invasion of *Drosophila suzukii* is accompanied by a large increase of transposable element load and a small number of putatively adaptive insertions

Vincent Mérel[1], Patricia Gibert[1], Inessa Buch[1], Valentina RodriguezRada[1], Arnaud Estoup[2], Mathieu Gautier[2], Marie Fablet[1], Matthieu Boulesteix[1], Cristina Vieira[1].

Co-corresponding authors: Matthieu Boulesteix - matthieu.boulesteix@univ-lyon1.fr & Cristina Vieira - cristina.vieira@univ-lyon1.fr

Affiliations:

1: Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, F-69622 Villeurbanne, France

2: CBGP, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France

Sommaire

2.1 Avant-propos

Ce travail a été réalisé dans le cadre de l'ANR SWING (Worldwide invasion of the **S**potted **WING** Drosophila). Un travail préliminaire sur un assemblage haute qualité du génome de *Drosophila suzukii* (PARIS et collab., 2020), a permis la création d'une banque de données de séquences répétées et une caractérisation du répétome. L'analyse de données de séquençages de 22 populations, 6 de l'aire native et 16 de l'aire envahie, a aidé à l'analyse des facteurs clé de la dynamique des ET ainsi que de leur potentiel rôle adaptatif. Les résultats présentés ici sous la forme d'un article ont été soumis dans le journal Molecular Biology and Evolution https://academic.oup.com/mbe/. Cette version non-définitive n'a pas encore été revue par les pairs. Elle est aussi disponible ici: https://www.biorxiv.org/.../2020.11.06.370932v1.

2.2 Abstract

Transposable Elements (TEs) are ubiquitous and mobile repeated sequences. They are major determinants of host fitness. Here, we portrayed the TE content of the spotted wing fly *Drosophila suzukii*. Using a recently improved genome assembly, we reconstructed TE sequences de novo, and found that TEs occupy 47% of the genome and are mostly located in gene poor regions. The majority of TE insertions segregate at low frequencies, indicating a recent and probably ongoing TE activity. To explore TE dynamics in the context of biological invasions, we studied variation of TE abundance in genomic data from 16 invasive and six native populations (of *D. suzukii*). We found a large increase of the TE load in invasive populations correlated with a reduced Watterson estimate of genetic diversity $\widehat{\theta_W}$, a proxy of effective population size. We did not find any correlation between TE contents and bio-climatic variables, indicating a minor effect of environmentally induced TE activity. A genome-wide association study revealed that ca. 5,000 genomic regions are associated with TE abundance. We did not find, however, any evidence in such regions of an enrichment for genes known to interact with TE activity (e.g. transcription factor encoding genes or genes of the piRNA pathway). Finally, the study of TE insertion frequencies revealed 15 putatively adaptive TE insertions, six of them being likely associated with the recent invasion history of the species.

Keywords: *Drosophila suzukii*, Transposable Elements, Biological Invasion, Populations, Adaptation, PoolSeq.

2.3 Introduction

Transposable Elements (TEs) are selfish genetic elements. Despite being mostly neutral or deleterious, they persist and proliferate in populations by copying and pasting themselves in genomes (CHARLESWORTH et CHARLESWORTH, 1983; DOOLITTLE et SAPIENZA, 1980; ORGEL et CRICK, 1980). The interest for those sequences considerably rose in the 2000's, with the discovery of some TE insertions having a functional, and potentially adaptive, effect on their host (DABORN et collab., 2002; MI et collab., 2000; NIU et collab., 2019). The parallel completion of the first sequencing projects confirmed TE ubiquity and largely contributed to the growing interest for such sequences (C. ELEGANS SEQUENCING CONSORTIUM, 1998; LANDER et collab., 2001; SCHNABLE et collab., 2009).

The nature and intensity of TE deleterious effects may vary with their genomic localization (MÉREL et collab., 2020a). First, TEs close to genes can alter their function. Second, TEs in highly recombining regions, are more likely to promote ectopic recombination, i.e. recombination between more-or-less identical sequences inserted at different locations in the genome. Third, recessive deleterious TEs are more likely to impact fitness when located on a chromosome in a hemizygous state (e.g. the X chromosome in males in a XY sex determination system). The strength of selection acting against TEs hence depends on the genomic region and may result in a local variation of TE density. In agreement with such expectations, TE density was found to be negatively correlated with gene density and local recombination rate in several species (BARTOLOMÉ et collab., 2002; BOISSINOT et collab., 2001). On the other hand, studies focusing on the *D. melanogaster* genome did not reveal a systematic lower TE content on the X-chromosome, which is hemizygous in males (CRIDLAND et collab., 2013; KOFLER et collab., 2012).

TE insertion frequencies reflect both TE activity and the selection acting upon them. Low frequency TE insertions are likely to be recent, or strongly selected against, or both. Conversely, high frequency TE insertions are likely to be old and only weakly subjected to purifying selection. As mentioned previously, TEs that are in the vicinity of genes and/or located in highly recombining regions are expected to be selected against. Accordingly, TE insertion frequencies were found to be negatively correlated with recombination rate and distance to the nearest gene in *D. melanogaster* (KOFLER et collab., 2012). In Drosophila, the overall distribution of TE frequencies seems compatible with an active repeatome (HILL, 2019; KOFLER et collab., 2015b). For example 80% of the insertions have a frequency lower than 0.2 in *D. melanogaster* and its close relative *D. simulans* (KOFLER et collab., 2015b).

Between population variation of TE content has been reported in various intraspecific studies. So far, the factors underlying such differences remain unclear. The effective population size (N_e) may play a prominent role in modulating TE contents. Considering that TEs are mostly deleterious, and that small N_e leads to a less efficient purifying selection, small N_e should be associated with high TE content (LYNCH et CONERY, 2003). In support for this hypothesis LYNCH et CONERY (2003) found a significant correlation between genome size and estimates of the scaled mutation rate $\vartheta = N_e\mu$ (with μ the mutation rate) across populations representative of various species. At the intraspecific level, if a higher TE content in some populations has sometimes been suggested to result from a reduction of their N_e (GARCÍA GUERREIRO et collab., 2008; GARCÍA GUERREIRO et

FONTDEVILA, 2011; TALLA et collab., 2017), to our knowledge the above expected correlation has not been reproduced at this evolutionary scale.

Variation in TE content may also rely on changes in TE activity in relation with the environment (STAPLEY et collab., 2015; VIEIRA et collab., 1999). In Drosophila, several laboratory experiments suggest that TE activity may respond to the environment (GARCÍA GUERREIRO, 2012; HORVÁTH et collab., 2017), but in natura studies considering the whole repeatome remain rare and a possible confounding effect of the demographic history cannot be excluded (LERAT et collab., 2019). Finally, the host genotype may explain intraspecific variation of TE abundance. For instance, in Drosophila, several studies found different levels of activity among isogenic lines (BIÉMONT et collab., 1987; DÍAZ-GONZÁLEZ et collab., 2011; PASYUKOVA et NUZHDIN, 1993).

The study of intraspecific variations in TE content and the underlying determining factors is valuable as TEs may also be important for adaptation (DABORN et collab., 2002; NIU et collab., 2019; VAN'T HOF et collab., 2016). Although some TE insertions exhibit a strong signal of positive selection and have been thoroughly validated experimentally, only few studies aimed at identifying putatively adaptive insertions at a genome-wide level (GONZÁLEZ et collab., 2008; LI et collab., 2018; RECH et collab., 2019; RISHISHWAR et collab., 2018). In addition, most of these studies deal with *D. melanogaster* (BLUMENSTIEL et collab., 2014; GONZÁLEZ et collab., 2010, 2008; RECH et collab., 2019). The most comprehensive of these studies analyzed genomic data on 60 worldwide natural *D. melanogaster* populations and reported 57 to 300 putatively adaptive insertions (depending on the degree of evidence considered) among the ~800 polymorphic insertions identified in the reference genome (RECH et collab., 2019). Considering that approximately twice as many non reference TE insertions as reference insertions may segregate in a single population (KOFLER et collab., 2015b), quite a high number of TE-induced adaptations is therefore expected. However, it remains unclear how important TEs are as substrates of adaptation considering the paucity of studies and their focus on reference genome insertions.

Invasive species provide a unique opportunity to study the combined effect of in natura N_e variations and environmental variations both on TE abundance and TE adaptive potential. Invasive populations often go through demographic bottlenecks allowing to test for an effect of N_e on TE abundance (ESTOUP et collab., 2016). Individuals from invasive populations also encounter new environmental conditions, allowing to test for an effect of bio-climatic variables on TE abundance. Because of the need of colonizing individuals to adapt to new environmental conditions, biological invasions are often used to study rapid contemporary adaptation (LAVERGNE et MOLOFSKY, 2007; ROLLINS et collab., 2015). Yet, the particular role of TEs in the rapid adaptation of invasive species remains speculative. In particular, TEs have been proposed to explain, at least in part, the paradox of invasive species, i.e. the successful adaptation to a new environment despite a reduced genetic diversity caused by small founder population sizes (ESTOUP et collab., 2016; MARIN et collab., 2020; STAPLEY et collab., 2015). In response to environmental changes, TE sequences may be recruited and affect the expression of nearby genes. Furthermore, if a higher activity of TE is induced in response to environmental changes, the insertions could thus result in genetic variation, and potentially beneficial alleles.

In this paper, we focused on the spotted wing fly *D. suzukii*, a close relative of *D. melanogaster*, displaying the highest reported TE content among Drosophila (SESSEGOLO et collab., 2016). *D.*

suzukii is native from Asia and has invaded independently the American and European continents where it was introduced probably in the late 2000's (FRAIMOUT et collab., 2017). Using the recently released high-quality genome assembly Dsuz-WT3_v2.0 based on Long PacBio Reads (PARIS et collab., 2020), we constructed a de novo TE database and found that TE represented 47% of the genome. We further assessed TE insertion frequencies and TE abundance in 22 worldwide populations representative of the native area (n=6) and of the two main invaded areas in Europe (n=8) and America (n=8). The study of TE frequencies showed that the repeatome is highly active in *D. suzukii*: 75% of insertion segregated at a frequency < 0.25. We found that the TE content was significantly higher in invasive populations and was correlated with a reduction of N_e. Finally, controlling for population structure, a genome scan conducted on polymorphic TE insertions identified 15 putatively adaptive TE insertions.

2.4 Results

2.4.1 A highly repeated reference genome

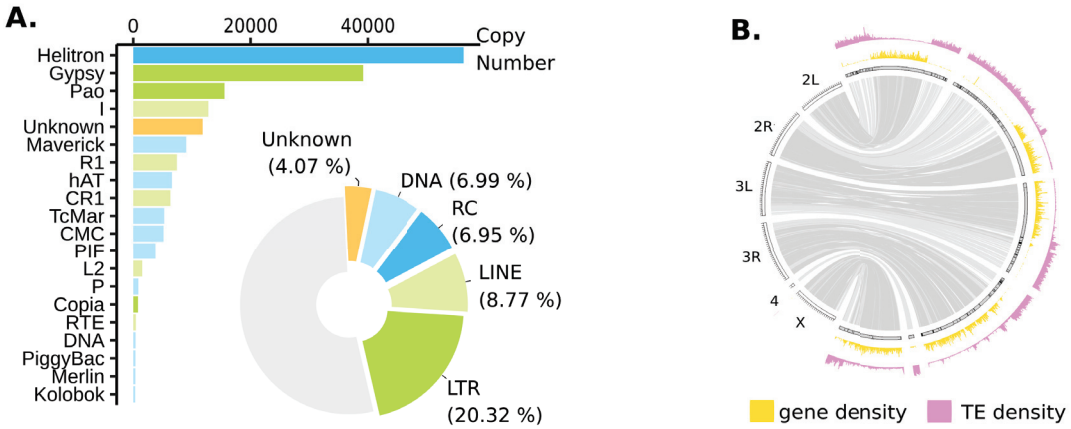


Figure 2.1: **Main features of the TE content in the *D. suzukii* reference genome. A.** TE copy numbers and TE genomic occupancy. Barplot representing TE copy numbers for the 20 TE superfamilies displaying the highest copy numbers. Piechart illustrating genomic sequence occupancy of each TE order (in percentages of the assembly). Class I TEs are shown in green (light green for LINEs and darker green for LTR Elements). Class II TEs are shown in blue (light blue for DNA and darker blue for Rolling Circles (RC)). Non repeated sequences are shown in gray. **B.** Distribution of TEs and genes. TE density (pink outer graph) and gene density (yellow inner graph) are shown for windows of 200 kb. The maximum value of gene density is 54. The maximum number of TE fragments is 713. Syntenic relationships with *D. melanogaster* assembly are shown inside using light links for regions of low gene density (< 7 genes per 200 kb) and dark links for regions of high gene density (≥ 7 genes per 200 kb). Contigs are surrounded by black strokes. Ticks on *D. melanogaster* assembly are separated by one Mb.

We found that the high-quality *D. suzukii* assembly Dsuz-WT3_v2.0 of PARIS et collab. (2020) is characterized by a high TE content. Overall, 47.07% of the reference assembly is annotated as repeated sequences (fig. 2.1A). In terms of genomic occupancy, LTR is the predominant TE order with more than 20% of the sequence assembly corresponding to these elements, then LINEs (8.77%), DNA elements (6.99%), and RC (6.95%). 4.07% of the assembly is occupied by unknown repeated sequences. At a lower hierarchical level, the three most represented superfamilies are Gypsy, Helitron and Pao, corresponding to 13.65%, 6.95% and 6.44% of the assembly, respectively

(supplementary table 2.S1). The average percentage of genomic occupancy per superfamily is 1.88%. Regarding TE copy numbers, the top three superfamilies are Helitron, Gypsy and Pao (56,493, 39,189 and 15,555 copies, respectively) (fig. 2.1A). The average number of copies per superfamily is 4,963. Syntenic relationships with *D. melanogaster* genome have been established for 212 of the 546 contigs of *D. suzukii* assembly. A total of 241 Mb of the 268 Mb assembly have a clearly identified counterpart in the *D. melanogaster* genome (fig. 2.1B, supplementary table 2.S2). Considering the observed bimodal distribution of gene density, we partitioned the *D. suzukii* assembly into gene-rich regions (≥ 7 genes per 200 kb; 121.8 Mb) and gene-poor regions (< 7 genes per 200 kb; 108 Mb) (fig. 2.1B, supplementary fig. 2.S1). TE fragment density also follows a bimodal distribution: 127.4 Mb correspond to TE-rich regions (≥ 165 TE fragments per 200 kb) and 102.4 Mb to TE-poor regions (< 165 TE fragments per 200 kb) (fig. 2.1B, supplementary fig. 2.S2). TE-rich regions are enriched in gene-poor regions, and TE-poor regions are enriched in gene-rich regions ($\chi 2 = 786.47$, df = 1, p-value < 2.2x10-16). We did not find any difference in mean TE density between autosomal and X-linked contigs ($\hat{\mu}_{Autosomes}$=172.00, $\hat{\mu}_{X-linked}$=151.93, W = 78900, p-value = 0.11). This conclusion holds when comparing autosomal and X-linked contigs as defined in Paris et collab. (2020) using a female-to-male read mapping coverage ratio ($\hat{\mu}_{Autosomes}$=176.11, $\hat{\mu}_{X-linked}$=150.09, W = 79088, p-value = 0.38). However, when considering only gene-rich regions, the mean TE density was far higher for X-linked contigs ($\hat{\mu}_{Autosomes}$=65.31, $\hat{\mu}_{X-linked}$=107.54, W=47394, p-value < $2.2x10^{-16}$). Once again, this conclusion holds when using autosomal and X-linked contigs as defined by Paris et collab. (2020) ($\hat{\mu}_{Autosomes}$ = 65.34, $\hat{\mu}_{X-linked}$ = 107.07, W = 47557, p-value < $2.2x10^{-16}$).

2.4.2 An active repeatome in the Watsonville reference population

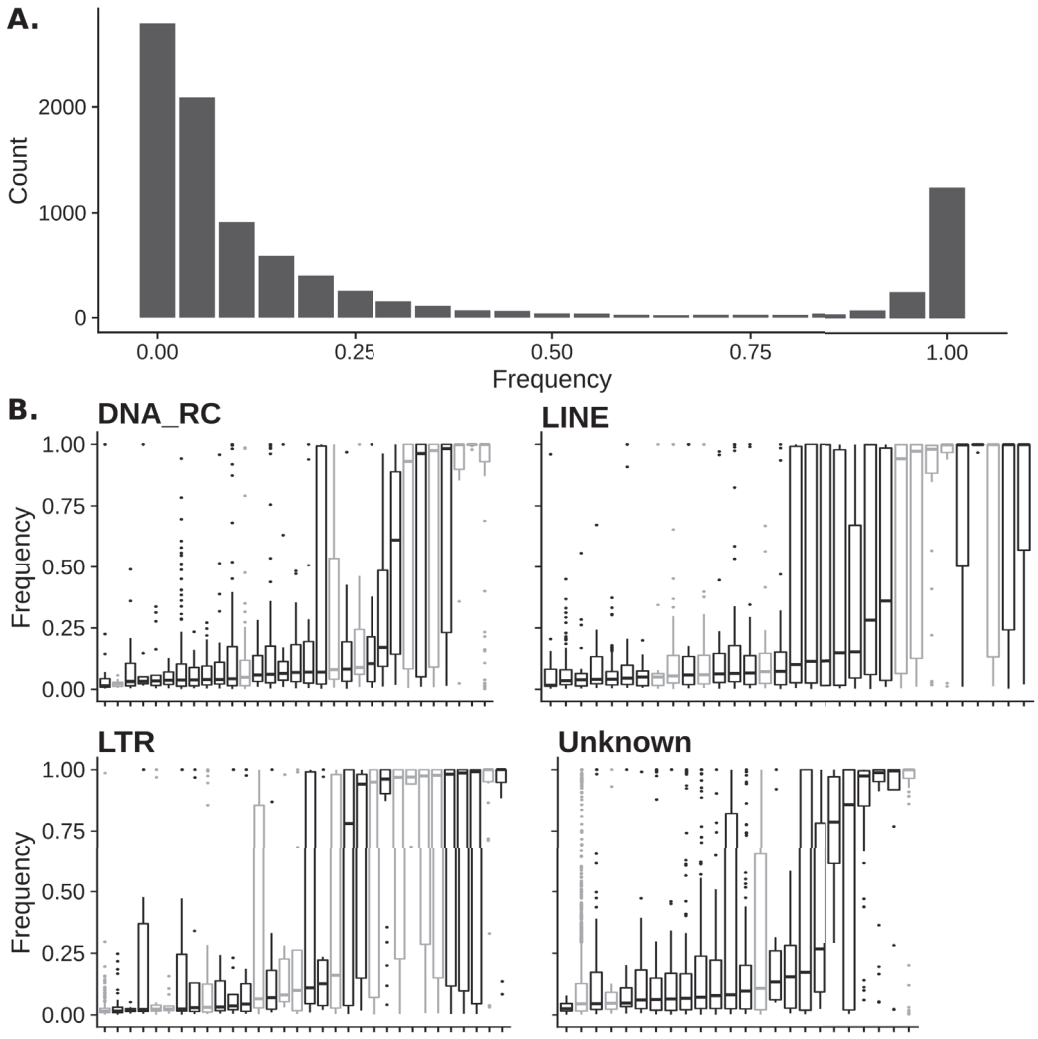


Figure 2.2: **TE activity in the *D. suzukii* reference population from Watsonville (USA). A.** Frequency distributions of TE insertions. **B.** Population frequencies for each TE family (in black) or pseudofamily (in gray). Only families/pseudofamilies with more than 10 insertions in the reference population are shown. DNA and Rolling Circles (RC) have been grouped for graphical reasons.

The female used to establish the WT3 isofemale strain corresponding to the genome assembly was collected in Watsonville (CA, USA) (PARIS et collab., 2020). To thoroughly evaluate TE activity in this reference population, we assessed TE insertion frequencies in a PoolSeq sample of 50 *D. suzukii* individuals from Watsonville. Because TEs are mostly deleterious, rare TE insertions are likely to be recent insertions, not yet eliminated by selection, whereas fixed TE insertions are presumably old insertions weakly submitted to selection. It is worth stressing that, for the study of TE frequencies and abundances, we first used simulated PoolSeq data to validate our pipelines and to evaluate their performance and their sensibility to parameters such as sequencing coverage or number of individuals (see supplementary methods for details).

A total of 9,256 insertions were recovered in the reference population. The frequency distribution is approximately U-shaped (fig. 2.2A) with a majority of insertions segregating at low frequency ($N = 6934$, $f < 0.25$). 1,642 insertions are found at high frequency, in the reference population ($f \geq 0.75$). Only a minority of insertions are of intermediate frequency ($N = 680$, $0.25 \ f < 0.75$). Among the 654 families/pseudofamilies found in the whole dataset, 473 were present in the reference population. 102 belonged to the DNA order, 98 to the LINE order, 175 to the LTR order, 46 to the RC

and 52 were Unknown. Only 119 TE families/pseudofamilies presented more than 10 insertions: 25 DNA families/pseudofamilies, 32 LINEs, 32 LTR, 6 RC and 24 Unknown. The vast majority of these families presented a median frequency lower than 0.25 (N = 80) (fig. 2.2B). Only four families displayed a median frequency between 0.25 and 0.75. Finally, 35 families had a median frequency superior or equal to 0.75. We did not find evidence that the number of TE families in these categories differed between TE orders (supplementary table 2.S3; $\chi 2 = 4.94$, df = 8, p-value = 0.76). However, the mean frequency was slightly different ($\hat{\mu}_{\text{DNA}} = 0.30$, $\hat{\mu}_{\text{LINEs}} = 0.31$, $\hat{\mu}_{\text{LTR}} = 0.46$, $\hat{\mu}_{\text{RC}} = 0.22$, $\hat{\mu}_{\text{Unknown}} = 0.16$, Kruskall-Wallis $\chi 2 = 92.35$, df = 4, p-value < $2.2\text{x}10^{-16}$). TE insertion frequencies were not evenly distributed along the assembly: mean TE insertion frequency was considerably lower in gene-rich windows. ($\hat{\mu}_{rich} = 0.13$, $\hat{\mu}_{poor} = 0.72$, W = 18863, p-value < $2.2\text{x}10^{-16}$; supplementary fig. 2.S3).

2.4.3 Demography as driver of TE contents in *D. suzukii* populations

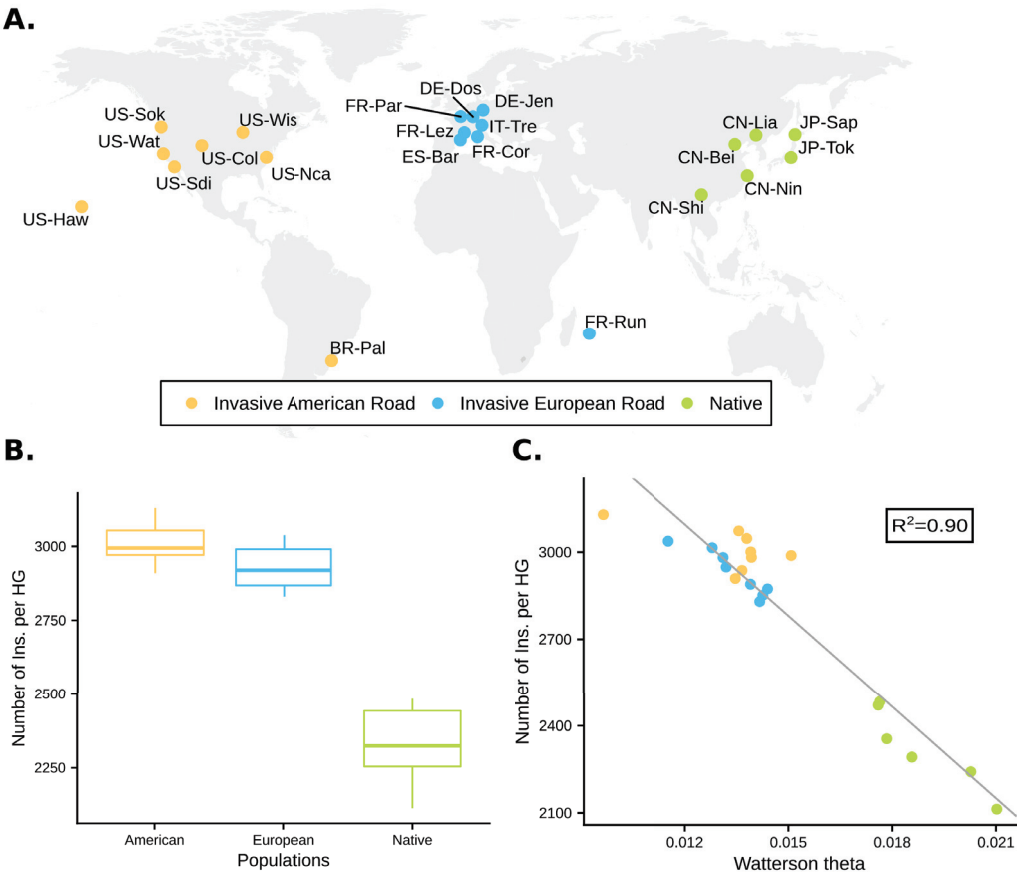


Figure 2.3: **TE dynamics in native and invasive *D. suzukii* populations. A.** Geographic location and historical status of the 22 *D. suzukii* population samples genotyped using a pool-sequencing methodology. Population samples from the native range are in green and those from the invaded range are in orange (American invasion route) or blue (European invasion route) (FRAIMOUT et collab., 2017). **B.** TE content in *D. suzukii* populations, as the numbers of insertions per haploid genome (HG). **C.** Correlation between TE content and Watterson's theta in *D. suzukii* population samples.

Our estimation of TE abundance in the 22 genotyped *D. suzukii* populations (fig. 2.3A) indicates substantial variation across populations, with significantly more TEs in invasive than in native populations and a strong correlation with the Watterson estimate of genetic diversity obtained

from SNPs corresponding to a proxy of population effective size (fig. 2.3B, C). The mean number of insertions per Haploid Genome (HG) and per population was 2,793, ranging from 2,113 in the Chinese population CN-Nin to 3,129 in the Hawaiian population (US-Haw). There was a significant effect of the continent on the mean number of families/pseudofamilies per population: American and European populations had more families/pseudofamilies than native populations ($\hat{\mu}_{America}$ = 470, $\hat{\mu}_{Europe}$ = 468, $\hat{\mu}_{Asia}$ = 453, Kruskal-Wallis $\chi 2$ = 10.505, df = 2, p-value = 0.0052). American and European populations also had more insertions per HG than native populations ($\hat{\mu}_{America}$ = 3008, $\hat{\mu}_{Europe}$ = 2928, $\hat{\mu}_{Asia}$ = 2326, Kruskal-Wallis $\chi 2$ = 14.4, df = 2, p-value = 7.3×10^{-4}). We found a negative linear correlation between the total number of insertions per HG and per population and the Watterson estimate of genetic diversity obtained from SNPs $\widehat{\theta_W}$, a proxy of population effective size (t = -13.415, df = 20, p-value = 1.8×10^{-11}, fig. 2.3C). The variation of explains a large proportion of the variance in the total number of insertions per HG across the populations (R^2 = 0.90). The correlation remains significant when considering only native populations (t = -5.22, df = 4, p-value = 6.4×10^{-3}), or only invasive populations (t = -3.06, df = 14, p-value = 8.6×10^{-3}), or only European populations (t = -5.46, df = 6, p-value = 1.6×10^{-3}), but not when considering only American populations (t = -1.89, df = 6, p-value = 0.11). The correlation between the number of insertions per HG per population and was also assessed individually for the 83 TE families/pseudofamilies showing an amplitude of variation superior or equal to 3 copies per HG. After a Benjamini-Hochberg correction for multiple testing, we found a significant correlation for 63 TE families (p-adjusted < 0.05).

2.4.4 Environmental and genotypic effects on TE abundance

Because $\widehat{\theta_W}$ did not explain all the observed variation in TE abundance among the 22 sampled populations, we tested the effect of two other factors: the environmentally induced changes in TE activity and the genetically determined changes.

To test for an effect of environmentally induced changes in TE activity, we used Partial Mantel tests. We tested the correlation between 19 bioclimatic variables and TE family abundance, for the 83 TE families showing an amplitude of variation superior or equal to 3 copies per HG, correcting for population structure. After correction for multiple testing we did not find any significant correlation (Benjamini-Hochberg correction for multiple testing, p-adjusted < 0.05).

To evaluate the effect of genetic variation on TE abundance we performed a genome-wide scan for association using methods controlling for population structure. To that end we relied on the 13,530,656 bi-allelic variants (mostly SNPs) previously described on the same data set (Olazcuaga et collab., 2020) and searched for association with the population abundance of the 83 TE families/pseudofamilies mentioned above using the BayPass software. Globally, we found 4,856 genomic regions showing evidence of association with population abundance of at least one TE family. Each region spanned at least 1 kb on the reference assembly and included one or several significant SNP/InDel separated by less than 1 kb (significance threshold: Bayes Factor (BF) >20). On average each region was associated with the number of insertions per HG of 1.37 families (min=1, max=69) and contained 2.40 SNPs/InDels (min=1, max=49). 306 (6.30%) regions overlapped with repeated sequences as annotated in the reference genome, which is less than expected by drawing SNP/InDel associated regions randomly ($\hat{\mu}$ = 9.22, $quantile_{0.025}$ = 7.60%, $quantile_{0.975}$ = 11.16%;

supplementary fig. 2.S4A). Only 14 of these regions contain a TE of the same family/pseudofamily as the TE abundance they were associated with. Regarding genes, 2,843 (58.55%) regions were associated with at least one gene, which is less than expected under random expectations ($\hat{\mu}$ = 66.97, $quantile_{0.025}$ = 62.40%, $quantile_{0.975}$ = 70.76%; supplementary fig. 2.S4B). Due to their known role in the activity of TEs, we further searched for enrichment in genes encoding transcription factors and piRNA pathway effectors among the genes located within our candidate regions. We did not observe any significant enrichment in genes encoding transcription factors (Observed: 13.63%, Expected: , $quantile_{0.025}$ = 12.00%, $quantile_{0.975}$ = 20.60%; supplementary fig. 2.S4C) nor in genes involved in the piRNA pathway (Observed: 0.33%, Expected: , $quantile_{0.025}$ = 0.00%, $quantile_{0.975}$=1.18%; supplementary fig. 2.S4D). Among the top 10 regions, corresponding to the regions associated with the highest number of TE families/pseudofamilies, two appeared to be non genic, four could not be attributed to *D. melanogaster* genome, three were associated with the mitochondrial genome and one was associated with *blot*, a member of the sodium- and chloride-dependent neurotransmitter symporter family (https://flybase.org/reports/FBgn0027660).

2.4.5 A small number of putatively adaptive TE insertions

Insertion	Statistics	Highest freq.	f	Taj D	Gene vicinity	Outlier SNP nearby	A/X	Order
1	C_2^{Am} XtX	JP-Sap	0.766	FALSE	*ASPP*	F	A	Unknown
2	C_2^{WW}	US-Haw	0.387	FALSE	*dia*	F	A	Unknown
3	C_2^{WW}	DE-Dos	0.229	FALSE	-	T	A	DNA
4	C_2^{WW}	FR-Cor	0.816	NA	NA	F	X	Unknown
5	C_2^{WW}	US-Wat	0.670	FALSE	*inaE*	F	X	Unknown
6	C_2^{WW}	US-Sok	0.359	FALSE	-	F	X	DNA
7	XtX	US-Sok	0.775	NA	*Mical*	F	A	DNA
8	XtX	JP-Sap	0.908	FALSE	*CG30015*	F	A	Unknown
9	XtX	US-Col	0.804	FALSE	-	F	A	Unknown
10	XtX	US-Sok	0.790	FALSE	*CR31386*	F	A	Unknown
11	XtX	JP-Tok	0.854	FALSE	-	F	A	Unknown
12	XtX	JP-Sap	0.609	FALSE	*Dop1R2*	F	A	Unknown
13	XtX	JP-Sap	0.617	FALSE	*jing*	F	A	Unknown
14	XtX	US-Sok	0.825	FALSE	*CG14282*	F	A	Unknown
15	XtX	JP-Sap	0.859	NA	*GATAe*	F	A	Unknown

Table 2.1: **Description of the 15 putatively adaptive insertions.** Each insertion is an outlier when considering one or a combination of the global differentiation statistics (XtX) and statistics contrasting allelic frequencies between native populations and populations of the invasive American road (C_2^{Am}) or populations of the invasive European road(C_2^{Eu}) or all invasive populations (C_2^{WW}).
Note.—The fourth column indicates whether a SNP potentially evolving under positive selection had been detected less than 5 kb away in Olazcuaga et collab. (2020) (F=False, T=True). The fifth column indicates whether the insertion is located on an autosomal (A) or X-linked contig (X).

We investigated the presence of putatively adaptive insertions using a genome scan combining three methods controlling for population structure implemented in BayPass (Olazcuaga et collab., 2020). First, we assessed overall differentiation (based on the XtX statistics). Second, we studied allele frequencies differences between two groups of populations (based on the C_2 statistics): American invasive vs native populations (C_2^{Am}), European invasive vs native populations (C_2^{Eu}),

all invasive vs native populations ($C_2{}^{WW}$). Third, we carried out genome-wide association with each of the 19 bioclimatic variables (based on the BF).

The genome scan was conducted on 7,004 polymorphic TE insertions (MAF > 0.025, 5,944 autosomal insertions and 1,060 X-linked insertions treated separately). We identified a total of 15 putatively adaptive insertions (12 located on autosomal and three on X-linked contigs) (table 2.1; fig. 2.4). Nine of these insertions were outliers when considering the global differentiation statistics XtX. Note that their frequencies were distinct between native Chinese (low frequencies) and native Japanese populations (high frequencies). One insertion was an outlier for both the XtX and $C_2{}^{Am}$ statistics. Finally, the last five insertions were outliers for the $C_2{}^{WW}$ statistics. No significant association was found between TE insertion frequencies and the 19 bioclimatic variables investigated. One of the 15 putatively adaptive insertions was close (i.e., 399 bp away) to a SNP/InDel that had previously been identified in a region potentially associated with *D. suzukii* invasive success (table 2.1) (OLAZCUAGA et collab., 2020). For one insertion we did not find any homologous regions in *D. melanogaster*, four others were in genomic regions without any genes, and the ten remaining were associated with genes.

We further investigated signatures of selection around candidate insertions by estimating local Tajima's D statistics in the SNP/InDel dataset. Low values of Tajima's D indicate an excess of rare mutations, one possible signature of a selective sweep due to positive selection. To test if each of our candidate insertions were associated with selective sweeps, we computed the linear correlation between its frequency and local Tajima's D values (supplementary fig. 2.S5). Five statistically significant correlations were found corresponding to the insertions n°4, 9, 10, 12 and 15 (Pearson's product-moment correlation, $p < 0.05$). Only a single insertion was associated with an extreme local Tajima's D (insertion n°15; Tajima's D < $\text{quantile}_{0.05}$), and only for a single population. The visualization of Tajima's D at a larger scale (i.e., 10 kb upstream - 10 kb downstream the insertion) confirms the lack of strong effect of the investigated insertions on Tajima's D (supplementary fig. 2.S6). It is worth noting that, if the effect of our candidate TE insertion on Tajima's D is globally low, a close investigation of Tajima's D suggests that, at least in some cases, it is the absence rather than the presence of the insertion that may be adaptive. As a matter of fact, while the correlation implying an extreme local Tajima's D was negative, the four other significant correlations between local Tajima's D and insertion frequency were positive.

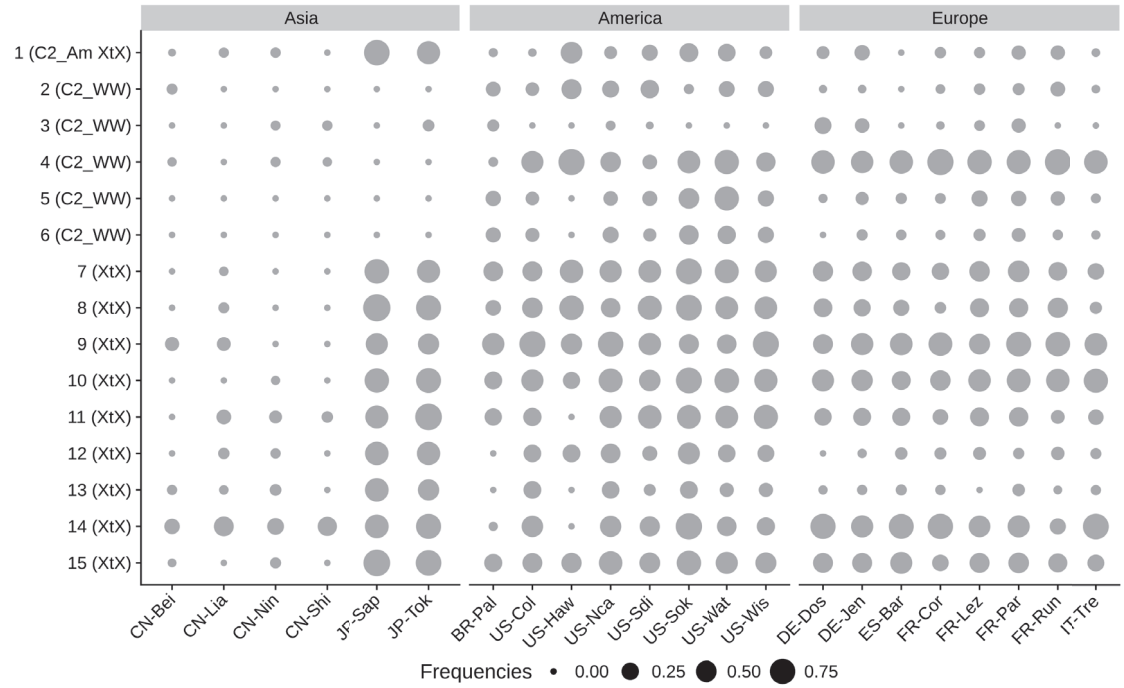


Figure 2.4: **Frequencies of each of the 15 putatively adaptive insertions in the 22 *D. suzukii* populations.** Insertion number is indicated on the left together with the associated BayPass statistics. XtX corresponds to a global differentiation statistic, C_2 to a statistic contrasting allelic frequencies between native populations and populations of the invasive American road ($C_2{}^{Am}$), or populations of the invasive European road ($C_2{}^{Eu}$), or all invasive populations ($C_2{}^{WW}$).

2.5 Discussion

For most species the repeatome is still a poorly known genomic compartment and much remains to be understood regarding its variability, dynamics, functional and fitness impacts. This is all the more important given that TEs appear to be ubiquitous, prompt to invade new genomes (Kofler et collab., 2015a), and they may drastically impact the host phenotype (Daborn et collab., 2002; Nikitin et Woodruff, 1995; Van't Hof et collab., 2016). Here we capitalized on a recently generated long-reads genome assembly and a large set of populational PoolSeq data (Olazcuaga et collab., 2020; Paris et collab., 2020) to thoroughly portray the TE content of the non model invasive species *D. suzukii*.

2.5.1 An abundant, unevenly distributed and active repeatome

The observed 47% of TEs in the genome of *D. suzukii* confirmed the outlier position of this species within the Drosophila genus regarding the global amount of TEs. Our estimate is somewhat higher than those reported in previous studies in *D. suzukii* (Chiu et collab., 2013; Ometto et collab., 2013; Sessegolo et collab., 2016). Considering that the assembly of repeats is often impossible using short paired-end (PE) reads (Rius et collab., 2016), it is not surprising that we recovered more TEs in a long reads genomic assembly than previous studies investigating TE contents using PE reads assemblies (Chiu et collab., 2013; Ometto et collab., 2013). In addition, we here performed a de novo reconstruction of TE sequences, which allowed us to identify more TE families/pseudofamilies, as compared to the previous research work based on the same assembly (35%) (Paris et collab.,

2020). Overall, de novo reconstruction of TE sequences from long read assemblies, such as the 15 Drosophila species assemblies recently generated using nanopore sequencing (MILLER et collab., 2018), should greatly improve our knowledge of TE diversity in Drosophila.

In agreement with the gene disruption hypothesis and observations in a variety of species (BARTOLOMÉ et collab., 2002; MEDSTRAND et collab., 2002; WRIGHT et collab., 2003), we observed a depletion of TE copies in gene-rich regions of the *D. suzukii* genome. Although it is likely that TEs are strongly selected against in these regions due to their negative effect on gene function or expression (LEE et KARPEN, 2017), it is also possible that TE copies are depleted in these regions because they promote ectopic recombination. In agreement with the latter hypothesis, gene-rich regions are also known to display high recombination rate in *D. melanogaster* (ADAMS et collab., 2000). The generation of a genomic map of recombination rates in *D. suzukii* would be needed to disentangle the respective effects of ectopic recombination and gene disruption.

At the chromosomal scale, we did not find a lower density of TEs on the X chromosome compared to autosomes. This pattern indicates that, if X-linked recessive insertions are more efficiently selected against than autosomal insertions, the effect on TE abundance is either low or balanced by another process. When comparing only gene-rich regions, we even found a higher density of TEs on the X chromosome than on autosomes. Three non-mutually exclusive explanations can be invoked: (i) there may be a higher insertion rate on the X chromosome, similar to what was previously found in *D. melanogaster* (ADAMS et collab., 2000); (ii) the recombination rate may be lower on the X chromosome, and thus a stronger Muller's ratchet; and (iii) the strength of selection may be reduced by a smaller effective population size for the X chromosome.

Similarly to what has been found in *D. melanogaster* and *D. simulans* (KOFLER et collab., 2015b) and to what is probably common among Drosophila species (HILL, 2019), the pattern of TE insertion frequencies in *D. suzukii* is compatible with an active repeatome. We found differences in the mean insertion frequency between TE orders, which suggests differences in activity but could also result from variation in the strength of purifying selection acting against the different orders (LEE et KARPEN, 2017; PETROV et collab., 2003). Considering the trap model of TE dynamics (i.e. a model in which newly invading TEs are quickly inactivated by host defense (KOFLER et collab., 2015b; ZANNI et collab., 2013)), an active repeatome suggests a recurrent turnover of TEs, potentially due to horizontal transfer events. Investigating TE activity in *D. melanogaster* and *D. simulans*, Kofler and colleagues KOFLER et collab. (2015b) suggested that such a turnover is influenced by the colonization history of those species. They propose that the high activity of DNA transposons in *D. simulans* results from horizontal transfer events from *D. melanogaster* during *D. simulans* worldwide colonization. In agreement, we detected more families/pseudofamilies in invasive populations of *D. suzukii* than in the native ones, suggesting that new TE families may have been acquired during the recent colonization of new areas. However, because the TE database used here relies on a reference genome obtained from individuals originating from America (i.e. from the Watsonville population), one may expect to find much more families/pseudofamilies in American than European populations. Yet this is not what we observed. This could be due to admixture between American and European populations. However, population genetics studies have shown that gene flow between the two continents is limited if not absent (FRAIMOUT et collab., 2017). It is thus possible that, for technical reasons, we are simply missing some families that are less

abundant in the Asian native range of the species. The comparison of long read assemblies of genomes generated from individuals originating from the three continents (Asia, America and Europe) should help shedding light on this issue.

2.5.2 Demography, rather than environment or genotype, drives TE content

In agreement with the Lynch and Connery hypothesis (LYNCH et CONERY, 2003), we found that the TE content in *D. suzukii* is negatively correlated with the Watterson estimate of genetic diversity which may be viewed as a proxy of the population effective size N_e. The negative correlation between and TE content was significant when considering only European invasive populations, invasive populations as a whole, or only native populations, but was not significant when considering only American invasive populations. Although a few studies suggest an increase of TE content following colonization (GARCÍA GUERREIRO et collab., 2008; GARCÍA GUERREIRO et FONTDEVILA, 2011; NARDON et collab., 2005; TALLA et collab., 2017), to our knowledge it is the first time that a correlation between TE content and N_e is found at the intraspecific level. Although several factors may affect N_e, the variation observed is likely to result from demographic processes. Indeed, both European and American invasive populations have encountered bottlenecks (FRAIMOUT et collab., 2017). In agreement with this idea, the invasive population from Hawaii, which experienced the strongest bottleneck (FRAIMOUT et collab., 2017), showed the smallest values. It is interesting to note that the negative correlation between and TE content remains significant when considering only native populations suggesting that other demographic event than bottleneck may also be involved (e.g. different stable effective population sizes and gene flow patterns).

Our analysis is controlled for sequencing bias, i.e. coverage and insert size, and we are confident in the biological significance of the correlation observed here. However, it is worth stressing that our dataset of TE insertions corresponds to a small fraction of the repeatome. Indeed, the mean number of insertions per HG per population is markedly below the number of TE copies recovered in the reference genome. We believe that this is due to an impossibility to properly call TE insertions when TEs are too close or even nested (VENDRELL-MIR et collab., 2019). It is thus possible that the negative correlation that we found here exists only for some part of the genome. Especially it is likely that regions of low TE density, where most of TE insertions are polymorphic, display the strongest answer to a reduction of selection efficacy. This is simply because polymorphic insertions can increase in frequency while fixed insertions cannot. One could also argue that the efficiency of selection is a function of the product between N_e and s (with s the selection coefficient). Therefore, the effects of a reduction of Ne should be especially marked in regions where selection against TEs is strong, such as TE-poor / gene-rich regions.

We found no significant effect on TE abundance for all the 19 environment variables tested. This might be surprising at first sight given the large number of studies showing an association between TE activity and external factors, such as temperature or viral infection (GARCÍA GUERREIRO, 2012; HORVÁTH et collab., 2017; ROY et collab., 2020; RYAN et collab., 2016) Several factors may explain this discrepancy. First, it is important to notice that, in Drosophila, most of these studies rely on lab experiments, some of them exploring environmental conditions unlikely in natura (see (GARCÍA GUERREIRO, 2012) for a review). To our knowledge none of these studies established a link between TE activity and natural environment without any possible confounding effect

from population structure and demographic features. Second, as often in Drosophila, most of such research works were carried out on the same particular species, *D. melanogaster*, so that so far we do not know much about interspecific variability. Third, although partial Mantel tests allowed revealing 15 significant correlations between TE abundance and environmental variables in *A. thaliana* populations (QUADRANA et collab., 2016), we consider our results as conservative, especially regarding the long discussion about the statistical performance of partial Mantel tests (DINIZ-FILHO et collab., 2013). More sophisticated statistical methods may be needed to tackle such relationships into more details.

Considering that several studies on Drosophila suggest a genotype effect on TE activity (ADRION et collab., 2017; BIÉMONT et collab., 1987; DÍAZ-GONZÁLEZ et collab., 2011; PASYUKOVA et NUZHDIN, 1993), we performed a GWAS on TE abundance to assess this effect in natural populations and identify the genomic regions involved. Overall, we found ca. 5,000 genomic regions associated with TE abundance. These regions were not enriched in transcription factor genes nor genes of the piRNA pathway. As far as we know, no such GWAS study has been carried out in Drosophila populations. Our results are somewhat similar to those found in A. thaliana, in which although a strong causal link between one transcription factor and the abundance of two TE families was found, no enrichment for any particular function was observed (QUADRANA et collab., 2016). Comparative genomics between closely related species may help identify a general pattern. Especially, one could lead the same study using available *D. melanogaster* PoolSeq data (KAPUN et collab., 2018), and focus on genes identified in both *D. melanogaster* and *D. suzukii*, as they might be likely to play a key role in the modulation of TE activity.

2.5.3 A potential adaptive role for a limited number of TEs

Similar to studies investigating TE adaptive potential in *D. melanogaster* populations (GONZÁLEZ et collab., 2010, 2008; RECH et collab., 2019), we found several putatively adaptive TE insertions in our *D. suzukii* dataset. Overall, we found 15 insertions, six of which likely to have eased the worldwide invasion of *D. suzukii*. It is important to note that we are probably missing some insertions, and thus likely underestimating the number of adaptive insertions sites.

Overall, we did not capture a strong signal of a selective sweep near the candidate adaptive TE insertions. This may be due to overall large effective population sizes as suggested in (OLAZCUAGA et collab., 2020), but also to the fact that Tajima's D is unlikely to detect soft selective sweep, i.e. adaptation from standing variation or multiple successive beneficial mutations (PENNINGS et HERMISSON, 2006). An appealing perspective would be to sequence candidate regions in individual strains and use a haplotype-based analysis. For example, the recently introduced Comparative Haplotype Identity (xMD) statistics (LANGE et POOL, 2016; VILLANUEVA-CAÑAS et collab., 2017) has been shown to perform well for soft sweeps. If the effect of our candidate TE insertion on Tajima's D is globally low, it highlighted the possibility that the absence rather than the presence of the insertion may be adaptive, at least for some of our candidate insertions. More specifically, for four insertions a positive correlation was found between local Tajima's D and insertion frequency. However, the only extreme local Tajima's D was found in the population where the putatively adaptive insertion is at its highest frequency, indicating that it is probably the insertion itself rather than the absence that might be adaptive.

One added value to our analysis based on GWAS is that the same type of analysis has been carried out using SNPs/InDel (OLAZCUAGA et collab., 2020). The authors of this study found 204 markers strongly associated with invasion success distributed over the whole genome. If we compare this number to our six TE insertions, it seems unlikely that TEs solely may explain the genetic paradox of invasive species (STAPLEY et collab., 2015). It is worth noting that the level of variation remains high in invasive *D. suzukii* populations (FRAIMOUT et collab., 2017). Hence, it would be interesting to carry out similar analyses in invasive species that experienced a more intense depletion of genetic variation during invasion (PRENTIS et collab., 2009; ROUX et collab., 2011; ZHANG et collab., 2010) to assess whether TEs are more likely to be adaptive in invasive populations with low levels of genetic diversity.

At first sight our finding of 15 putatively adaptive polymorphic insertions in worldwide populations of *D. suzukii* contrasts with the 41 to 300 putatively adaptive polymorphic insertions found in worldwide populations of *D. melanogaster* (RECH et collab., 2019). The difference is even more blatant considering that we analyzed 7,004 polymorphic insertions, against 800 in (RECH et collab., 2019). This suggests a largely higher rate of TE induced adaptations during *D. melanogaster* invasion and this despite the much larger, still active and diverse repeatome of *D. suzukii*. This discrepancy could have several non-exclusive explanations. First, it may be due to historical differences between the two species. *D. melanogaster* experienced a relatively slow and ancient worldwide invasion that started from Africa about 15,000 ya, whereas *D. suzukii* came out from its native range in Asia only a few decades ago (FRAIMOUT et collab., 2017; STEPHAN et LI, 2007). Second, the discrepancy may result from intrinsic species differences with respect to the repeatome contents. For example, *D. melanogaster* TEs could possess more environment responsive sequences that might be co-opted by the host. Third, it may be due to differences in the methodology used for the two species. Our analysis relies essentially on the research of overly differentiated TEs across populations with a correction for population structure (GAUTIER, 2015; OLAZCUAGA et collab., 2020), whereas in the analysis used for *D. melanogaster* there is no direct methodological control for population structure. In the *D. melanogaster* study (RECH et collab., 2019), a TE insertion is considered as putatively adaptive if it is present at high population frequency (from 10% to 95%), and is located in genomic regions where recombination rate -and so selection efficacy - is high (ca. 300 putatively adaptive insertions). Further evidence is collected using a combination of three haplotype-based tests to detect selective sweeps in the vicinity of candidates, and statistical treatments based on F_ST estimations (with 84 insertions confirmed by at least one test). Applying our statistical methodologies to the *D. melanogaster* dataset, which also consist in PoolSeq data, would help to determine if methodology differences can explain the observed discrepancy. Finally, one could ultimately rely on experimental evolution, applying the same selective pressure to different Drosophila species, to test for an impact of intrinsic species differences on TE adaptive potential.

Our study of TE induced adaptation strongly calls for a validation of candidate insertions. Allele specific expression assays would allow evaluating if these insertions affect nearby gene expression (GONZALEZ et collab., 2009). This would consist in testing a difference of nearby gene expression between the two alleles of an F1 hybrid between strains with and without the insertion. While such test should control for genotype effect, as compared to a simple test of differential expression between strains, it does not preclude for an effect of a SNP/InDel close to the insertion. Using a CRISPR-Cas9 methodology would also allow (in)validate that the TE(s) of interest is the causative

agent of gene expression change and would allow direct testing for a phenotypic effect.

2.5.4 Conclusion

Our study illustrates the value of an approach combining a long reads based genome assembly, a de novo reconstruction of TE sequences, and PoolSeq population data, to characterize the repeatome of a non model species. Our set of analyses especially highlighted that the particularly large *D. suzukii* repeatome is probably active and shaped by purifying selection, similar to that of *D. melanogaster*'s. Additional data, such as local recombination rate, would also help us shed light on the nature of selection acting on TEs. The analysis of TE abundance variations in invasive and native populations suggests that a reduction of purifying selection intensity, in response to demographic processes, can significantly increase TE content. Our study also indicates that positive selection may act on TE insertions in response to selective factors that remains to be determined. Experimental validation will allow to (in)validate a functional impact of our putatively adaptive insertions. Overall, the natural extent of the trends we uncovered here should be explored into more details, for instance through the application of similar methods to other (invasive) species that would allow to evaluate the impact of a stronger bottleneck on both TE content increase and TE adaptive potential.

2.6 Materials & Methods

Creation of a TE database

A TE database was created by merging previously established consensus of Drosophila TE families and de novo reconstructed consensus of *D. suzukii* TE families. The previously established consensus were obtained by extracting all Drosophila consensus annotated as DNA, LINE, LTR, Other, RC, SINE and Unknown from Dfam and Repbase databases (release 2016-2018 for both) (HUBLEY et collab., 2016)(https://www.girinst.org/repbase/). Full LTR element sequences were reconstructed by merging LTRs and their internal parts. De novo reconstruction was performed using an assembly of an American strain from Watsonville, sequenced using PacBio long reads technology, and the REPET package (v2.5) (FLUTRE et collab., 2011; PARIS et collab., 2020). Unless otherwise specified, the options were used as in the default configuration file. Briefly, the genome assembly was cut into batches and aligned to itself using blastn (ncbi-blast v2.2.6) (ALTSCHUL et collab., 1990). High-scoring Segment Pairs (HSPs) were clustered using Recon (v1.08) and Piler (v1.0) (BAO et EDDY, 2002; EDGAR et MYERS, 2005). A structural detection step was performed using LTRHarvest from the GenomeTools package (v1.5.8) (ELLINGHAUS et collab., 2008; GREMME et collab., 2013). LTRHarvest-produced sequences were clustered using blastclust. Consensus sequences were created for each cluster using MAP (HUANG, 1994). Additional consensus sequences were generated using RepeatScout (v1.0.5) (PRICE et collab., 2005). All consensus, i.e. from Recon, Piler, LTRHarvest and RepeatScout, were further submitted to a filtering step. Sequences were retained only if they produced at least 3 hits against the genome assembly with at least 98% query coverage (blastn, blast 2.6.0+). Structural and coding features were identified and used to classify consensus (see HOEDE et collab. (2014) for classification details, the used libraries were ProfilesBankForREPET_Pfam27.0_-

GypsyDB.hmm, repbase20.05_aaSeq_cleaned_TE.fsa, repbase20.05_ntSeq_cleaned_TE.fsa). Single satellite repeats, potential host genes and unclassified sequences were filtered out. Since REPET can easily mis-annotate any pair of repeats separated by a spacer as TRIM or LARD, those sequences were also removed (ARKHIPOVA, 2017). Remaining sequences were further annotated by homology to previously established consensus of Drosophila TE families. Homology was determined using RepeatMasker (-cutoff 250, v 1.332) (http://www.repeatmasker.org/). We followed the rules below: 1) if all hits belonged to the same superfamily, the sequence was annotated as corresponding to that particular superfamily and order; 2) if hits from different superfamilies were observed the sequence was considered as ambiguous; 3) without any hit, the sequence was annotated as unknown. Ambiguous sequences were manually curated, sequences which could be unambiguously attributed to one superfamily according to hits and proteic domains were kept (proteic domains were investigated using NCBI Conserved Domain Search (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Finally, consensus were clustered in families using UClust (-id 0.80, -strand both, –maxaccepts 0 –maxrejects 0; v11.0.667) (EDGAR, 2010). The annotation, superfamily and order, attributed to each cluster, i.e. each family, is the annotation of the longest sequence in the cluster. The generated TE database is accessible at: https://github.com/vmerel/Dsu-TE.

Annotation of the reference genome

To recover TE fragments and TE genomic sequence occupancy, the reference genome assembly was masked using RepeatMasker and the above TE database (-gccalc, -s, -a, -cutoff 200, -no_is, -nolow, -norna, -u; v 1.332) (http://www.repeatmasker.org/). TE density was evaluated as the number of TE fragments completely within non overlapping genomic windows of 200 kb. TE copies were reconstructed from TE fragments using OneCodeToFindThemAll (BAILLY-BECHET et collab., 2014). Gene density was computed from a run of augustus (–species=fly, –strand=both, –genemodel=complete; v2.5.5) (STANKE et collab., 2008) as the number of genes completely within non overlapping genomic windows of 200 kb. Promer was used to generate alignments between *D. melanogaster* and *D. suzukii* assemblies and establish syntenic relationships (MUMmer v3.23) (KURTZ et collab., 2004). *D. melanogaster* masked assembly was downloaded from UCSC Genome Browser (dm6; http://hgdownload.soe.ucsc.edu/goldenPath/dm6/bigZips/). *D. suzukii* masked assembly was retrieved from RepeatMasker output (see above). The promer output was filtered out using the delta-filter module in order to obtain a one-to-one mapping of reference to query (-q, -r). A file containing alignment coordinates for alignments of minimum length 100 bp, and in which overlapping alignments were merged, was generated with the show-coords module (-b, -L 100, -r). Because the abundance of repeated sequences and the use of masked assemblies may result in multiple small alignments, alignments separated by less than 20 kb were merged using a custom script. Note that only alignments implying the 2L, 2R, 3L, 3R, X and 4 chromosomes of *D. melanogaster* were kept at this step and if a *D. suzukii* contig aligned to several *D. melanogaster* chromosomes only the best pair was conserved (i.e. the pair producing the longest alignment). A graphical visualization of the results was produced using Circos (KRZYWINSKI et collab., 2009).

Fly samples and pool sequencing

Pool-sequencing (PoolSeq) data originate from OLAZCUAGA et collab. (2020) where the detailed associated protocol is described. Briefly, adult wild flies were sampled between 2013 and 2016 from 22 localities of both native and invasive areas (fig. 3A) (FRAIMOUT et collab., 2017). Six samples were collected in the native Asian area, more precisely in four Chinese and two Japanese localities. The remaining 16 samples were chosen to be representative of two separate invasion roads: the American invasion road and the European invasion road. The American invasion road is represented by one Hawaiian sample, one Brazilian sample and six samples from the United States. The European invasion road corresponds to two German samples, four French samples (including one from La Réunion Island), one Italian sample and one Spanish sample. For each population sample, DNA extraction was performed from the thoraxes of 50 to 100 flies and used to prepare paired-end (PE) libraries (insert size of 550 bp). PE sequencing was achieved using a HiSeq 2500 from Illumina to obtain 2×125 bp reads. Reads were trimmed using the trim-fastq.pl script in the PoPoolation package (–min-length 75, –quality-threshold 20; v1.2.2) (KOFLER et collab., 2011).

TE frequency pipeline

To obtain TE insertion frequencies in PoolSeq samples a calling of TEs was done using PoPoolationTE2 (KOFLER et collab., 2016), the reference genome and the newly constructed database. To make sure that no reads from TE sequences could map on the masked assembly, TE reads were simulated, mapped on the masked assembly and aligned positions were also masked. Reads simulation was performed using the script create-reads-for-te-sequences.py (KOFLER et collab., 2016): reads of 125 bp reads, coverage of 1024 X per TE sequence in the database. Because we do not expect a split read based TE calling tool such as PoPoolationTE2 to accurately call for insertions shorter than the insert size, TE sequences shorter than 500 bp were removed before calling. Moreover, as PoPoolationTE2 filters out insertions with reads mapping on more than one family, families with cross-mapping were grouped in pseudofamilies. Two families were brought together if at least 1% of reads from one sequence of the first family were mapped on a sequence of the second family (read simulation: 125 bp reads, coverage of 100 X per consensus). Concerning the TE calling, reads were mapped using bwa bwasw (v0.7.17) (LI et DURBIN, 2010) and paired-end information restored using the se2pe script provided with the PoPoolationTE2 package (v1.10.04) (KOFLER et collab., 2016). One unique ppileup file was generated with all samples specifying a minimum mapping quality of 15. The remaining modules of PoPoolationTE2 were used as follow: identifySignatures: –mode joint, –signature-window minimumSampleMedian, –min-valley minimumSampleMedian, –min-count 2; updatestrand: –map-qual 15, –max-disagreement 0.5; frequency; filterSignatures –min-coverage 10, –max-otherte-count 2, –max-structvar-count 2; pairupSignatures –min-distance -200, –max-distance 300. The final output contained frequencies in the 22 populations for each called TE insertion. See supplementary methods for the validation work on simulated data.

TE abundance pipeline

TE abundances, as the numbers of insertions per HG per population, were estimated in PoolSeq samples by summing insertion frequencies in each sample. Since this pipeline also relies on the estimation of TE frequencies in PoolSeq samples, it is very similar to the TE frequency pipeline. However, the last steps were modified to account for differences in coverage and insert sizes between samples and to allow an unbiased comparison of TE abundance across samples. After the ppileup step the following analyses were performed: subsamplePpileup: –target-coverage 30; identifySignatures –mode separate, –signature-window minimumSampleMedian, –min-valley minimumSampleMedian, –min-count 2; updatestrand: –map-qual 15, –max-disagreement 0.5; frequency; filterSignatures: –min-coverage 10; –max-otherte-count 2; –max-structvar-count 2; pairupSignatures: –min-distance -200; –max-distance 300. See supplementary methods for the validation work on simulated data.

Evaluation of population genetics statistics

We estimated Watterson's theta ($\widehat{\theta_W}$) and Tajima's D statistics in non-overlapping 1000 bp windows using PoPoolation (v1.2.2) (KOFLER et collab., 2011). Forward and Reverse trimmed reads were mapped separately using bwa aln (-o 2 -d 12 -e 12 -n 0.01; v0.7.17) (Li and Durbin 2010). A paired-end alignment file was generated using bwa sampe. Reads were filtered for a minimum mapping quality of 20 and a pileup file generated with samtools (v1.7) (LI et collab., 2009). Each pileup file was split into two files: one corresponding to autosomal contigs and another corresponding to X-linked contigs (autosomal and X-linked contigs as determined in OLAZCUAGA et collab. (2020). PoPoolation was used as follows: –min-count 2 –min-coverage 8 –max-coverage 250 –min-qual 20. The pool-size argument was modified accordingly between autosomal and X-linked pileup.

Genome Wide Association Study with TE family abundance

All genome scans were performed using BayPass (v2.2) (GAUTIER, 2015; OLAZCUAGA et collab., 2020), a package aiming at identifying markers evolving under selection and/or associated to population-specific covariates, taking into account the shared history of the populations. For each SNP/InDel previously called in these PoolSeq samples (OLAZCUAGA et collab., 2020), we estimated 83 Bayes Factors (BF), reflecting their association with the number of insertions per HG of 83 families/pseudofamilies (based on a linear regression model). The 83 chosen TE families/pseudofamilies were those displaying an amplitude of variation of at least three insertions per HG across the complete dataset. To improve computing time BayPass was run on data subsets. Data concerning TE abundance was split into three subsets of 28, 28 and 27 families, respectively. For SNPs/InDel, we used the data subsets of OLAZCUAGA et collab. (2020), for which the 11,564,472 autosomal variants are divided into 154 subsets and the 1,966,184 X-linked variants into 26 subsets. Since we used the importance sampling algorithm implemented in Baypass to assess BFs, and single run estimations may be unstable, a total of three runs were performed for each combination of TE subsets-SNP/InDel subsets and the median of BFs computed (GAUTIER et collab., 2018). Note

that different pool size files were used for autosomal and X-linked variants to take into account differences in the number of autosomes and X chromosomes in each PoolSeq sample. In accordance to Jeffrey's rule, a SNP/InDel was considered as associated with a TE family/pseudofamily abundance for a BF superior to 20 deciban (dB) (JEFFREYS, 1961).

SNP/InDel locations were used to define genomic regions associated with TE abundance. Variants were gathered if separated by less than 1 kb. If the spanned genomic interval was less than 1 kb or if a variant could not be found, the region was obtained by adding 500 bp on both sides. For each region we looked for overlapping TEs using the RepeatMasker annotation (gff file, see Annotation of the reference genome). We also investigated gene content. First, we retrieved homologous regions in the *D. melanogaster* genome using BLAT against the *D. melanogaster* masked assembly downloaded from UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/goldenPath/dm6/bigZips/; BLAT v.36x4, -t=dnax -q=dnax). We then checked for genes overlapping the best hit subject sequence using the UCSC Genome Browser gff annotation file. Note that if the best hit score was lower than 100 we considered that no homologous region was retrieved. The number of transcription factor genes among the genes retrieved was obtained by comparing their IDs to those of the gene group Transcription factor on flybase (https://flybase.org/reports/FBgg0000745.html). Similarly, the number of genes involved in the piRNA pathway was obtained by comparing gene IDs to those listed in OZATA et collab. (2019). In order to test if the candidate regions were enriched in TEs we generated random expectations by applying the above to 1000 randomly selected SNPs 250 times. For computing time reasons, for genes, transcription factor genes, or genes involved in the piRNA pathway, we used 500 randomly selected SNPs 125 times.

Correlation between climatic variables and TE family abundance

Partial Mantel tests were used to test the correlation between bioclimatic variables and TE family abundance correcting for population structure (as in QUADRANA et collab. (2016)). 19 bioclimatic variables from the worldclim dataset (FICK et HIJMANS, 2017) were considered: annual mean temperature, mean diurnal range, isothermality, temperature seasonality, max temperature of warmest month, minimum temperature of coldest month, temperature annual range, mean temperature of wettest quarter, mean temperature of driest quarter, mean temperature of warmest quarter, mean temperature of coldest quarter, annual precipitation, precipitation of wettest month, precipitation of driest month, precipitation seasonality, precipitation of wettest quarter, precipitation of driest quarter, precipitation of warmest quarter, precipitation of coldest quarter. The 83 families with an amplitude of variation of at least three insertions per HG between populations were considered. The population structuring of genetic diversity is summarized by the scaled covariance matrix of population allele frequencies (Ω) estimated with Baypass, one autosomal subset randomly chosen was used (the correlation of the posterior means of the estimated Ω elements across SNP subsamples had previously been verified (OLAZCUAGA et collab., 2020)). Partial Mantel tests were conducted using the R package ecodist (GOSLEE et URBAN, 2007). P-values were further adjusted to account for multiple testing applying the Benjamini-Hochberg correction (BENJAMINI et HOCHBERG, 1995).

Screening for putatively adaptive TE insertions

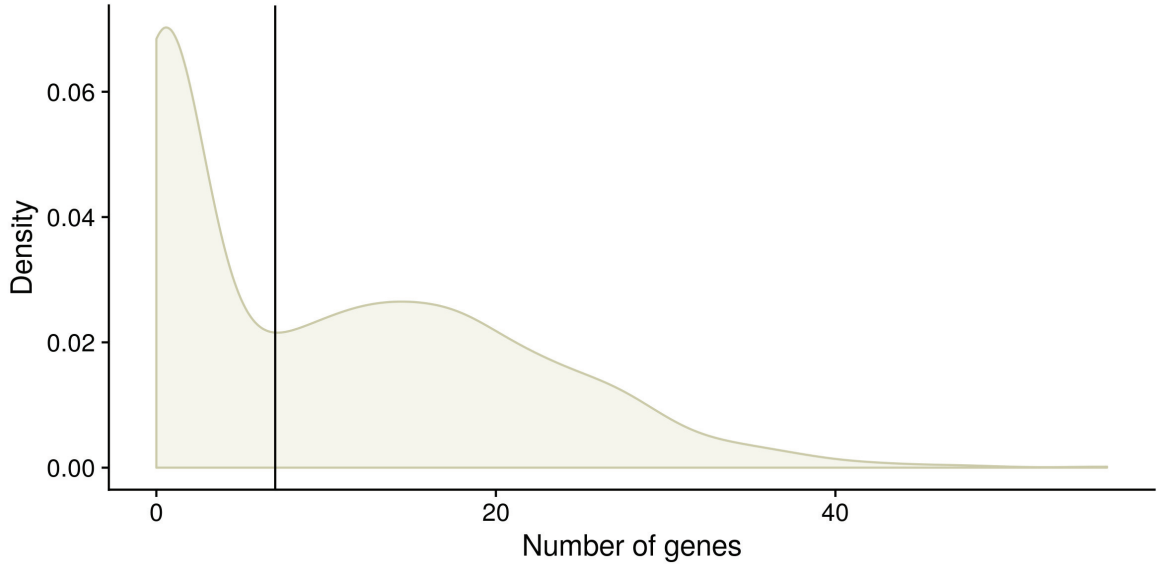
A genome scan for putatively adaptive TE insertions was performed using BayPass with the output of the TE frequency pipeline (v2.2) (GAUTIER, 2015; OLAZCUAGA et collab., 2020). Insertions with Minor Allelic Frequency (MAF)) inferior to 0.025 were removed before the analysis. Autosomal and X-linked contigs were analyzed separately. Three statistics were computed to detect putatively adaptive TE insertions: XtX, C_2 and the Bayes Factor (BF) for Environmental Association Analysis. Briefly, XtX corresponds to a global differentiation statistics, C_2 contrasts allelic frequencies between user-defined groups of populations, and BF measures the support of the association between a marker and a covariate (usually an environmental variable). Because Bayes Factor was computed using the importance sampling algorithm, and single run estimations may be unstable, BF were estimated as the median over five estimates obtained from independent runs of Baypass (GAUTIER et collab., 2018). In accordance to Jeffrey's rule, a BF superior to 20 deciban (dB) was considered as decisive evidence supporting an association (JEFFREYS, 1961). XtX and C_2 estimates came from one single run and simulation was used to determine a significance threshold. The R function simulate.baypass() provided within the BayPass package was used to simulate read count data (nSNP=10000, pi.maf=0). We used the physical coverage estimated from the ppileup file using the module stat-coverage of PoPoolationTE2 (KOFLER et collab., 2016). BayPass was run on this simulated dataset to estimate the null distribution of the XtX and the C_2 statistics. An insertion was considered as overly differentiated (for XtX) or associated to the tested contrast (for C_2) if the corresponding statistics exceeded the 99.9% quantile of the estimated null distribution. The populations whose frequencies were contrasted using the C_2 were: populations of the invasive American road and the native ones (${C_2}^{Am}$), populations of the invasive European road and the native ones (${C_2}^{Eu}$, invasive populations and the native ones (${C_2}^{WW}$). This choice was made according to the invasion roads inferred using microsatellite markers (FRAIMOUT et collab., 2017), the populations structure assessed with SNP/InDel markers called in these samples (OLAZCUAGA et collab., 2020) and the population structure assessed here with TE markers (supplementary fig. 2.S7). For each putatively adaptive insertion, gene vicinity in a 1 kb region centered on the insertion was investigated as described in the paragraph "Genome Wide Association Study with TE family abundance". The presence of the insertion in a region of selective sweep was assessed using Tajima's D. For the 22 populations, we investigated if the Tajima's D estimated in the 1 kb window containing this insertion was inferior to the quantile 0.05 of Tajima's D distribution in this population. More precisely, to prevent for a difference between autosome and X chromosome, autosomal insertions were compared to the autosomal Tajima's D distribution and X-linked insertions to the X chromosome Tajima's D distribution (with autosomal and X-linked contigs as defined in PARIS et collab. (2020). We also checked if the insertion was close to SNPs/InDels previously identified as potentially adaptive during *D. suzukii* invasion (considering a maximum distance of 5 kb) (OLAZCUAGA et collab., 2020).

2.7 Acknowledgements

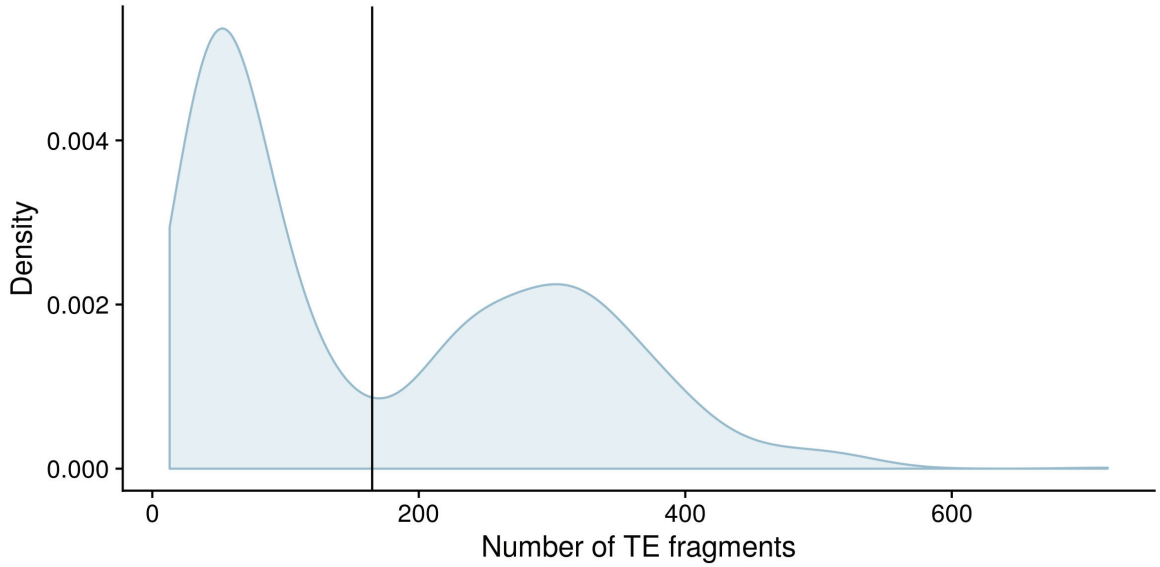
This work was supported by the French National Research Agency (ANR-16-CE02-0015-01 – SWING) and performed using the computing facilities of the CC LBBE/PRABI. We sincerely thank C. Mermet-Bouvier for technical help. We are also grateful to B. Prud'homme and F. Sabot for

constructive discussion about this article.

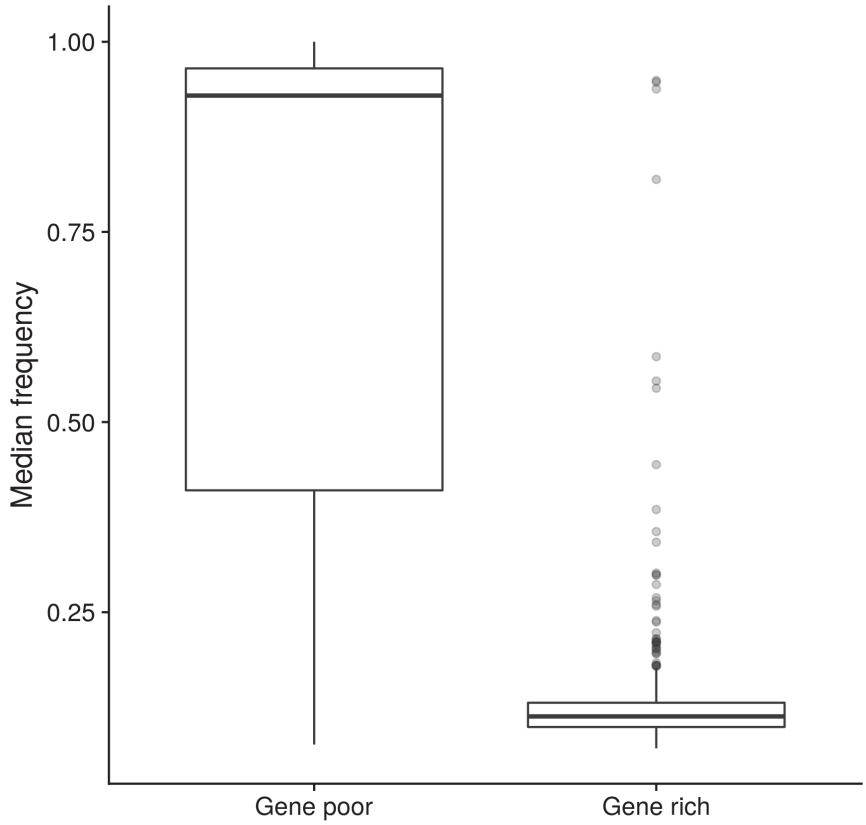
2.8 Supplementary Figures



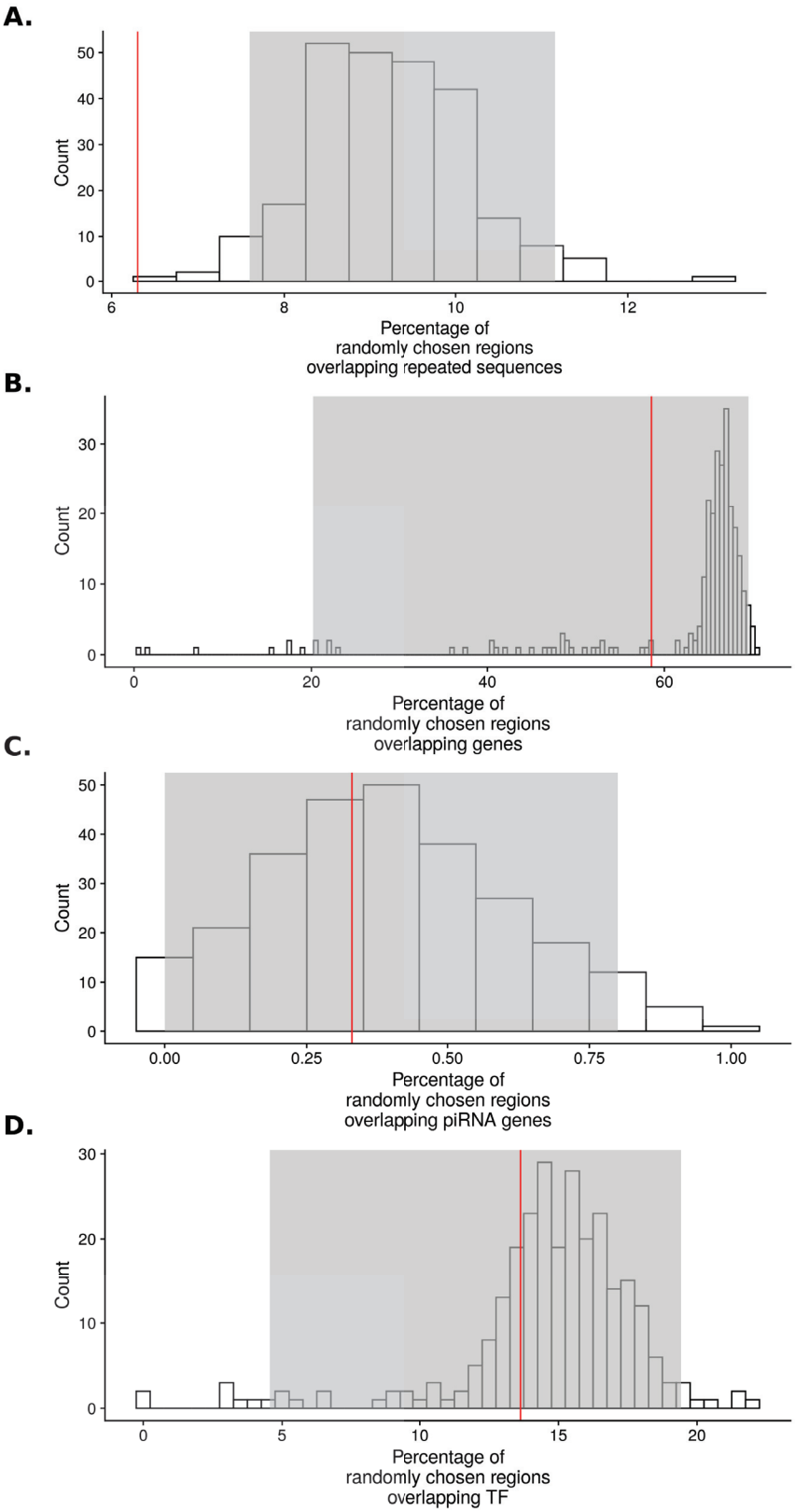
Supplementary figure 2.S1: **Distribution of the number of genes per 200 kb windows in *D. suzukii* assembly.**The vertical line correponds to x=7.



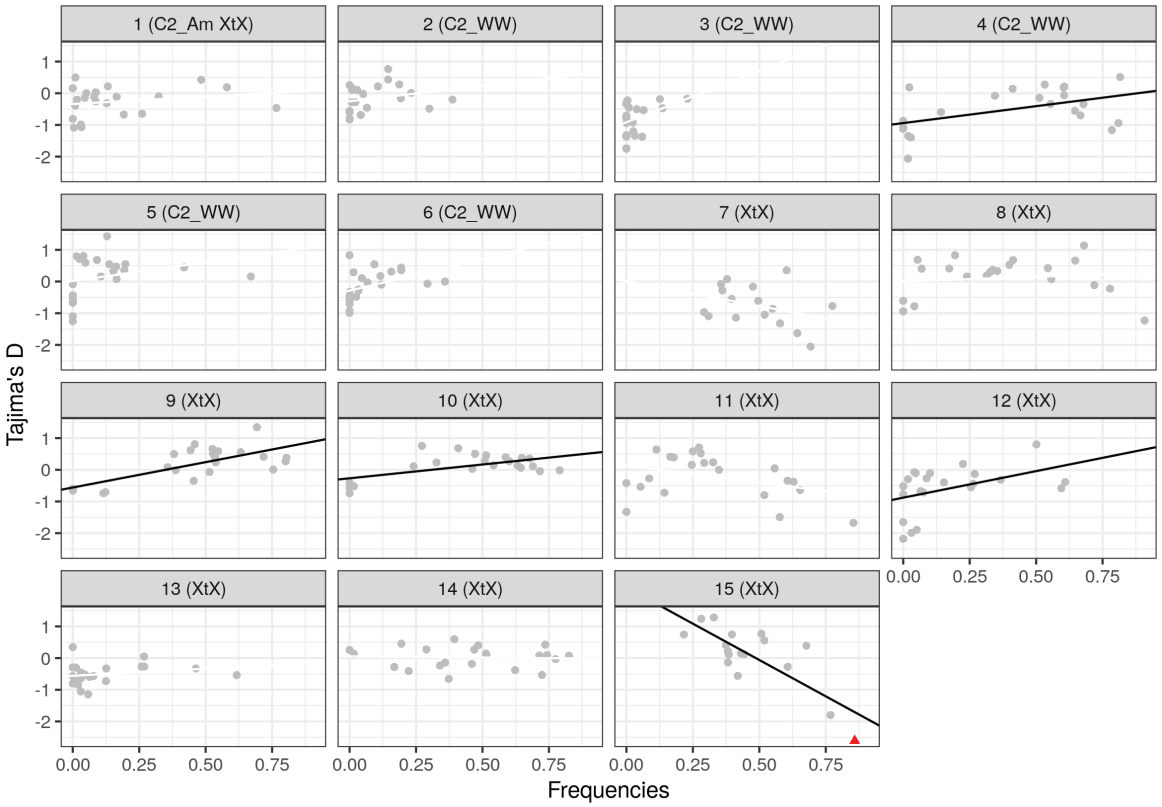
Supplementary figure 2.S2: **Distribution of the number of TEs per 200 kb windows in *D. suzukii* assembly.** The vertical line correponds to x=165.



Supplementary figure 2.S3: **Distribution of the median TE insertion frequency per 200 kb windows in *D. suzukii* assembly for gene-poor (< 7 genes per 200 kb) or gene-rich (≥ 7 genes per 200 kb) windows.** Frequencies were estimated in Watsonville reference population.

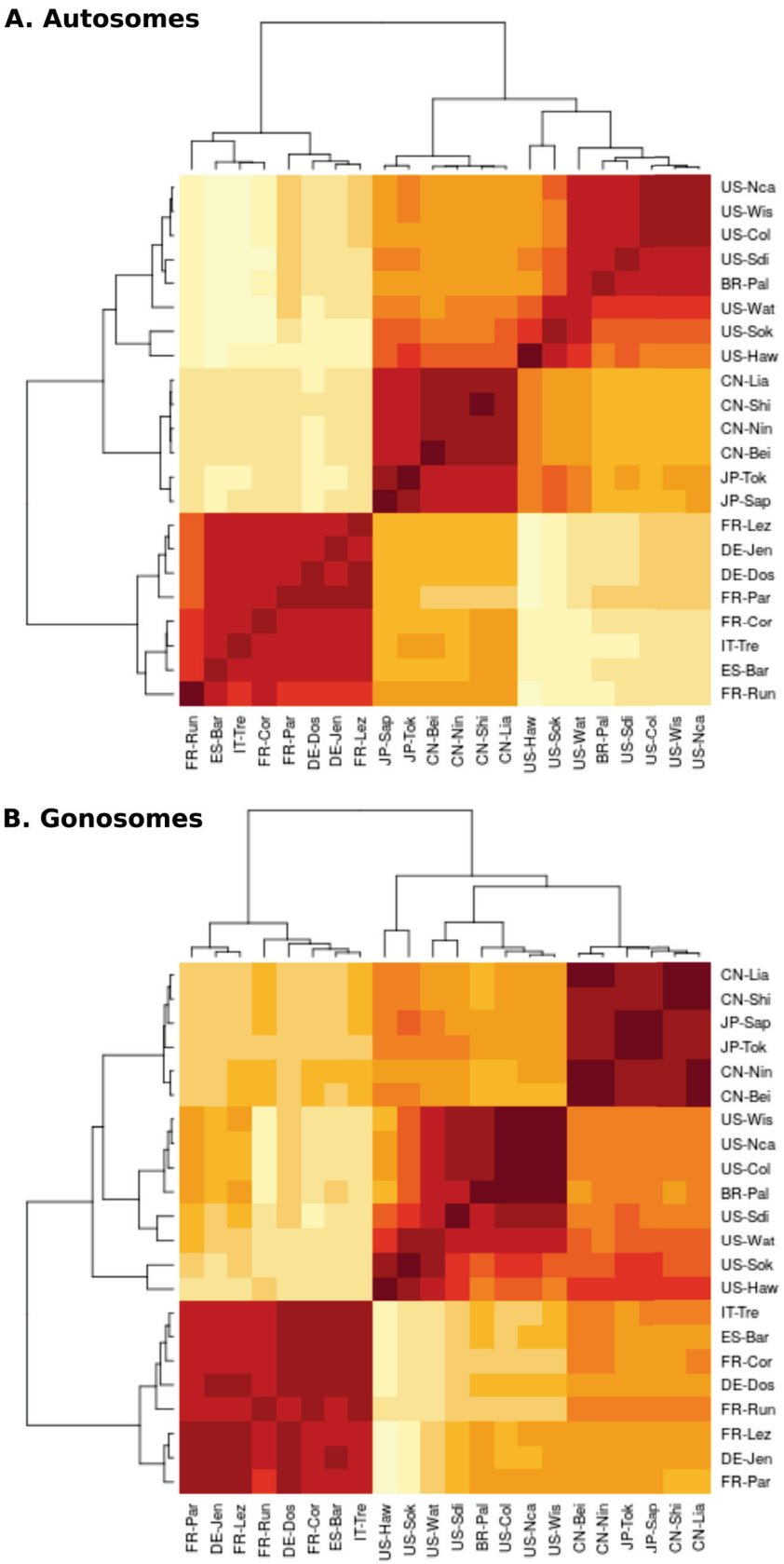


Supplementary figure 2.S4: **Distribution of the percentage of randomly chosen regions surrounding SNPs in *D. suzukii* assembly and overlapping: A. repeated sequences; B. genes; C. genes of the piRNA pathway; D. genes encoding transcription factors (TFs).**250 samples of 1000 regions were used to draw the distribution A., 150 samples of 500 regions for distributions B,C and D. The gray rectangle in the background delimites the portion of the distribution between quantile 2.5% and quantile 97.5%. The vertical red lines correspond to the observed percentage for regions associated with TE abundance.



Supplementary figure 2.S5: **Correlation between insertion frequencies and local Tajima's D estimates in the 22 *D. suzukii* populations for each of the 15 putatively adaptive insertions.** Each panel corresponds to one insertion and Tajima's D are estimated from the 1 kb window containing the insertion. Regression lines are drawn when linear correlations are significant (Pearson's product-moment correlation, $p<0.05$). The red dot indicates that local Tajima's D is inferior to quantile 5% in that population.

Supplementary figure 2.S6: **Tajima's D around the 15 putatively adaptative insertions.** Positions along the contigs (bp) are on the x axis and TE insertions are located at the vertical black lines. Each statistics is estimated using SNPs/InDels in a 1-kb genomic window. Asian populations are in green, American in red and European in blue.



Supplementary figure 2.S7: **Correlation plots of the scaled covariance matrices of population allele frequencies (Omega) among all 22 *D. suzukii* populations based on autosomal (A) and gonosomal (B) TE insertions.**

2.9 Supplementary Tables

Superfamily	Percentage
CMC	0.3995069
Copia	0.2302631
CR1	2.4076456
DNA	0.0493970
Gypsy	13.6487085
hAT	0.1893922
hAT?	0.0054341
Helitron	6.9451585
I	2.4830564
Kolobok	0.0517868
L2	0.4752952
Maverick	4.9174460
Merlin	0.0251339
MULE	0.0123812
P	0.0330418
Pao	6.4408019
Penelope	0.0018600
PIF	0.3977682
PiggyBac	0.0328470
R1	3.2406329
R2	0.0322243
RTE	0.1323261
Sola	0.0001112
TcMar	0.8718515
Unknown	4.0665738
Zator	0.0032957

Supplementary table 2.S1: Percentage of *D. suzukii* assembly occupied by each TE superfamily.

D. melanogaster chromosome	Mb of D. suzukii assembly
2L	51.9
2R	58.8
3L	45.6
3R	50.0
4	2.6
X	31.7

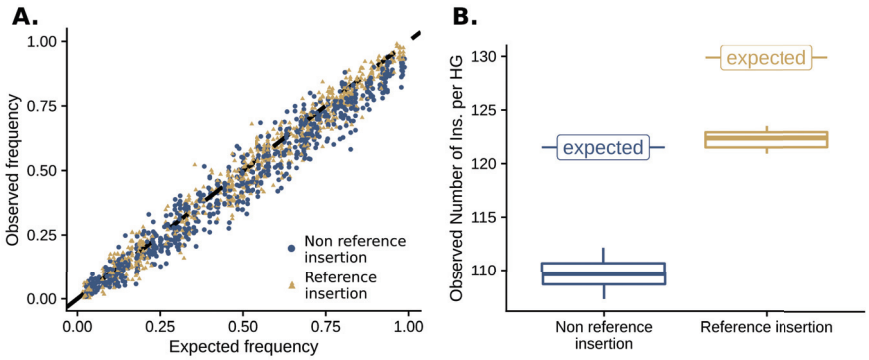
Supplementary table 2.S2: Number of Mb of *D. suzukii* assembly attributed to each of the *D. melanogaster* chromosomes.

	DNA	LINE	LTR	RC	Unknown
High f.	7	6	10	1	6
Intermediate f.	1	3	1	0	1
Low f.	22	25	26	7	17

Supplementary table 2.S3: Number of families with median of High (f>=0.75), Intermediate (0.25<=f<0.75), or Low (f<0.25) frequency in *D. suzukii* reference population for each TE order. Only families with more than 10 insertions are considered

2.10 Supplementary Methods

The accuracy of the TE calling procedure was validated by a simulation work using simulaTE (KOFLER, 2018). As a starting point, an artificial genome devoid of TEs was created by removing masked nucleotides in a randomly selected 1 Mb chunk of the masked assembly. The resulting genome was 607,100 bp long. A population of 1000 diploid individuals each displaying 500 insertions, of frequencies ranging from 0.01 to 0.99, was generated by inserting TE sequences in the artificial genome. A reference genome, containing 250 of these insertions was also created. This artificial population was used to simulate read data corresponding to each of the 22 PoolSeq samples. For each sample, x haploid genomes were drawn according to the exact number of individuals in the sample. Reads were simulated using simulaTE and mimicking the coverage and insert size of the original sample. We used the coverage estimated in OLAZCUAGA et collab. (2020) and the inner distance, i.e. insert size -2*Read length extracted from the ppileup file header. The standard deviation on the inner distance was set to 100 bp. The TE frequency and TE abundance pipelines described in the Materials and Methods section were then run on this dataset.



Supplementary figure 2.S8: **Validation of the TE frequency and TE abundance pipelines: expectations vs observations in a 22 samples simulated dataset mimicking the original dataset A.Estimation of TE insertion frequencies.** Observed frequencies in the simulated dataset are compared to expected frequencies. Insertions absent from the reference genome are shown in blue whereas insertions present in the reference genome are in gold. **B. Estimation of TE abundance.** Distribution of the numbers of haploid insertions for the 22 simulated samples for both non reference insertions and reference insertions. The expected numbers of insertions per haploid genome (HG), 121.5 for non reference insertions and 129.9 for reference insertions, are indicated by a horizontal segment.

A run of our pipelines on a simulated dataset mimicking the original *D. suzukii* dataset indicated that our methods estimate accurately TE insertion frequencies and TE abundances (as the numbers of TE insertions per haploid genome (HG) per population) with little variation between population samples. Regarding the TE frequency pipeline, overall 10,419 TE insertions were called in the simulated dataset (supplementary fig. 2.S8A). 10,353 of these were true positive (99.37%), 66 were false positive (0.63%). 647 insertions out of the 11,000 simulated were not recovered by PoPoolationTE2, corresponding to a false negative rate of 5.88%. The mean number of TE insertions called per sample was 473.59 (sd = 1.30), with an average of 470.59 true positives (sd = 1.30) and 3 false positives (sd = 0). The average number of false negatives was 29.40 (sd = 1.30). We found an effect of the presence of the considered insertion in the reference genome on the ability to be detected ($\chi 2$ = 32.19, df = 1, p-value = 1.4 x $10^{-}8$), insertions present in the reference genome being missed more often. The differences between expected and observed TE frequencies were

poorly explained by variations in number of individuals, coverage or inner distance between samples, or their interactions (R^2 = 0.56%, square-root transformed Y variable). Concerning the TE abundance pipeline, the mean number of insertions per haploid genome (HG) per sample was 234.48 (sd = 1.29) for an expectation of 251.41. On average 2.59 insertions per HG (sd = 0.28) were due to false positives. A mean of 109.63 non reference insertions per HG were recovered (sd = 1.37) over the 121.52 expected. On average, 122.26 reference insertions per HG were recovered (sd = 0.84) over the 129.89 expected (supplementary fig. 2.S8B). The difference between the mean number of insertions per HG and the expectation was higher for reference insertions compared to non-reference insertions (t = -21.924, df = 35.348, p-value < 2.2e-16). The difference between the observed mean number of insertions per HG and the expectation was poorly explained by differences in number of individuals, or coverage or inner distance between samples, or their interactions (F-statistic = 1.509, df = 7-14, p-value = 0.24, R^2 = 0.43%, adjusted R^2 = 0.145).

3 Massive and localized proliferation of Transposable Elements in *Drosophila suzukii* genome without genome-wide increase in dN/dS

Vincent Mérel[1], Théo Tricou[1], Nelly Burlet[1], Marie Fablet[1], Cristina Vieira[1], Annabelle Haudry[1]

Corresponding author: annabelle.haudry@univ-lyon1.fr

Affiliations:

1: Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, F-69622 Villeurbanne, France

Sommaire

3.1 Avant-propos

Ce travail a été réalisé afin de mieux comprendre les aspects de la dynamique des ET derrière le très fort pourcentage d'ET dans le génome de *D. suzukii*. En effet, une étude réalisée au laboratoire en 2016, montre que 33% du génome de *D. suzukii* sont occupés par des ET (SESSEGOLO et collab., 2016). C'est l'espèce de drosophile avec le plus fort pourcentage d'ET parmi les 26 de cette étude. Le travail d'annotation du génome de réference de *D. suzukii*, que j'ai effectué au début de ma thèse, a même porté ce pourcentage à 47% (voir Chapitre 2). C'est ce très fort pourcentage, couplé à l'observation que chez *D. biarmipes* seuls 20% du génome sont occupés par des ET, qui a motivé le travail presenté dans ce chapitre. En effet, la date de divergence entre *D. biarmipes* et *D. suzukii* ayant été estimée à ~7 millions d'années (OMETTO et collab., 2013), cela suggère une forte accumulation d'ET dans le génome de *D. suzukii* dans un passé proche. Nous avons donc commencé une étude de génomique comparative entre *D. suzukii* et sept espèces proches, en tirant parti notamment d'un accès à des données de séquençage privées du génome de *D. subpulchrella*, espèce ssupposée de *D. suzukii*. Initialement, l'objectif était l'étude de l'évolution quantitative et qualitative du contenu en ET dans le temps, à la recherche notamment des signes d'un potentiel burst, i.e. le déplacement d'un grand nombre de séquences d'ET à travers le génome sur un temps très court. Ce travail s'est étoffé par la suite, avec notamment l'étude de la variabilité intra-génomique dans l'accumulation des ET et celle d'un potentiel rôle d'un relâchement de la sélection purificatrice. Presenté ici sous la forme d'un article, ce chapitre à néanmoins vocation à être encore approfondi. La discussion sera notamment étoffée, et particulièrement le passage quant à l'utilisation du dN/dS pour tester l'hypothèse d'un relâchement de la sélection.

3.2 Abstract

Transposable Elements (TEs) are selfish genetic elements that can have major effect on host fitness. Despite their ubiquity, their dynamics of proliferation remains unclear. TEs may occupy different proportions of the genome, even between closely related species. In *suzukii* subgroup, a massive accumulation of TEs led to one of the highest genomic TE content reported in *Drosophila* so far: the one of the spotted-wing fly, *D. suzukii*. In this study, we reconstructed a phylogeny based on whole-genome data (2283 genes), and analyzed the TE content for eight closely related species. While approximately a third of *D. suzukii* genome, and that of its sister species *D. subpulchrella*, are composed of TEs (~100 Mb and 90 Mb respectively), their closest relative *D. biarmipes* has three times less TEs in its genome (~30 Mb). The estimated divergence among TE sequences suggest that a TE proliferation started 4 Millions years ago, which is also our estimate of the date of divergence between *D. suzukii-D. subpulchrella* ancestor and *D. biarmipes*. A comparison of *D. suzukii* and *D. biarmipes* long-reads based assemblies, indicates that the TE proliferation spared gene rich genomic regions but greatly affected surrounding regions. The analysis of evolutionary rates for 2283 orthologous genes did not reveal any link with a genome-wide relaxation of selection, suggesting a small effect of a weaker TE removal. Interestingly, we found the TE proliferation to be associated with strong positive selection acting on genes related to chromatin state, a possible answer to the localized proliferation.

3.3 Introduction

Transposable Elements (TEs) are selfish genetic elements. Mostly neutral or deleterious for their hosts, these sequences persist, and even proliferate, by copying and pasting themselves to other genomic locations (Burt et Trivers, 2006; Charlesworth et Charlesworth, 1983; Doolittle et Sapienza, 1980; Orgel et Crick, 1980). Ubiquitous and diverse (Wicker et collab., 2007), they have been proven to drastically impact genomes, e.g. genomic size (Elliott et Gregory, 2015; Sessegolo et collab., 2016; Vitte et Panaud, 2005), and phenotypes, e.g. lifespan (Nikitin et Woodruff, 1995). Although ubiquitous, TEs may occupy different proportions of the genome, even between closely related species. For instance, 12% of the nematod genome is composed of TE sequences against 85% for wheat (C. elegans Sequencing Consortium, 1998; Wicker et collab., 2018). In the *Oikopleura* genus, 15% of *O. dioica* genome is composed of TEs against 50% for *O. vanhoeffeni* (Naville et collab., 2019).

Such variations in TE content among species have long puzzled scientists. Studies revealed a complex dynamic of proliferation depending on a complex balance between transposition rates, selection against TEs, and genetic drift. The accumulation rate, which depends on the difference between the insertion rate and the removal rate, is not constant over time (Bergman et Bensasson, 2007). Some lineages may thus experience intense episodes of TE accumulation. For instance, in *Leptidea* the average rate of TE-driven genome expansion is close to 4 Mb/My, however a particular population of *Leptidea juvernica* is experiencing a rate of genome expansion of 72 Mb/My (Talla et collab., 2017). Importantly, TE accumulation may be modulated by TE elimination through small deletions. In salamanders, slower DNA loss in non-LTR retrotransposons, may contributes to

genomic gigantism (SUN et collab., 2012).

Because factors affecting the dynamics of TE proliferation vary in intensity at the genomic scale, one may expect a spatial heterogeneity in TE accumulation. Especially, the selection against TEs is expected to be stronger in gene vicinity, where TEs may alter the expression/function of nearby genes (MÉREL et collab., 2020a), and in highly recombining regions, where TEs promote recombination between non-homologous genomic fragments (ectopic recombination) (PETROV et collab., 2003). TEs should thus accumulate slowly in these regions. The intensity of transposition rate can also vary at the genomic scale, and may mitigate or reinforce the effect of local variation in the strength of selection on TE accumulation rate. Particularly, gene expression could render DNA more accessible to insertion in gene vicinity (FONTANILLAS et collab., 2007), or TE could also have evolve mechanisms to avoid insertion where they will be strongly counter-selected, e.g. within or close to genes (SULTANA et collab., 2017).

Because purifying selection is expected to slow down TE proliferation, and population effective size (N_e) is affecting selection efficacy ($N_e s$ with s the selection coefficient), N_e have been proposed as a key parameter controlling TE proliferation (LYNCH et CONERY, 2003). As selection efficacy is positively correlated to N_e, TEs are expected to accumulate in lineages where N_e is reduced. In agreement with this hypothesis, LYNCH et CONERY (2003) found a negative correlation between synonymous diversity (an estimator of $4N_e\mu$, μ being the mutation rate) and genome size among various species (LYNCH et CONERY, 2003).

Thereafter, several studies have shown that a reduction of N_e resulted in an accumulation of TEs (in subterranean *Aselloidea* (LEFÉBURE et collab., 2017); in polyploids plants (BADUEL et collab., 2019). However, it has also been shown that the correlation found by Lynch Connery was no longer true when taking into account phylogenetic inertia (WHITNEY et collab., 2010), as shared common evolutionary history has a strong impact on TE content. Furthermore, in Drosophila, the study of 12 distant species revealed that greater euchromatic TE abundance was associated with greater level of purifying selection (CASTILLO et collab., 2011). The role of genome-wide reduction of selection efficacy is thus still unclear.

Rather than genome-wide reduction in selection efficacy, one could also propose that it is only the reduction in selection efficacy of some particular genes that matter (CASTILLO et collab., 2011). Because of the overall potential effect of TEs, their hosts have evolved defense mechanisms. In Drosophila, a particular pathway called the piRNA pathway protects the germline against TE proliferation/mobilization (OZATA et collab., 2019). In brief, some loci constituted of more or less degraded TE sequences (called piRNA clusters), are transcribed in small RNAs called piRNAs. These piRNAs operate by sequence complementarity to silence TEs both at the transcriptional and post transcriptional level. Arguably, a reduction of selection efficacy on genes of the piRNA pathway could be associated with an increase of TE content.

The spotted-wing fly *D. suzukii* provides a great opportunity to investigate the dynamic of TE proliferation. Recent studies revealed that this invasive species has a much larger genome (~330-340 Mb) compared to others Drosophila species (219 Mb in average) (HJELMEN et collab., 2019; SESSEGOLO et collab., 2016), associated with a high TE content (~33-47% of the genome) (MÉREL et collab., 2020b; SESSEGOLO et collab., 2016). *D. biarmipes*, a close relative to *D. suzukii* (~7 Mya divergent) (OMETTO et collab., 2013), possesses a relatively small genome (~196 Mb) and

proportion of TEs (~20-25% of the genome) (KIM et collab., 2020; SESSEGOLO et collab., 2016). This suggests an important and quite recent proliferation of TEs.

In this paper, we compared the TE content of *D. suzukii* with those of seven close relatives to get insights into the TE dynamics underlying *D. suzukii* high TE content. In addition to *D. suzukii* and *D. biramipes*, our dataset comprises another species from the *suzukii* subgroup: *D. subpulchrella*. *D. subpulchrella* is expected to be *D. suzukii* sister species (ROTA-STABELLI et collab., 2020). We showed that, as *D. suzukii*, *D. subpulchrella* has a high TE content. We demonstrated that the accumulation of TEs started approximately 4 Mya, within *D. suzukii*-*D. subpulchrella* ancestor, and is probably still going on, at least in *D. suzukii*. The associated genomic expansion appears to be restricted to regions of low gene density. We did not find any evidence of genome-wide relaxation of selection using dN/dS estimations.

3.4 Results

3.4.1 A TE proliferation started less than 4 Mya

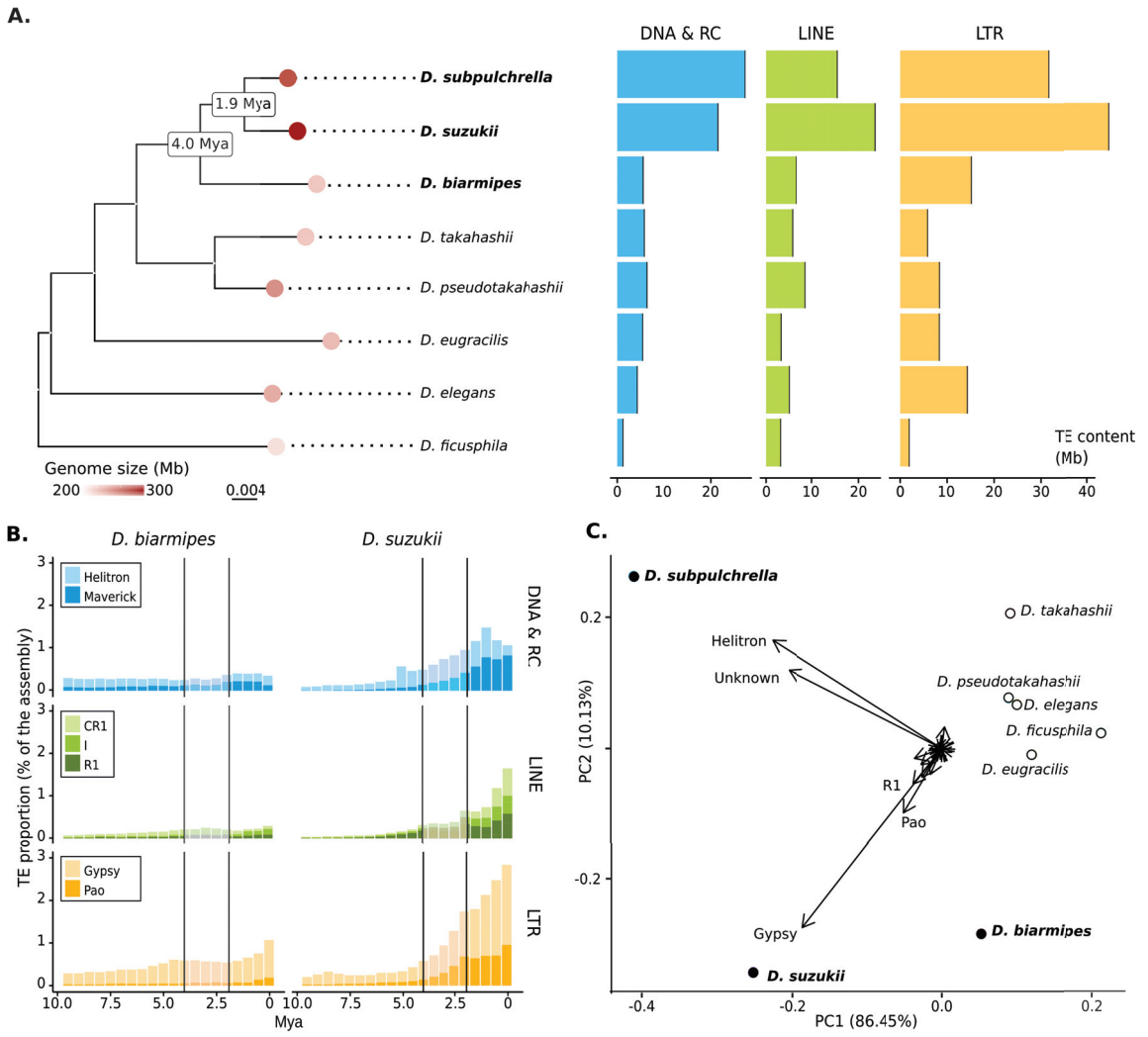


Figure 3.1: **Phylogenetic relationships, genome sizes and TE contents in *D. suzukii* and close relatives. A. Phylogenetic tree figuring genome sizes and overall amounts of TEs for each species.** The tree was reconstructed using 2283 orthologous genes using the LG+G4 model. Genome size was estimated using reads kmers occurrences distribution, and genomic TE content was estimated using deviaTE (WEILGUNY et KOFLER, 2019) (DNA and Rolling Circles (RC) elements in blue, LINEs in green and LTR elements in orange). Species of the *suzukii* subgroup are in bold. The black vertical segment bars represent standard deviation in the three different reads subsamples. Note that these bars are superimposed. **B. TE landscapes in *D. suzukii* and in *D. biarmipes*.** The percentage of each TE superfamily in the genome assembly estimated using RepeatMasker is represented as a function of its age. For graphical reasons, only superfamilies representing more than 1% of *D. suzukii* or *D. biarmipes* assembly are represented. Landscapes for unknown TE sequences are presented in supplementary fig. 3.S1). **C. PCA of TE superfamilies copy numbers.** Species of the *suzukii* subgroup are represented by full dots. Replicates are superimposed. The first axis (PC1), separating *D. suzukii* and *D. subpulchrella* from the other species, explains 86% of the total variation.

To get insight into the dynamics of TEs in *suzukii* subgroup, we combined phylogenetic and comparative genomics approaches. Using alignments of 2283 orthologous genes, we reconstructed phylogenetic relationships for three species of the *suzukii* subgroup and five closely related species. Our results confirmed that *D. suzukii* sister species is *D. subpulchrella* (ROTA-STABELLI et collab., 2020). We found a divergence date between the two species to be approximately 1.89 Mya (I C 95% =[1.88; 1.90], fig. 3.1A). *D. biarmipes* is a more distant relative of *D. suzukii* with a common ancestor living 3.97 Mya (I C 95% =[3.95; 3.99]). Note that, as expected from the analysis of morphological characteristics, mating behaviour and hybridization tests (DWIVEDI et GUPTA, 1981),

D. pseudotakashii appears to be closely related to *D. takahashii*.

For each species, genome size was estimated using reads kmers occurrences distribution, and genomic TE contents by countings PE reads mapping on a homemade TE database with the software deviaTE (WEILGUNY et KOFLER, 2019). Our analysis reveals that, with 29.48% (97.20 Mb out of a 329.67 Mb genome) and 29.67% (87.45 Mb out of a 294.67 Mb genome) of TEs respectively, *D. suzukii* and *D. subpulchrella* harbor a largely higher TE content than the six other species (20.63 ± 6.92 Mb). Especially the other genome of the *Suzukii* subgroup, namely *D. biarmipes* genome, contains only 28.49 Mb of repeated sequences. The higher TE content in *D. suzukii* and *D. subpulchrella* holds when considering individually DNA Rolling Circle (RC), LINEs or LTR elements (fig. 3.1A). Note that for each species we evaluated the TE content on three different PE reads subsamples and that we observe very little variation (with a maximum standard deviation on overall TE content of 0.0090 Mb in *D. pseudotakashii*).

To ensure the reliability of our results we also estimated TE content with another method, the one implemented in dnaPipeTE (GOUBERT et collab., 2015). If deviaTE estimates are markedly lower than dnaPipeTE estimates, overall we obtained a significant linear correlation between estimates of the two methods (t = 14.03, df = 6, p-value = 8.178e-06, supplementary fig. 3.S2). Furthermore, with both methods, TE estimates are strongly positively correlated with genome size ($t_{deviaTE}$ = 7.8737, $df_{deviaTE}$ = 6, p-$value_{deviaTE}$ = 0.0002222 ; $t_{dnaPipeTE}$ = 14.356, $df_{dnaPipeTE}$ = 6, p-$value_{dnaPipeTE}$ = 7.151e-06; supplementary fig. 3.S3), as it has been found several times know (HILL, 2019; SESSEGOLO et collab., 2016)

To get an absolute datation of the TE proliferation event that led to *D. suzukii* high TE content, we draw *D. suzukii* TE landscape, i.e. the genomic percentage of each TE superfamily as a function of insertion age (fig. 3.1B). Assuming that old TEs have accumulated mutations, and young TEs are similar to consensus sequences, insertion age is derived from the divergence to the consensus using Drosophila substitution rate. As the closest *D. suzukii* relative which did not underwent the TE proliferation according to its low TE content, *D. biarmipes* TE lanscape was also drawn for comparison. The results show that, whereas TE accumulation remained globally stable over the last 10 Mya in the genome of *D. biarmipes*, an amplification of TE content started in *D. suzukii* genome around 4 Mya, which corresponds to the estimated divergence time between *D. biarmipes* and *D. suzukii*. Remarkably, the accumulation of TEs gradually increased in intensity. This proliferation of TEs within *D. suzukii* concerns seven superfamilies: Helitron, Maverick, CR1, I, R1, Gypsy and Pao. In term of genome occupation, Gypsy shows the most important amplification. Note that for Helitron elements, the accumulation rate slowed down ~2 Mya.

To confirm that the TE proliferation started in the ancestor of *D. suzukii* an *D. subpulchrella*, and that the two species did not experience a more recent and independent increase in TE copy numbers, we compared TE content at a finest scale. We performed a Principal Component Analysis (PCA) on TE superfamilies copy numbers on the 8 species of the dataset (fig. 3.1C). The first axis, that separate *D. suzukii* and *D. subpulchrella* from the other species, explains 86.45% of the variance. This demonstrates a similar TE content in both species as compared with other species. It is congruent with a TE proliferation started preliminary to *D. subpulchrella* divergence with *D. suzukii*, as suggested by the TE landscape (Figure 3.1B). The contribution of the different variables indicates that Helitron, Unknown, Gypsy, R1, Pao are the most numerous elements in these two

species. The second axis (10.13% of the variance), opposing *D. suzukii* and *D. subpulchrella*, highlights the differences in TE composition between the two species: *D. suzukii* is characterized by more copies of Gypsy whereas more copies of Helitron and Unknown elements are found in *D. subpulchrella*. These "specific" elements have either been accumulated in each genome recently, after their divergence (1.9Mya), or have been inserted in the ancestor and differentially lost recently in only one of the species.

3.4.2 An uneven genomic proliferation

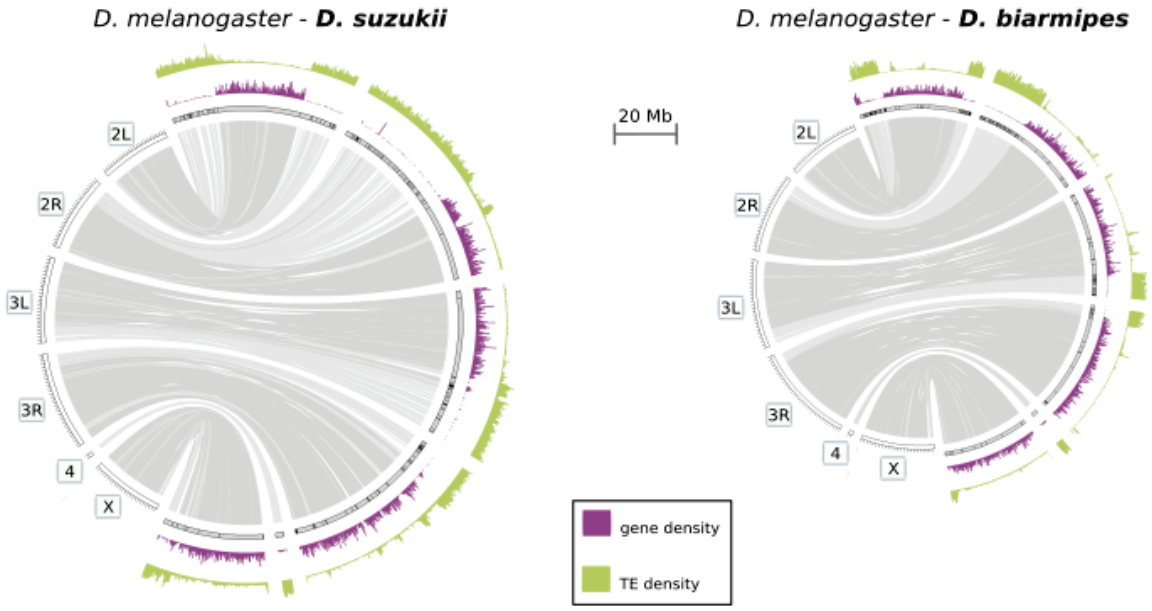


Figure 3.2: **TE density and gene density in *D. suzukii* (left) and *D. biarmipes* (right) assemblies.** TE density (green outer graph) and gene density (purple inner graph) are shown for windows of 200 Kb. The homologous relationships with *D. melanogaster* assembly are shown inside using light links for regions of low gene density (<6 genes per 200 Kb) and dark links for regions of high gene density (>=6 genes per 200 Kb). The maximum gene density of the plot is 56 for *D. suzukii* and 55 for *D. biarmipes*. The maximum number of TE fragments is 2233 for *D. suzukii* and 2499 *D. biarmipes*.

In order to characterize intra-genomic variations of TE accumulation, we first compared the distributions of TE and gene density between *D. biarmipes* and *D. suzukii* long-reads assemblies. We especially verified that, has previously found for *D. suzukii* MÉREL et collab. (2020b), *D. biarmipes* assembly can be divided in gene rich/TE poor and gene poor/TE rich windows. Second, we tested if gene rich and gene poor regions underwent the same expansion in *D. suzukii* lineage during the TE proliferation. Note that *D. suzukii* annotation come from MÉREL et collab. (2020b) and we established a similar annotation for *D. biarmipes*. Briefly, syntenic relationships between assembly contigs and *D. melanogaster* chromosomes were established using homology (KURTZ et collab., 2004), and distributions of TE and gene density in the assembly were estimated from a run of RepeatMasker and a run of augustus respectively (http://www.repeatmasker.org/)(STANKE et collab., 2008).

As previously reported for *D. suzukii* (MÉREL et collab., 2020b), *D. biarmipes* genome is highly structured regarding the distribution of genes and TEs (fig. 3.2). *D. biarmipes* distribution of gene

density in 200 kb genomic windows is bimodal (supplementary fig. 3.S4). Based on this distribution, and *D. suzukii* distribution of gene density, we chose a common threshold to divide both assemblies in gene rich and gene poor windows. This threshold was set to six genes per 200 Kb. Using this threshold, we found 127 Mb to be classified as gene rich and 31 Mb as gene poor for *D. biarmipes*, whereas 112 Mb were classified as gene rich and 108 Mb as gene poor for *D. suzukii*. Concerning TE density, 122 Mb are TE poor and 40 Mb are TE rich in *D. biarmipes* assembly (with TE poor windows having less than 165 TE fragments and TE rich windows 165 TE fragments or more; supplementary fig. 3.S5). 133 Mb are TE poor and 108 Mb are TE rich in *D. suzukii* assembly. Note that, if *D. suzukii* distribution of TE density is clearly bimodal, it is less clear for *D. biarmipes*. In *D. biarmipes* genome, TE rich windows are less abundant and highly variable in number of TE fragments. For both *D. biarmipes* and *D. suzukii* the classification of a window as TE rich or TE poor is not independant of its classification as gene rich or gene poor (X-squared = 121.11, df = 1, p-value < 2.2e-16, and X-squared = 199.2, df = 1, p-value < 2.2e-16). TE rich windows tend to be gene poor and TE poor windows tend to be gene rich.

To investigate if gene rich and gene poor regions in *D. suzukii* lineage underwent the same expansion during the TE proliferation we compared their size in *D. suzukii* long-reads assembly with the size of homologous regions in *D. biarmipes* long-reads assembly. To do so, we made the following assumption: *D. biarmipes* and *D. suzukii* genomic regions homologous to the same region of *D. melanogaster* are homologous. Homologous relationships with *D. melanogaster* genome have been established for 207 of the 661 contigs of *D. biarmipes*, i.e. 161 Mb out of 182 Mb of the assembly (fig. 3.2, supplementary table 3.S1). Based on gene density, we splitted *D. suzukii* assembly in 14 large regions: seven gene poor and seven gene rich (table 3.1). The size of the resulting regions ranges from 2.1 Mb to 34.0 Mb for gene poor regions ($\hat{\mu}$ = 14.0 Mb), and from 0.50 Mb to 39.1 Mb for gene rich regions ($\hat{\mu}$ = 20.5 Mb). The size of homologous regions in *D. biarmipes* assembly ranges from 1.4 Mb to 13.1 Mb for gene poor regions ($\hat{\mu}$ = 5.1 Mb), and from 0.3 Mb to 30.0 Mb for gene rich regions ($\hat{\mu}$ = 18.05 Mb). Gene rich regions in *D. suzukii* are slightly larger than homologous regions in *D. biarmipes*, with an average size ratio of 1.23 (sd=0.36). Gene poor regions in *D. suzukii* are markedly larger than homologous regions in *D. biarmipes*, with an average size ratio of 2.84 (sd=1.03). The size ratio was significantly different between gene poor and gene rich regions (t = 3.8928, df = 7.4205, p-value = 0.005313).

3.4.3 No genome-wide increase in dN/dS

To test if the observed TE proliferation was associated with a genome-wide relaxation of selection, we investigated the ratio of divergence at non-synonymous and synonymous sites, dN/dS or ω, for 2283 orthologous genes. In this set of genes conserved in Diptera, we expect a relaxation of selection to drive a higher fixation rate of weakly deleterious mutations, and thus a higher non-synonymous substitution rate and ω. We explored the possibility of a relaxation of selection in: **A.** *D. suzukii* and *D. supbulchrella* ancestor ; **B.** *D. suzukii*, *D. supbulchrella*, and their ancestor ; **C.** *D. suzukii* and *D. supbulchrella*. Note that testing different scenarii allow us to account for temporal variations in selective pressures. We perfomed likelihood ratio tests between models from codeml to compare ω values between lineages of interest (A, B and C) and the rest of the phylogeny. More specifically, for each of the 2283 genes, we performed two likelihood ratio tests. The first one,

Regions in *D. melanogaster*	Gene density in *D. suzukii*	Size of the region in *D. suzukii* (Mb)	Size of the region in *D. biarmipes* (Mb)	Size ratio
2L:6000000-6450000	gene poor	14.4	6.1	2.37
2L:22050001-23513712	gene poor	13.9	3.0	4.58
2R:0-6000000	gene poor	34.0	13.1	2.60
3L:23000000-28110227	gene poor	17.4	6.4	2.73
3R:0-5000000	gene poor	11.4	3.6	3.16
4:0-1100000	gene poor	2.1	1.8	1.16
X:21500000-23542271	gene poor	4.5	1.4	3.28
2L:0-5999999	gene rich	6.9	7.5	0.92
2L:6450001-22050000	gene rich	18.6	17.5	1.06
2R:6000001-25286936	gene rich	23.8	21.0	1.14
3L:0-22999999	gene rich	27.2	25.1	1.08
3R:5000001-32079331	gene rich	39.1	30.0	1.30
4:100000-1348131	gene rich	0.5	0.3	2.00
X:0-21499999	gene rich	27.1	25.0	1.09

Table 3.1: **Evolution of gene rich/gene poor regions size in *D. suzukii* lineage.** Seven gene poor and seven gene rich genomic regions were manually delimited in *D. suzukii* assembly zooming in on fig. 3.2. The homologous regions in *D. melanogaster* are indicated, and were used to retrieve homologous regions in *D. biarmipes* (assuming that *D. biarmipes* and *D. suzukii* genomic regions homologous to the same region of *D. melanogaster* are homologous). For each region delimited in *D. suzukii* assembly, its size, as well as the size of the homologous region in *D. biarmipes* assembly, is indicated. The last column corresponds to the ratio of size between the region in *D. suzukii* and its homologous in *D. biarmipes*.

			Number of rejected H0		
Test	Interpretation	Lineage	Total	$\omega_{Lineage} > \omega_{Tree}$	$\omega_{Lineage} < \omega_{Tree}$
One-ratio (M0) vs two-ratio	Difference on the overall signature of selection	**A**	69 (2245)	51	18
		B	271 (2271)	187	84
		C	258 (2272)	171	87
M2 rel vs CmC	Divergent evolution of some sites	**A**	360 (2133)	190	170
		B	550 (2131)	252	298
		C	440 (2054)	163	277

Table 3.2: **Variations of ω in *suzukii* subgroup.** Two likelihood-ratio tests contrasting null and alternative models from codeml were used to assess variations of selection in lineages of *suzukii* subgroup (YANG, 2007). The first one, one-ratio (M0) vs two-ratio, assumes that all sites evolve similarly, and detects differences in overall ω. The second one, M2 rel vs CmC, relaxes the hypothesis of unvarying evolution along the gene, and identifies divergent evolution at a given, and estimated, proportion of sites. These two tests were performed on the 2283 orthologs recovered in all the eight studied species. Each test was performed focusing on three different lineages: **A.** *D. suzukii* and *D. subpulchrella* ancestor, **B.** *D. suzukii*, *D. subpulchrella* and their ancestor, **C.** *D. suzukii* and *D. subpulchrella.* For each test, the total number of rejections of the null hypothesis is reported. This number is divided between cases where ω is greater, or less, in the considered lineage as compared to the rest of the tree. Note that values of ω > 10 were removed, so the total number of genes per test, indicated in brackets, vary.

one-ratio (M0) vs two-ratio, assumes that all sites evolve similarly, and detects differences in overall ω. The second one, M2 rel vs CmC, relaxes the hypothesis of unvarying evolution along the gene, and identifies divergent evolution at a given, and estimated, proportion of sites.

Considered lineages have a difference of overall ω with the rest of the tree for only 3-12% of the

genes tested (one-ratio (M0) vs two-ratio test, table 3.2), indicating an absence of genome-wide relaxation of selection. When significantly different, overall ω values tend to be superior in studied lineages (66-74% of the significant genes, 2-8% of the genes tested). This suggests either a weaker purifying selection or strong positive selection. Actually, few of the genes with superior ω values in studied lineages display strong signature of selection, i.e. ω > 1, 2 over 51 in the lineage A, 0 over 187 in the lineage B and 0 over 171 in the lineage C. The study of overall ω therefore indicate a plausible relaxation of selection but for less than 10% of the genes.

Relaxing the hypothesis that all sites evolve similarly (M2 rel vs CmC test), we found a higher proportion of genes with divergent evolution in lineages of interest. Between 17% and 26% of genes present a proportion of sites evolving differently (table 3.2). This proportion does not necessarily display a higher ω in studied lineages (37-53% of the significant genes, 8-12% of the genes tested). It suggests once again an absence of genome-wide relaxation of selection. Moreover, regardless of the lineage, the majority of genes with a higher ω have a ω>1 (135 over 190 in the lineage A, 186 over 252 in the lineage B and, 112 over 163 in the lineage C), suggesting a strong positive selection rather than a relaxation of selection.

3.4.4 Gene Ontology of genes with a different dN/dS

Because a relaxation of selection acting on genes implied in the defense against TEs, like those of the piRNA pathway, may be sufficient to explain the observed proliferation, we: 1) tested for an enrichment in Gene Ontology (GO) terms among genes with a greater ω in candidate lineages; and 2). investigated ω in genes of the piRNA pathway specifically.

Our GO enrichment analysis was run independently for each lineage of interest with three different ontologies: Biological Process (BP), Cellular Component (CC) and Molecular Function (MF). Genes with a greater overall ω in the lineages of interest than in the rest of the tree (one-ratio (M0) vs two-ratio test), were enriched for 15 GO terms (supplementary table 3.S2). Only two terms were shared between different lineages : "insulin receptor signaling pathway" and "positive regulation of growth" (shared between lineages B and C). None of the significant GO terms recall the defense against TEs. For all the significant genes, ω is smaller than one in both lineages of interest and the tree. ω estimates are always greater in lineages of interest suggesting a relaxation of selection.

Genes with greater ω in the lineages of interest for some sites only (M2 rel vs CmC test), were enriched in 21 GO terms (supplementary table 3.S3). Four terms are related to chromatin: "chromatin", "chromatin binding", "chromatin assembly or disassembly" and "regulation of chromatin organization". The latter is significant for all of the lineages considered. "Chromatin" and "chromatin assembly or disassembly" are significant for the lineages B and C. There is six significant genes in these chromatin related categories: *Bre1*, *Ctr9*, *dre4*, *ebi*, *polybromo*, *XNP*. Sites with variable rates of evolution have low ω values in the phylogenetic tree, suggesting evolution under purifying selection. However, in at least one lineage of interest a high ω is found, indicating strong positive selection (e.g. ω = 5.43 for *polybromo*).

In our dataset of 2283 orthologous genes, we retrieved only three genes of the piRNA pathway (out of 33): *mino*, *papi* and *tapas*. Considering overall ω values (one-ratio (M0) vs two-ratio test),

our results indicate no difference of evolutionary rate in the lineages of interest for these genes (supplementary table 3.S4). However, if we relax the hypothesis that all sites evolve at the same rate (M2 rel vs CmC test), some sites have different rates of evolution for *papi* and *tapas*. *papi* is a gene involved in negative regulation of transposon integration (noa, b). *tapas* is implied in negative regulation of transposition and piRNA biosynthetic process (noa, a). Regarding *papi*, some sites evolved under strong positive selection in lineage A, *D. suzukii-D. supbulchrella* ancestor ($\omega_{Lineage\mathrm{A}}/\omega_{Tree}$=2.10/0.40). On the contrary, in lineages B (*D. suzukii*, *D. supbulchrella*, and their ancestor), and C (*D. suzukii* and *D. supbulchrella*), we obtained a signal of strong purifying selection ($\omega_{Lineage\mathrm{B}}/\omega_{Tree}$=0.15/0.45 ; $\omega_{Lineage\mathrm{C}}/\omega_{Tree}$=1.0x10^3/0.48). These results suggest an event of quick positive selection and purifying selection afterward. Concerning *tapas*, some sites were under stronger purifying selection in lineages B and C as compare to the rest of the tree ($\omega_{Lineage\mathrm{B}}/\omega_{Tree}$=1.0x10^3/0.49 ; $\omega_{Lineage\mathrm{C}}/\omega_{Tree}$=1.0x10^3/0.48) but not in lineage A. It indicates purifying selection after divergence between *D. suzukii* and *D. supbulchrella*. Overall, on this small set of genes of the piRNA pathway, we observed a stronger positive selection and/or a stronger purifying selection in *suzukii* subgroup, but no signal of a relaxation of selection.

3.5 Discussion

Despite TEs being powerful mutators, with potentially large effects on their hosts, their dynamics in genomes remains unclear. In this paper we used sequencing data from *D. suzukii* and seven close species to portray the TE proliferation behind the high TE content in *D. suzukii* genome.

3.5.1 Spatial heterogeneity in the TE induced genome expansion

The study of TE genomic distribution evolution has long been limited by the difficulty of comparing genomes between closely related species. The recent development of long-read sequencing technologies, allowing to generate near-chromosomal assemblies even for non-model species, pushed this limit further. In our study, we found both *D. suzukii* and *D. biarmipes* long-reads based assemblies to be divided between mutually exclusive gene rich/TE poor and gene poor/TE rich regions (see also MÉREL et collab. (2020b)). It is thus likely that this spatial structure already existed in the common ancestor of the two species, and so before the TE proliferation that occured in *D. suzukii* lineage after its divergence with *D. biarmipes*. We showed that gene rich and gene poor regions were not affected in the same way by the TE proliferation. While the size of gene rich regions in *D. suzukii* is only 1.23 times larger than homologous regions in *D. biarmipes*, gene poor regions are nearly three times larger (x 2.84 in average).

Recently the comparison of long-reads genomic assemblies for three species of the *D. simulans* complex (CHAKRABORTY et collab., 2020), also demonstrated a relative conservation of euchromatic gene rich regions, and that variations of TE content were mostly restricted to heterochromatic gene poor regions. It suggests that the accumulation of TEs may be constrained in gene rich regions for the whole *melanogaster* group, to whom the *D. simulans* complex and the *suzukii* subgroup belong. A conservation of gene rich regions sizes is expected because a strong purifying selection is acting against TEs in gene vicinity (MÉREL et collab., 2020a). Moreover, purifying selection against

ectopic recombination may also contribute to prevent the accumulation of TEs in in gene rich regions, as in *D. melanogaster* recombination rate is higher in these regions (ADAMS et collab., 2000). The investigation of the recently produced long-reads based assemblies of 101 drosophilid species should help testing to which extent gene rich regions are conserved in Drosophila genus (KIM et collab., 2020).

3.5.2 The study of dN/dS did not reveal any relaxation of selection

Our analysis of the ratio of non-synonymous to synonymous substitution rates (dN/dS or ω) does not support the hypothesis of a genome-wide relaxation of selection to explain the high TE content in *D. suzukii* and *D. subpulchrella* genomes. While there may be other explanations to the observed TE accumulation, it is important to remember that some authors call for caution regarding the use of dN/dS to study evolutionary differences between lineages (WEBER et collab., 2014). In particular, dN/dS may be affected by: 1) fluctuations in the frequency of codons (GUÉGUEN et DURET, 2018), 2) too short divergence times (MUGAL et collab., 2014; WEBER et collab., 2014), 3) saturation at the third position of each codon. However, one may argue that these biases are negligible here. First, because in Drosophila codon usage has been found to be relatively stable (VICARIO et collab., 2007), one may expect a small impact of fluctuations in codon frequency on our results. Secondly, with respect to the divergence time effect, it is difficult to assess how much this factor might have affected our study. Indeed, both statistical and biological biases come into play when the divergence times are too short (WEBER et collab., 2014). However, it seems that the major bias, i.e. taking into account polymorphism in the calculation of the dN/dS, fades after about 10 N_e generations (MUGAL et collab., 2014). In our study, the time of divergence between our species of interest (*D. biarmipes*, *D. subpulchrella* and *D. suzukii*), is comprised between 30 and 60 N_e generations. Thirdly, concerning a possible saturation at the third position of each codon, it is reasonable to assume that it is negligible at the time scale of our study [1].

Our results indicate that the observed TE proliferation did not result from a relaxation of selection on genes acting against TE activity. We found signatures of strong positive selection/purifying selection for genes of the piRNA pathway. We also detected signatures of strong positive selection for genes related to chromatin state. These observations suggest that the TE proliferation drove modifications in evolutionary pressures acting toward these genes, rather than the opposite. Indeed, it is plausible that the accumulation of TEs in gene poor regions selected for a stronger compaction of chromatin in these regions, and thus perhaps a reduced expression of TEs. It is also likely that the strong TE proliferation selected for a lower TE activity and drove positive selection/purifying selection on genes of the piRNA pathway. However, it is important to notice that we worked on a reduced set of genes (2,283 against ca. 15,500 in *D. melanogaster*), with only three genes of the piRNA pathway out of the 33 listed in OZATA et collab. (2019). In the future improved assemblies should allow to work on a larger set of genes, and to assess the generalizability of our results.

[1] Here we are going to verify that dS values low to ensure that saturation is indeed negligible in our dataset, and replace this sentence by a more appropriate one

3.5.3 Alternatives to the relaxation of selection hypothesis

We would like to discuss three alternatives to the relaxation of selection hypothesis. The first alternative would imply horizontal transfer of TEs. According to the trap model of TE dynamic, few generations are needed for an invading TE to be inactivated by insertion(s) in piRNA cluster(s) (KOFLER et collab., 2018; ZANNI et collab., 2013). A recurrent acquisition of TEs by horizontal transfer over the last four million years could therefore explain their abundance in *D. suzukii* and *D. subpulchrella* genomes. The main flaw in this hypothesis, is the difficulty in giving a straightforward explanation for the necessary high rate of horizontal transfer. Castillo and collaborators proposed that, in Drosophila, a high rate of TE transfer could result from very large population size (CASTILLO et collab., 2011). They suggest that population TE exposure rates scale with the population size. However, if *D. suzukii-D. subpulchrella* lineage was characterized by large population size, one may expect an impact on ω. In our study, we found no evidence that ω was particular in this lineage.

Another hypothesis, that could be put forward to explain the observed TE proliferation, is a reduction in the relative size of piRNA clusters (KOFLER, 2020). A simulation work suggests that the cumulative size of piRNA clusters must exceed 0.2% of the genome to effectively stop the proliferation of TEs. One could thus propose that a reduction in piRNA cluster size in *D. suzukii* and *D. biarmipes* ancestor triggered the observed TE proliferation. The sequencing of *D. biarmipes* and *D. suzukii* piRNAs would make it possible to locate piRNA clusters in these two species and to test this hypothesis. If this hypothesis is correct, the cumulative size of piRNA clusters should be smaller in *D. suzukii*.

A last hypothesis is the evolution of greater insertion rates and/or insertion preferences in *D. suzukii-D. subpulchrella* lineage since its divergence with *D. biarmipes*. For instance, some TEs could have evolved mechanisms to insert preferentially in gene poor regions, i.e. where purifying selection is weak, or outside of piRNA clusters. This would have progressively expanded parts of the genome favorable toward TEs, and facilitated their accumulation. It would be consistent with the gradual increase of TE accumulation rate observed. TE evolution, and the necessary differences in evolutionary rate between TEs, could also explain discrepancies in accumulation rate. For example, it may explain the intense proliferation of *Gypsy* as compared to others elements. Because of their important size, these elements are more likely to promote ectopic recombination and thus to be efficiently removed by selection (PETROV et collab., 2003). Moreover, in both *D. melanogaster* and *D. simulans*, LTR elements such as *Gypsy*, have been found to be more heavily targeted by repressive heterochromatin marks, and more intensely selected against, potentially because of the effect of repressive heterochromatic marks on nearby gene expression (LEE et KARPEN, 2017). Considering the last two points, the intense proliferation of *Gypsy* is unexpected, except if it achieved greater insertion rates or insertion preference toward genomic regions where selection is weak. The alignment of TE sequences and an analysis of evolutionary rates could tell us if we should investigate further the hypothesis of TE evolution or dismiss it. If such hypothesis is true, one would except to find a stronger signal of positive selection on TE sequences in branches of the phylogenetic tree were the TE accumulation is intense. Additionaly, we could expect a correlation between the strength of positive selection and the accumulation rate among TE superfamilies.

3.5.4 Conclusion

Our study illustrates the complex dynamics behind the high percentage of TEs in *D. suzukii* genome. In particular, we showed that it results from a long-term accumulation of TEs. This accumulation started about four million years ago, approximately the date of divergence between *D. biarmipes* and *D. suzukii*-*D. subpulchrella* ancestor. Our results highlight the spatial heterogeneity in TE accumulation, and show that gene rich regions are globally spared. It is very likely to be due to intra-genomic variations in the efficacy of purifying selection, and emphasizes the necessity to consider genomic heterogeneity when studying TE dynamics. Investigating dN/dS, we did not find any evidence that the observed TE proliferation results from a genome-wide relaxation of selection. On the opposite it may have triggered positive selection and/or stronger purifying selection on genes involved in chromatin conformation and genes of the piRNA pathway. Two appealing alternatives to the relaxation of selection hypothesis would be: 1) a reduction of the genomic part occupied by piRNA clusters, and 2) an evolution of greater insertion rate/insertion preferences by TEs. An investigation of piRNA cluster size and TE evolutionary rate among species of *suzukii* subgroup may provide insights in the forces driving TE accumulation at the interspecific scale.

3.6 Material Methods

Genomic assemblies

In order to extract coding sequences and study phylogenetic relationships among *D. suzukii* and its close relatives, 6 public assemblies were retrieved (*D. biarmipes*-GCA_000233415.2, *D. elegans*-GCA_000224195, *D. eugracilis*-GCA_000236325.2, *D. ficusphila*-GCA_000220665.2, *D. suzukii*-GCA_-000472105.1, *D. takahashii*-GCA_000224235.2), and 2 assemblies were built (*D. pseudotakashii* and *D. subpulchrella*) from private paired-end reads libraries using IDBA (Peng et collab., 2012).

Phylogenetic reconstruction

Obtained assemblies were masked using RepeatMasker (–species drosophila) (http://www.repeatmasker.org/). 2683 single copy genes were identified in masked assemblies using BUSCO (v3.0.2, -m genome -sp fly) (Waterhouse et collab., 2018). The lineage dataset used was diptera_odb9 ($N_{species}$=25, N_{BUSCOs}=2799). Multiple sequence alignments were performed using MACSE for the 2,283 single copy genes present in all the 8 masked assemblies. (v2.03, -prog alignSequences -optim 1) (Ranwez et collab., 2018). Ambiguously aligned regions were removed from the alignments using Gblocks (v091b, -b1=7 -b2=9 -b3=10 -b4=2 -b5=h) (Talavera et Castresana, 2007). Maximum-likelihood phylogenies were inferred under the LG+ G4 partitioned model using IQ-Tree (v1.6.2 -s concat -spp partition -nt AUTO -bb 1000 -alrt 1000 -bnni -wbt) (Nguyen et collab., 2015). Divergence time between *D. suzukii* and *D. subpulchrella*, as well as between *D. suzukii* and *D. biarmipes*, were estimated using BEAST (v1.10.4, Yang96 model; strict clock; Tree prior=Speciation: Yule Process; length of MCMC chain 5000000). The substitution rate was set to 0.011 as estimated on 176 nuclear genes in Drosophila (Tamura et collab., 2004). A consensus tree was obtained using the module

treeannotator. Confidence intervals were computed as upper and lower values within the 95% highest probability density.

Reads processing

Genome size and TE content of *D. suzukii* and close species were investigated using paired-end reads. In addition to the two private paired-end reads datasets mentioned above (*D. pseudotakahashii, D. subpulchrella*), the following accessions were used: SRR345536, SRR345540, SRR345543, SRR345541, SRR942805, SRR345539 (*D. biarmipes, D. elegans, D. eugracilis, D. ficusphila, D. suzukii, D.takahashii*). First, low quality positions were removed from reads using UrQt (v1.0.17, –pos both –t 20) (MODOLO et LERAT, 2015). Then, read size was unify across samples to 100 bp (FASTX-Toolkit v0.0.13, http://hannonlab.cshl.edu/fastx_toolkit/). Finally, putative mitochondrial reads, i.e. reads mapping on various mitochondrial genomes of Drosophila, were filtered out (fastq screen v0.14.0; –aligner bwa, see SESSEGOLO et collab. (2016) for the mitochondrial assemblies used) (WINGETT et ANDREWS, 2018).

Genome size estimates

Genome sizes were estimated using reads kmers occurrences distribution. Kmers occurrences were counted using Jellyfish count module (v2.3.0; -C -m 21) (MARÇAIS et KINGSFORD, 2011). Histogram were retrieved with the histo module (-h 3000000). findGSE provided genome size estimates using the obtained distribution (v0.1.0) (SUN et collab., 2018).

TE database

A TE database was created by merging previously established consensus of Drosophila TE families, and consensus reconstructed de novo from high-quality genome assemblies of *D. biarmipes*, *D. eugracilis* and *D. suzukii* (MILLER et collab., 2018; MÉREL et collab., 2020b; PARIS et collab., 2020). The previously established consensus were obtained by extracting all Drosophila consensus annotated as DNA, LINE, LTR, Other, RC, SINE and Unknown from Dfam and Repbase databases (release 2016-2018 for both). Full LTR retroelements sequences were reconstructed by merging LTR and their internal part. *D. suzukii* sequences come from MÉREL et collab. (2020b). *De novo* reconstruction was performed for *D. biarmipes* and *D. eugracilis* as in MÉREL et collab. (2020b) (please refer to this publication for details). Briefly, REPET was used in combination with LTRHarvest and RepeatScout to detect repeated sequences and reconstruct consensus (ELLINGHAUS et collab., 2008; FLUTRE et collab., 2011; PRICE et collab., 2005). The latter were further annotated by homology to Drosophila known TEs using RepeatMasker (http://www.repeatmasker.org/). Finally, previously established consensus and *de novo* reconstructed consensus, were clustered in families using UClust (-id 0.80 -strand both –maxaccepts 0 –maxrejects 0; v11.0.667) (EDGAR, 2010). The longest sequence only is kept, leading to a database with one consensus by TE family.

TE content

TE content was evaluated from paired-end reads using two independent methods. Both methods estimate TE content as the proportion of reads mapping to TE sequences. The first one reconstructs TE sequences from the assembly of a subsample of reads (such as coverage in the subsample <1X) (GOUBERT et collab., 2015), whereas the second method uses a previously established database (WEILGUNY et KOFLER, 2019). The second method allows to quantify TE abundance at the family level. The first method, implemented in dnaPipeTE, was used on a single file containing merged paired-end reads, using genome size estimates of findGSE, and using the above mentioned TE database for annotation (v1.3; -genome_coverage 0.15 -sample_number 2). Non TE, or ambiguous, categories, i.e. Low_Complexity, na, others, rRNA, Satellite, Simple_repeat and Tandem_repeats, were removed from the results. Note that the MITE category was empty. The second method, implemented in deviaTE, was used on a 10X subsample of merged paired-end reads (v0.3.7; –rpm).

Long-read based assemblies annotation

For *D. biarmipes* long-read based assembly (MILLER et collab., 2018), TE content, gene content, and syntenic relationships with *D. melanogaster* assembly, were investigated as in MÉREL et collab. (2020b) (please refer to this publication for details). Briefly, TE fragments were recovered using RepeatMasker with the TE database mentionned above (http://www.repeatmasker.org/). TE density was computed as the number of TE fragments completely within non overlapping genomic windows of 200 Kb. From a run of augustus gene density was evaluated as the number of genes completely within non overlapping genomic windows of 200 Kb (STANKE et collab., 2008). Syntenic relationships between *D. biarmipes* and the chromosomes of *D. melanogaster* were derived from a pipeline based on Promer generated alignments of the two masked assemblies (KURTZ et collab., 2004). *D. melanogaster* masked assembly was downloaded from UCSC Genome Browser (**?**). Data for *D. suzukii* come from MÉREL et collab. (2020b). A graphical visualization of the results was produced using Circos (KRZYWINSKI et collab., 2009).

TE landscapes

To study TE ages in *D. biarmipes* and *D. suzukii* genomes the so called TE landscapes were investigated. For each TE fragment recovered using RepeatMasker (http://www.repeatmasker.org/), the divergence to the family consensus was computed using a script from KAPUSTA et SUH (2017) with RepeatMasker output .align file. The percentage of divergence to the consensus sequence is considered a proxy for age: old TEs have accumulated mutations, young TEs are similar to consensus sequences. The age of insertions was inferred from estimated divergence using the Drosophila substitution rate mentioned above, i.e. 0.011 (TAMURA et collab., 2004).

Evolutionary rate analysis

In order to test for a relaxation of selection, we investigated the dN/dS ratio, or ω, for 2283 genes independently. We expect a relaxation of selection to drive a higher fixation rate of weakly

deleterious mutations, and thus a higher non-synonymous substitution rate and ω. We looked for a relaxation of selection in: **A.** *D. suzukii* and *D. subpulchrella* ancestor solely; **B.** this ancestor as well as in *D. suzukii* and *D. subpulchrella*; **C.** *D. suzukii* and *D. subpulchrella*. We used codeml as implemented in the ete3 package (v3.1.1) (HUERTA-CEPAS et collab., 2016; YANG, 2007). More precisely, two likelihood ratio tests between codeml models were conducted for each gene. The first likelihood ratio test (one-ratio (M0) vs two-ratio test), assesses if the overall ω is different between a lineage of interest (here A, B or C) and the rest of the tree. This test aims at detecting differences on the overall signature of selection. It compares a model in which all the branches evolve at the same rate (One-ratio or M0 model), with a model in which branches of interest may evolve differently (two-ratio model). The second likelihood ratio test , assesses if, for a given and estimated proportion of sites, ω is different between a lineage of interest (here A, B or C) and the rest of the tree. It thus relaxes the hypothesis that all sites evolve similarly. It compares a "site model" (M2a rel) with a "clade model" (CmC). These two models have three categories of sites of proportions p_0, p_1, $p_2 = 1 - p_0 - p_1$ and ω values such as $\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 > 0$. For the site model the three sites categories have one proportion and ω value for all the tree. For the clade model ω_2 varies between branches of interest and the rest of the tree. All runs of codeml were performed with CodonFreq set to 4, and estFreq set to 1, in order to use a Muse and Gaut style codon model (MUSE et GAUT, 1994).

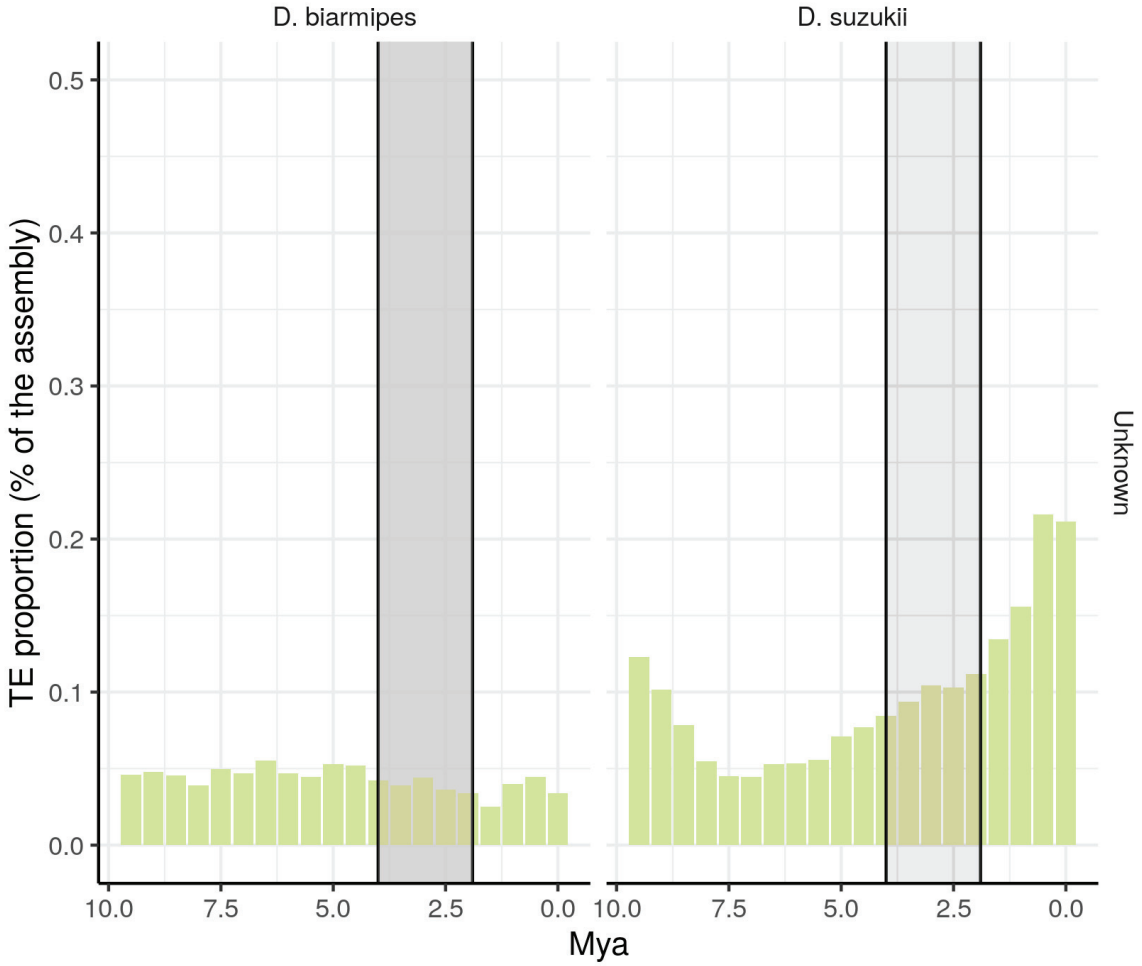
Gene Ontology GO analysis

To assess if genes were differentially impacted according to their function, GO enrichment analyses were performed using topGO (v1.0) (ALEXA et RAHNENFUHRER). As a prerequisite, genes were annotated blating the sequence of the outgroup species, i.e. *D. ficusphila*, against *D. melanogaster* CoDing Sequences (dmel-all-CDS-r6.32.fasta; blat v36x4;-t=dnax -q=dnax). The best hit was retained to annotate the query. Enrichment tests were performed for genes with a greater overall ω (one-ratio (M0) vs two-ratio test), and genes with a greater ω for a given proportion of sites only (M2 rel vs CmC test). These test were performed independently for each lineage and three different ontologies were tested: Biological Process (BP), Cellular Component (CC) and Molecular Function (MF). topGO was used with the following parameters: nodeSize=5, algorithm = "weight01", statistic = "Fisher". Because the piRNA pathway play an important role in the defense against TEs, we paid special attention to corresponding genes (we used the list of genes in OZATA et collab. (2019)).

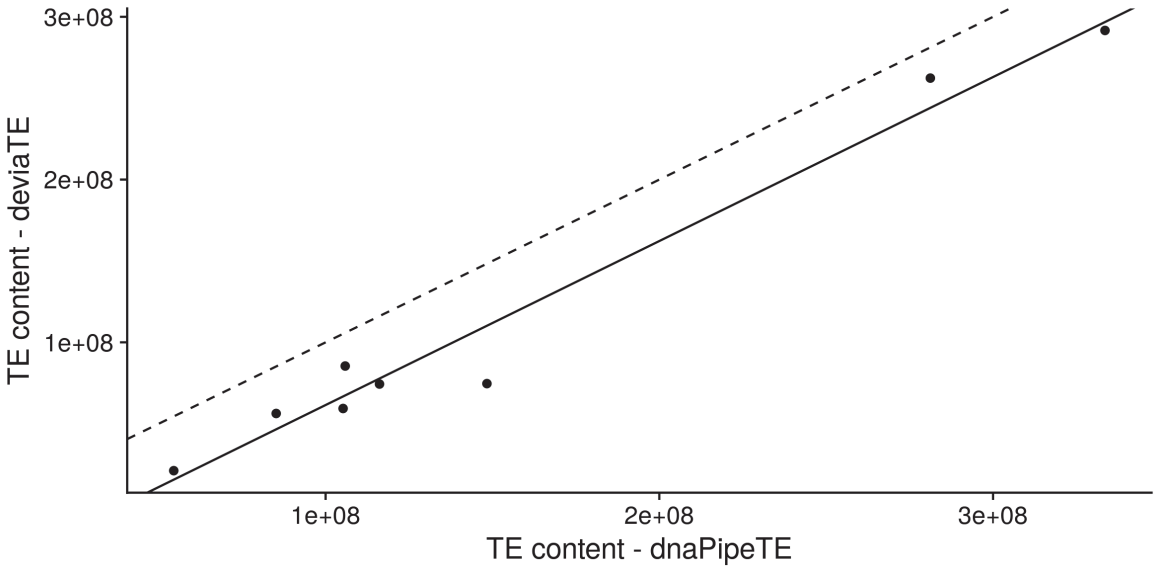
3.7 Acknowledgements

This work was supported by the French National Research Agency (ANR-16-CE02-0015-01 –SWING) and performed using the computing facilities of the CC LBBE/PRABI.

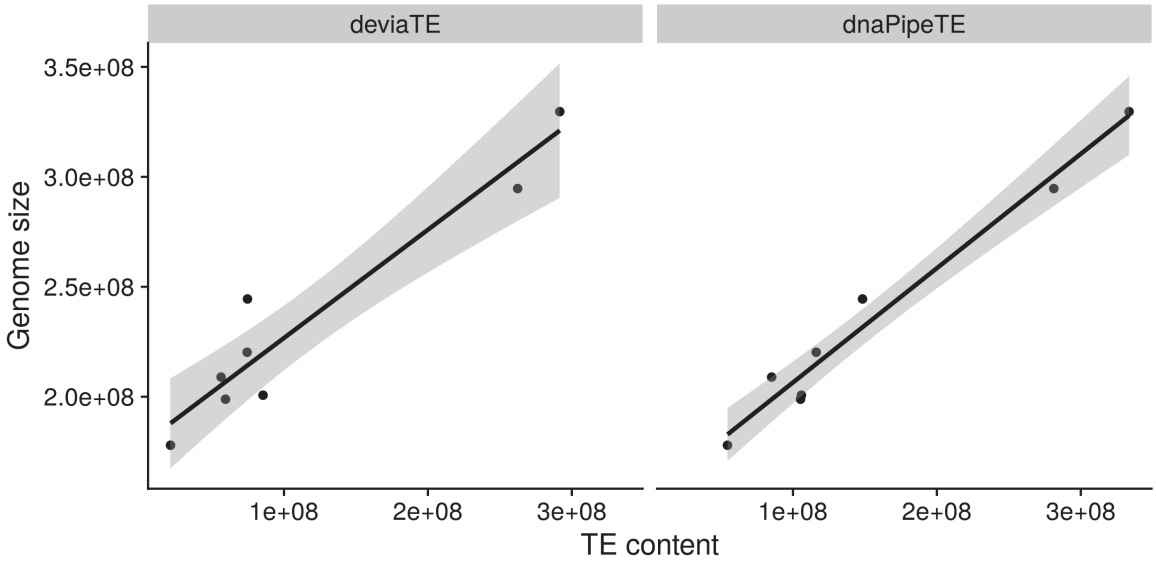
3.8 Supplementary Figures



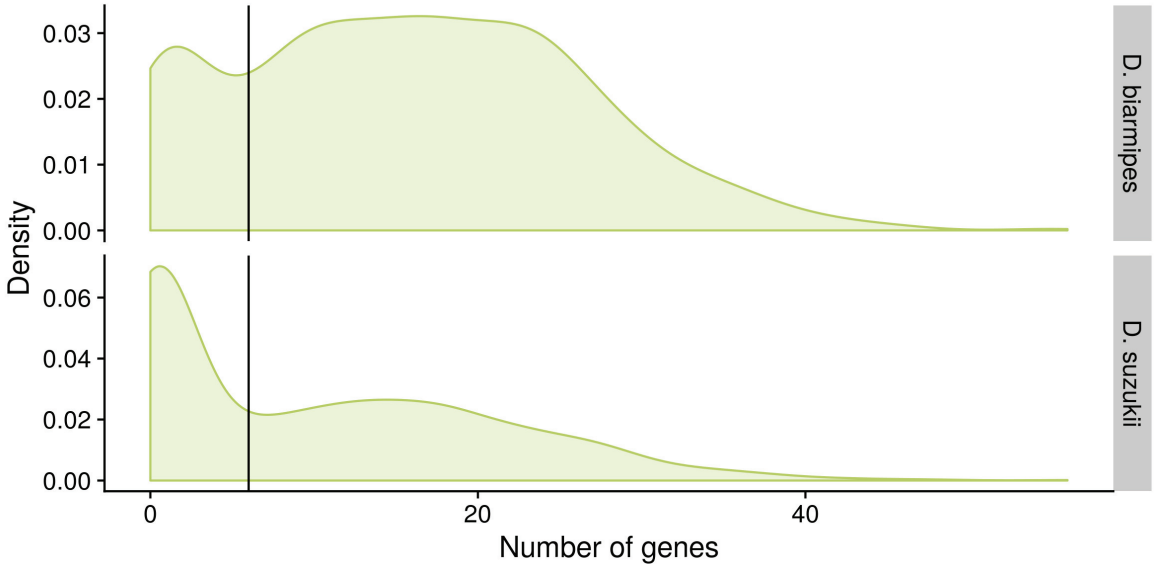
Supplementary figure 3.S1: **TE landscapes for unknown sequences in *D. suzukii* and in *D. biarmipes*.** The percentage of the genome assembly estimated using RepeatMasker is represented as a function of its age.



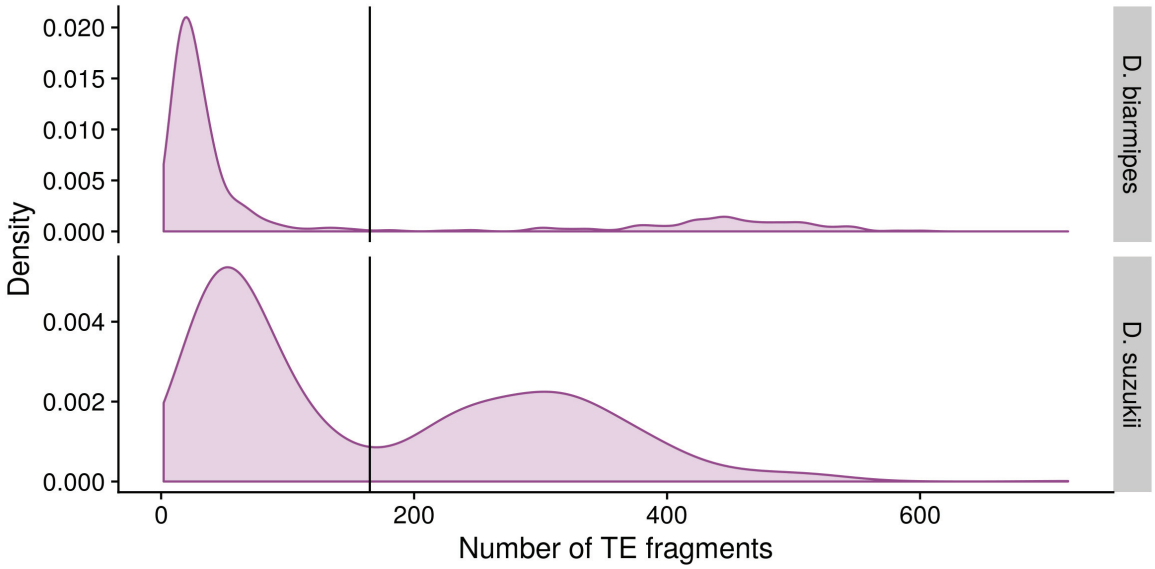
Supplementary figure 3.S2: **Correlation between deviaTE and dnaPipeTE estimates of the overall TE content.** Values on both axis are in bp. The unbroken line corresponds the regression equation (-39269809 + 1.00759x). The dashed line corresponds to y=x.



Supplementary figure 3.S3: **Correlation between genome size and TE content estimates from deviaTE (left) and dnaPipeTE (right).** Values on both axis are in bp. The unbroken line corresponds the regression equation (deviaTE : -3.19e+08 + 1.85x; dnaPipeTE: -2.84e+08 + 1.87x)



Supplementary figure 3.S4: **Distribution of the number of genes per 200 Kb windows in *D. biarmipes* and *D. suzukii* assemblies.** The vertical line corresponds to x=6.



Supplementary figure 3.S5: **Distribution of the number of TE fragments per 200 Kb windows in *D. biarmipes* and *D. suzukii* assemblies.** The vertical line corresponds to x=165.

3.9 Supplementary Tables

D. melanogaster chromosome	Mb of *D. biarmipes* assembly
2L	33.9
2R	34.2
3L	31.5
3R	33.8
4	1.9
X	26.2

Supplementary table 3.S1: **Number of Mb of *D. biarmipes* assembly attributed to each of the *D. melanogaster* chromosomes.**

Term	p	Genes ($\omega_{Lineage}/\omega_{Tree}$)	Ontology	Lineage
ATP binding	0.0076	Mlh1 (0.41/0.08) Pms2 (0.3/0.09)	MF	A
ATPase activity	0.0121	Mlh1 (0.41/0.08) Pms2 (0.3/0.09)	MF	A
double-stranded DNA binding	0.0147	Mlh1 (0.41/0.08) Pms2 (0.3/0.09)	MF	A
DNA repair	0.0080	Mlh1 (0.41/0.08) Pms2 (0.3/0.09)	BP	A
insulin receptor signaling pathway	0.0310	lin-28 (0.73/0.14) raptor (0.13/0.08)	BP	B
positive regulation of growth	0.0450	14-3-3epsilon (0.25/0.02) raptor (0.13/0.08)	BP	B
organelle assembly	0.0450	B9d1 (0.53/0.18) Rab7 (0.18/0.02)	BP	B
DNA-binding transcription factor activity, RNA polymerase II-specific	0.0300	Ets96B (0.25/0.06) ovo (0.12/0.04)	MF	C
oogenesis	0.0053	14-3-3epsilon (0.29/0.02) lin-28 (0.73/0.14) ovo (0.12/0.04) Rab7 (0.31/0.02) raptor (0.13/0.08)	BP	C
regionalization	0.0163	14-3-3epsilon (0.29/0.02) ovo (0.12/0.04) Rab7 (0.31/0.02)	BP	C
insulin receptor signaling pathway	0.0312	lin-28 (0.73/0.14) raptor (0.13/0.08)	BP	C
animal organ development	0.0328	14-3-3epsilon (0.29/0.02) asrij (0.24/0.07) Ctr9 (0.04/0.01) ovo (0.12/0.04) Rab7 (0.31/0.02) raptor (0.13/0.08)	BP	C
positive regulation of growth	0.0452	14-3-3epsilon (0.29/0.02) raptor (0.13/0.08)	BP	C
regulation of cell morphogenesis	0.0452	ovo (0.12/0.04) raptor (0.13/0.08)	BP	C

Supplementary table 3.S2: **Gene-ontology enrichment analysis for genes with a greater overall ω in the lineages of interest than in the rest of the tree (one-ratio (M0) vs two-ratio test).** Significant GO terms are indicated together with the corresponding p-value. For each term, significant genes are indicated with the estimated ω ($\omega_{Lineage}/\omega_{Tree}$). Three different ontologies were considered: Biological Process (BP), Cellular Component (CC) and Molecular Function (MF). Studied lineages correspond to **A.** *D. suzukii* and *D. subpulchrella* ancestor, **B.** *D. suzukii, D. subpulchrella* and their ancestor, **C.** *D. suzukii* and *D. subpulchrella*

Term	p	Genes ($\omega_{Lineage}/\omega_{Tree}$)	Ontology	Studied
DNA repair	0.0200	Mlh1 (3.63/0.29) mus81 (0.9/0.33) Pms2 (2.08/0.34)	BP	A
regulation of protein localization	0.0450	papi (2.1/0.4) Ranbp21 (0.62/0.34)	BP	A
regulation of protein organization	0.0450	dre4 (3.27/0.32) ebi (1.84/0.42)	BP	A
polytene chromosome	0.0073	dre4 (0.58/0.36) ebi (0.55/0.42) polybromo (5.43/0.07) XNP (3.64/0.08)	CC	B
chromatin	0.0448	Bre1 (1.09/0.39) Ctr9 (1.42/0) polybromo (5.43/0.07) XNP (3.64/0.08)	CC	B
chromatin binding	0.0120	dre4 (0.58/0.36) ebi (0.55/0.42) polybromo (5.43/0.07)	MF	B
ATPase activity	0.0430	Mlh1 (1.8/0.3) polybromo (5.43/0.07) XNP (3.64/0.08)	MF	B
regulation of protein organization	0.0004	Ctr9 (1.42/0) dre4 (0.58/0.36) ebi (0.55/0.42) polybromo (5.43/0.07)	BP	B
chromatin assembly or disassembly	0.0085	dre4 (0.58/0.36) polybromo (5.43/0.07) XNP (3.64/0.08)	BP	B
post-embryonic appendage morphogenesis	0.0098	ebi (0.55/0.42) ovo (8.55/0.09) polybromo (5.43/0.07)	BP	B
imaginal disc-derived appendage morphogenesis	0.0098	ebi (0.55/0.42) ovo (8.55/0.09) polybromo (5.43/0.07)	BP	B
histone modification	0.0260	Bre1 (1.09/0.39) Ctr9 (1.42/0) ebi (0.55/0.42)	BP	B
nucleosome organization	0.0260	dre4 (0.58/0.36) polybromo (5.43/0.07) XNP (3.64/0.08)	BP	B

regulation of intracellular signal transduction	0.0287	asrij (2.35/0.14) ebi (0.55/0.42) XNP (3.64/0.08)	BP	B
chromatin	0.0190	Bre1 (0.87/0.43) Ctr9 (1.96/0.01) polybromo (7.46/0.06) XNP (7.54/0.08)	CC	C
organelle part	0.0470	B9d1 (2.34/0.18) Bre1 (0.87/0.43) Ctr9 (1.96/0.01) Dis3 (4.96/0.05) polybromo (7.46/0.06) Psf1 (0.35/0.32) thoc6 (2.24/0.08) XNP (7.54/0.08)	CC	C
3'-5' exonuclease activity	0.0350	Dis3 (4.96/0.05) Not1 (4.11/0.72)	MF	C
catalytic activity, acting on DNA	0.0350	polybromo (7.46/0.06) Psf1 (0.35/0.32)	MF	C
gene silencing	0.0380	polybromo (7.46/0.06) XNP (7.54/0.08)	BP	C
chromatin assembly or disassembly	0.0380	polybromo (7.46/0.06) XNP (7.54/0.08)	BP	C
regulation of chromatin organization	0.0380	Ctr9 (1.96/0.01) polybromo (7.46/0.06)	BP	C

Supplementary table 3.S3: **Gene-ontology enrichment analysis for genes with a proportion of sites with a greater ω in the lineages of interest than in the rest of the tree (M2 rel vs CmC test).** Significant GO terms are indicated together with the corresponding p-value. For each term, significant genes are indicated with the estimated ω ($\omega_{Lineage}/\omega_{Tree}$). Three different ontologies were considered: Biological Process (BP), Cellular Component (CC) and Molecular Function (MF). Studied lineages correspond to **A.** *D. suzukii* and *D. subpulchrella* ancestor; **B.** *D. suzukii, D. subpulchrella* and their ancestor; **C.** *D. suzukii* and *D. subpulchrella*

Test	Interpretation	Lineage	Gene	p_{adj}	ω Lineage	ω Tree
One-ratio (M0) vs two-ratio	Difference in the overall signature of selection	**A**	mino	0.58	0.032	0.070
			papi	0.23	0.022	0.067
			tapas	0.84	0.13	0.16
		B	mino	0.83	0.058	0.069
			papi	0.62	0.052	0.072
			tapas	0.076	0.10	0.16
		C	mino	0.97	0.071	0.068
			papi	0.10	0.026	0.072
			tapas	0.067	0.091	0.16
M2 rel vs CmC	Divergent evolution of some sites	**A**	mino	0.72	0.00037	0.41
			papi	**0.0021**	2.10	0.40
			tapas	1.00	0.0001	0.0001
		B	mino	1.00	0.38	0.40
			papi	**0.0028**	0.15	0.45
			tapas	**0.0048**	0.0001	0.49
		C	mino	0.80	0.66	0.33
			papi	**0.00093**	0.00010	0.48
			tapas	**0.0052**	0.0001	0.48

Supplementary table 3.S4: **Evolution of ω for genes of the piRNA pathway in *suzukii* subgroup.** Two likelihood-ratio tests contrasting null and alternative models from codeml were used to assess variations of selection in lineages of *suzukii* subgroup (YANG, 2007). The first one, one-ratio (M0) vs two-ratio, assumes that all sites evolve similarly, and detects differences in overall ω. The second one, M2 rel vs CmC, relaxes the hypothesis of unvarying evolution along the gene, and identifies divergent evolution at a given, and estimated, proportion of sites. Each test was performed focusing on three different lineages: **A.** *D. suzukii* and *D. subpulchrella* ancestor; **B.** *D. suzukii*, *D. subpulchrella* and their ancestor; **C.** *D. suzukii* and *D. subpulchrella*.

4 Discussion générale

Sommaire

L'objectif de ma thèse était l'étude de la dynamique des ET chez l'espèce non modèle *Drosophila suzukii*. Une étude bibliographique a permis dans un premier temps de brosser un portrait général des connaissances sur les ET chez le genre Drosophila. Le contenu en ET chez *D. suzukii* ayant été relativement peu étudié, son analyse a fait l'objet d'un travail préliminaire. Ce travail a révélé une très importante proportion d'ET dans le génome (~47 %), ainsi qu'une répartition non-homogène et une vraisemblable activité de ces ET. L'étude de 22 populations de *D. suzukii* a quant à elle suggéré un rôle important des variations de taille efficace (N_e) dans la dynamique des ET: de petits N_e sont associés à un contenu en ET important. L'analyse du polymorphisme d'insertions dans ces populations a permis de révéler 15 insertions potentiellement adaptatives. Enfin, l'analyse de données de séquençage pour *D. suzukii* et 7 espèces proches, suggère que, indépendamment de ce qui a été observé dans les populations, *D. suzukii* a vu les ET proliférer dans son génome au cours des 4 derniers millions d'années. Cette prolifération s'est faite de manière localisée, et l'étude du dN/dS indique une absence de relâchement de la sélection à l'échelle du génome. De manière générale, ce travail fourni une analyse détaillée des ET chez *D. suzukii* et de leur dynamique.

4.1 Contenu en ET et intensité de la sélection

Les travaux réalisés au cours de cette thèse suggèrent deux épisodes d'augmentation du contenu en ET dans le passé récent de *D. suzukii*: une accumulation progressive au cours des quatre derniers millions d'années, et une plus intense lors de l'invasion des continents Européen et Américain dans les années 2000. Les taux d'accumulation observés dans ces deux cas présentent des ordres de grandeurs différents, et les mécanismes sous-jacents sont potentiellement différents. En effet, nos résultats suggèrent un rôle de la réduction de l'efficacité de la sélection sur l'augmentation du contenu en ET au cours du processus invasif uniquement. Ceci pourrait toutefois aussi être expliqué par l'existence de certains biais méthodologiques.

4.1.1 Une accumulation plus rapide lors de l'invasion

L'étude, dans le génome séquencé de *D. suzukii*, de la divergence des séquences des copies d'ET à leur consensus (le « *TE landscape* »), indique une accumulation de 88 Mb d'ET au cours des 4 derniers millions d'années, soit environ 22 pb par an. En effet, 88 Mb de l'assemblage génomique est occupé par des copies d'ET de séquences proches de leur consensus suggérant une faible accumulation de mutation depuis leurs insertions.

L'étude du polymorphisme d'insertion chez 22 populations naturelles, quant à elle suggère une accumulation de 642 insertions par génome haploïde au cours du processus d'invasion. Si on considère un début d'invasion en 2008 et un échantillonnage entre 2013 et 2016, ceci correspond à environ 160 insertions par an, ou 160 000 pb par an pour une taille moyenne d'insertion de 1 000 pb.

Même en prenant en compte que le chiffre de 22 pb par an pour l'accumulation au niveau interspécifique pourrait-être légèrement sous estimé du fait du caractère partiel de l'assemblage [1], le taux d'accumulation observé lors de l'invasion est manifestement plus important. Ceci est

[1] D'après les estimations de taille de génome en cytométrie en flux il manque environ 40-70 Mb à l'assemblage

d'autant plus vrai que le nombre d'insertions accumulées lors du processus d'invasion est vraisemblablement sous-estimé, du fait notamment de la difficulté de détecter toutes les insertions dans les régions du génome riches en ET.

4.1.2 dN/dS: un bon proxy de la taille efficace ?

Nos travaux de simulations montrent que, du fait du caractère récent de l'invasion, $\widehat{\theta_W}$ n'était pas nécessairement égal à $4N_e\mu$ dans les populations de *D. suzukii* à la date de leur échantillonnage (voir Annexe 4)[1]. Toutefois ils montrent aussi que la diminution de $\widehat{\theta_W}$ dans les populations invasives reflète vraisemblablement bien une réduction transitoire de la taille des populations lors de l'invasion. Ainsi $\widehat{\theta_W}$ semble un bon proxy des variations récentes de taille efficace dans les populations de *D. suzukii*. La corrélation observée entre $\widehat{\theta_W}$ et le nombre d'insertions par génome haploïde suggère donc un impact de la taille efficace sur le contenu en ET.

Les résultats du Chapitre 3 en revanche indiquent que l'augmentation du contenu en ET observée au cours des derniers millions d'années n'est pas nécessairement liée à une réduction de la taille efficace. En effet, elle n'est pas associée à une élévation du dN/dS à l'échelle du génome. Or dans cette thèse, on suppose que, toutes choses étant égales par ailleurs, le dN/dS doit-être inférieur dans les populations avec une taille efficace importante du fait d'une sélection purificatrice plus forte.

Contrairement à celle de $\widehat{\theta_W}$, l'utilisation du dN/dS comme proxy de la taille efficace n'a pas été validée par simulation. Certains auteurs invitent à la prudence quant à l'utilisation du dN/dS pour l'étude des différences d'évolution entre lignées (Weber et collab., 2014). Notamment, le dN/dS peut-être affecté par 1) des fluctuations de la fréquence des codons (Guéguen et Duret, 2018), 2) des temps de divergence trop courts (Mugal et collab., 2014; Weber et collab., 2014), 3) une saturation à la troisième position de chaque codon. Chez les oiseaux Weber et collab. (2014) n'ont pas trouvé d'effet clair de ces facteurs sur le dN/dS. Si on ne peut pas affirmer de la même manière que ces facteurs n'aient pas affectés nos résultats, on peut en discuter l'éventualité. Premièrement, des travaux indiquent que l'usage du code est relativement stable au sein du genre Drosophila (Vicario et collab., 2007), ceci suggère un faible impact des fluctuations de la fréquence des codons sur nos résultats. Deuxièmement, en ce qui concerne un effet du temps de divergence, il est difficile d'évaluer dans quelle mesure ce facteur pourrait avoir affecté notre étude. En effet, des biais statistiques aussi bien que biologiques entrent en jeux quand les temps de divergence sont trop courts (Weber et collab., 2014). Il semble toutefois que le biais majeur que constitue la prise en compte du polymorphisme dans le calcul du dN/dS, s'estompe après environ 10 N_e générations. Or dans notre étude, le temps de divergence entre nos espèces d'intérêts (*D. biarmipes*, *D. subpulchrella* et *D. suzukii*), est comprise entre 30 et 60 N_e générations. Troisièment, au sujet d'une éventuelle saturation à la troisième position de chaque codon, s'il est potentiellement envisageable de faire l'hypothèse qu'elle est négligeable à l'échelle de temps dans laquelle se place notre étude, on pourrait aussi envisager d'utiliser le ratio Kr/Kc plutôt que le ratio dN/dS. Le ratio

(Hjelmen et collab., 2019; Sessegolo et collab., 2016)

[1] Ces travaux ont été réalisés en réponse à une interrogation d'un des relecteurs sollicité par le journal auquel l'article presenté Chapitre 2 a été soumis. En effet, ce relecteur a fait état de doutes quand au fait que les variations de $\widehat{\theta_W}$ observées soient le reflet de variations de taille des populations. Ces travaux ont vocation à être intégrés dans la prochaine version du manuscrit

Kr/Kc correspond au ratio substitutions radicales/substitutions conservatives. Les substitutions d'acide aminé radicales, contrairement aux substitutions conservatives, altèrent la polarité ou le volume du résidu et sont donc plus susceptibles d'être négativement sélectionnées. L'utilisation de Kr/Kc permettrait d'outrepasser les phénomènes de saturation.

4.1.3 Et si ce n'était pas la taille efficace ?

Si, à nos yeux, l'augmentation du contenu en ET concomitante au processus d'invasion est très vraisemblablement due à une réduction de la taille efficace, l'accumulation d'ET observée au cours des quatre derniers millions d'années, pourrait elle avoir un autre moteur. L'absence d'augmentation du dN/dS constatée ne serait alors pas nécessairement artefactuelle. On peut émettre au moins trois hypothèses alternatives à la réduction de taille efficace: 1) un fort taux d'acquisition d'ET par transfert horizontal, 2) une réduction relative de la taille des clusters de piRNA et 3) l'évolution d'un plus fort taux d'insertion/de préférences d'insertions. Ces trois hypothèses sont non mutuellement exclusives. La combinaison des hypothèses 2) et 3) notamment, permet l'élaboration d'un scénario plausible d'après les résultats observés. En effet, l'évolution par les ET d'un mécanisme d'évitement des clusters de piRNA aurait pu faciliter leur prolifération, entraîner une réduction de la taille relative des clusters, faciliter encore la prolifération des ET, et ainsi de suite. Un tel scénario présente l'avantage d'expliquer l'accélération progressive du taux d'accumulation d'ET observée. Le séquençage des petits ARN (voir Section 4.4.1), et l'étude des pressions de sélection sur les séquences d'ET (voir Section 4.4.2) pourraient permettre de mieux comprendre les facteurs à l'œuvre ici.

4.2 ET, sélection positive et adaptation

Au-delà de l'étude du contenu en ET, l'étude du polymorphisme d'insertions a permis d'explorer l'existence d'insertions évoluant sous sélection positive/potentiellement adaptatives. Bien que notre jeu de données soit vraisemblablement incomplet, du fait notamment de la difficulté à estimer le polymorphisme d'insertions avec des données de séquençage (voir Section 4.5.1), nos résultats apportent des éléments quant à la contribution de la sélection positive à la dynamique des ET, et à la contribution des ET à l'adaptation.

4.2.1 Contribution de la sélection positive à la dynamique des ET

4.2.1.1 Une faible contribution

Avec 15 insertions potentiellement adaptatives sur 7004 sites polymorphes, nos résultats confirment une faible contribution de la sélection positive à la dynamique des ET. Globalement, nos résultats sont en accord avec les observations faites chez une variété d'organismes, et qui eux aussi suggèrent que seule une faible proportion d'insertion évolue sous sélection positive. Parmi ces observations on trouve notamment: la prédominance des insertions en faibles fréquences (Kofler et collab., 2015b; Stritt et collab., 2017), la déplétion en insertions dans les régions riches en

gènes/et ou fortement recombinantes (KOFLER et collab., 2012; WRIGHT et collab., 2003), ou encore la plus forte fréquence des SNP (BOURGEOIS et collab., 2020; LOCKTON et collab., 2008).

De façon intéressante, nos résultats suggèrent que la contribution de la sélection positive à la dynamique des ET puisse être l'augmentation en fréquence de l'allèle « sans insertion ». Historiquement, les études ce sont plutôt concentrées sur le cas opposé, c'est à dire une sélection de l'allèle « avec insertion » (GONZÁLEZ et collab., 2008). Ici, notre scan génomique, basé sur la différenciation et le contraste de fréquences, ne discrimine pas les deux cas. Sur 15 sites d'insertion candidats, quatre corrélations positives ont été trouvées entre les fréquences d'insertions et le D_{Taj} estimé localement. Le D_{Taj} étant attendu faible dans les régions évoluant sous sélection positive, ceci indique une potentielle sélection de l'allèle « sans insertion ». On peut notamment imaginer l'existence d'un évenement d'excision adaptatif. Il est important toutefois de rappeler ici qu'aucun D_{Taj} extrêmement faible n'a été retrouvé associé à une insertion en faible fréquence. Á l'avenir l'utilisation de scans génomiques ne discriminant pas parmi les allèles « avec ou sans insertion », couplée à un examen approfondi des fréquences d'insertions, pourrait permettre de découvrir des cas plus flagrants d'absence d'insertion évoluant sous sélection positive. Leur analyse détaillée permettrait alors une meilleure compréhension des processus à l'oeuvre (e.g. sélection de l'absence d'insertion ou sélection d'un évenement d'excision ?).

4.2.1.2 La délicate comparaison entre études

Il est tentant de comparer nos résultats avec les autres études ayant recherché des insertions potentiellement adaptatives à l'échelle du génome. Parmi les plus récentes on trouve celle de LI et collab. (2018) chez *A. thaliana* (2 insertions évoluant potentiellement sous sélection sur 2311 insertions polymorphes), celle de RECH et collab. (2019) chez *D. melanogaster* (300/1223), et celle de BOURGEOIS et collab. (2020) (4/339149) chez *Anolis carolinensis*. En terme de ratio entre le nombre d'insertions candidates et le nombre total d'insertions polymorphes, notre estimation est 2.47 fois supérieure à celle obtenue chez *A. thaliana*, 1156 fois inférieure à celle obtenue chez *D. melanogaster*, et 181 fois supérieure à celle obtenue chez *A. carolinensis*. Si ces différences peuvent sembler importantes, il est important d'une part de noter que, à l'exception de *D. melanogaster* (0.25), ces ratios sont tous très faibles (<0.0021), et d'autre part que tout ou partie de ces différences est potentiellement imputable à des différences de méthode [1]. Par exemple, en considérant que pour quatre loci la sélection agit peut-être sur l'absence d'insertion plutôt que la présence, le ratio obtenu chez *D. suzukii* n'est plus que 1.81 fois supérieur à celui obtenu chez *A. thaliana*. Une part de l'importante différence avec le ratio obtenu chez *A. carolinensis* est potentiellement imputable au fait que les 3394149 insertions polymorphiques mentionnées dans cette étude comporte des insertions avec MAF inférieur à 0.025 ce qui n'est pas le cas de nos 7004 insertions (649875 ont été filtré avant de réaliser notre scan génomique). Enfin, l'écart le plus important, celui avec l'étude réalisée chez *D. melanogaster* est potentiellement explicable par l'absence d'un contrôle direct pour des facteurs démographiques qui peuvent affecter la fréquence des insertions.

[1] Autant que possible les chiffres mentionnés sont ceux « mis en avant » par les auteurs, c'est à dire que si dans une publication les auteurs considèrent potentiellement adaptatives les insertions qui répondent aux conditions A et B, alors c'est le nombre d'insertions qui remplissent ces deux conditions qui sera utilisé. Il existe des différences importantes dans le niveau de preuve exigé par chaque auteur pour considerer une insertion comme potentiellement adaptative

4.2.1.3 Quid de l'espèce et de son histoire ?

Au-delà de l'aspect méthodologique, il est tout à fait vraisemblable qu'une part des différences observées entre les diverses études soit liée à l'histoire de l'espèce étudiée et à ses caractéristiques intrinsèques.

Toutes les espèces sur lesquelles portent les études mentionnées ont colonisé de nouveaux environnements relativement récemment, c'est d'ailleurs pourquoi elles ont été choisies pour analyser le potentiel adaptatif des ET. Ces espèces ne présentent toutefois pas la même histoire. La différence entre la proportion d'insertions candidates obtenue dans nos travaux et celle obtenue dans l'étude faite chez *D. melanogaster* pourrait ainsi être expliquée par une dynamique d'invasion différente. En effet, alors que *D. suzukii* est sortie d'Asie il y a quelques dizaines d'années (FRAIMOUT et collab., 2017), *D. melanogaster* est sortie d'Afrique il y a environ 15 000 ans avec un rythme d'invasion relativement lent (STEPHAN et LI, 2007). On peut donc imaginer que les événements de transposition adaptatifs ne se soient pas « encore produits » chez *D. suzukii*. Cet argument n'explique toutefois pas le très faible nombre d'insertions candidates obtenu dans l'étude chez *A. carolinensis*. En effet, cette étude cible des événements adaptatifs qui seraient survenus lors d'une invasion commencée il y a 2-3 millions d'années. Le très faible nombre d'insertions candidates obtenu chez *A. carolinensis* pourrait par contre s'expliquer par un faible contraste entre les environnements colonisés (Sud-Est des États-Unis sauf Floride) et l'environnement d'origine (Floride).

Des différences dans les caractéristiques intrinsèques des espèces, pourraient aussi avoir leur rôle à jouer. Ainsi les variations dans la proportion d'insertions évoluant potentiellement sous sélection positive entre les différentes espèces, pourraient être liées à la possession par certaines espèces de certains ET avec des séquences régulatrices au fort potentiel adaptatif. Il est aussi possible que certaines espèces présentent plus d'insertions adaptatives du fait d'un taux de transposition plus important. Les possibilités sont multiples. Une piste potentielle pour explorer le rôle des caractéristiques intrinsèques aux espèces serait de soumettre des espèces différentes à une même expérience d'évolution expérimentale et d'appliquer ensuite la même méthodologie pour une recherche d'insertions candidates.

4.2.2 Contribution des ET à l'adaptation

En plus de la question de la proportion d'insertions évoluant sous sélection positive, on peut s'intéresser au nombre d'insertions adaptatives sur le nombre de variants adaptatifs totaux. C'est-à-dire d'une certaine manière de la contribution des variants de type ET à l'adaptation. Jusqu'à aujourd'hui on dispose de peu d'éléments à ce sujet.

Le premier scan génomique à la recherche d'insertions potentiellement adaptatives évoque une insertion adaptative toutes les 2 000 - 25 000 générations chez *D. melanogaster* (GONZÁLEZ et collab., 2008). Ce chiffre est comparé par les auteurs à celui d'une substitution adaptative tous les 3 000 générations (MACPHERSON et collab., 2007). Cela correspond à un taux d'adaptation induit par les ET de 0.12 à 1.5 fois celui induit par les SNP.

Chez *D. suzukii*, une étude réalisée avec les mêmes méthodes et les mêmes données que celles utilisées dans cette thèse mais basée sur l'étude des SNP (OLAZCUAGA et collab., 2020), nous permet

de faire une comparaison similaire. En contrastant les fréquences alléliques entre populations invasives et natives, en utilisant la statistique C_2, les auteurs trouvent 204 SNP potentiellement adaptatifs. La même analyse sur nos insertions d'ET faisait ressortir six insertions potentiellement adaptatives. On obtient donc un taux d'adaptation induit par les ET de 0.029 fois inférieur à celui induit par les SNP.

Dans l'ensemble les résultats suggèrent une contribution des ET à l'adaptation, comparable à inférieure à celle SNP. Il est important de prendre en compte qu'un SNP adaptatif peut entraîner avec lui une insertion non adaptative et inversement, faussant ainsi les calculs effectués dans ce paragraphe. Toutefois, chez *D. suzukii*, la comparaison de la localisation des SNP et des insertions potentiellement adaptatives effectuée dans le Chapitre 2 suggère un faible taux de co-localisation. En effet, seule une insertion est proche d'un SNP potentiellement adaptatif.

4.2.3 Validation expérimentale des insertions potentiellement adaptatives

Notre scan génomique à la recherche d'insertions potentiellement adaptatives, est appelé à être suivi par une validation expérimentale des insertions candidates. Ce travail, dont l'objectif est de démontrer l'impact fonctionnel de ces insertions, a en réalité déjà commencé. Il s'effectue notamment en collaboration avec H. Henri, C. Régis, et S. Martinez qui participent à la recherche des insertions candidates dans les lignées de *D. suzukii* disponibles au laboratoire.

4.2.3.1 Recherche des insertions dans les lignées de laboratoire

Une approche Réaction en chaîne par polymérase (ou *Polymerase Chain Reaction* en anglais) (PCR) est en effet actuellement en cours pour déterminer la présence ou l'absence de nos 15 insertions candidates parmi les 34 lignées de *D. suzukii* au laboratoire. La conception d'amorces PCR telles que l'une soit complémentaire de l'insertion et l'autre de la région flanquant l'insertion permet de vérifier la présence d'une insertion dans l'ADN de nos lignées. Cette méthode à été validée sur trois insertions estimées fixées dans les populations de *D. suzukii* par notre pipeline d'analyse bioinformatique. Les données de présence/absence que nous obtiendrons permettront la réalisation d'un test d'EAS.

4.2.3.2 Test d'Expression Allèle Spécifique (EAS)

En effet, à partir d'une lignée avec et d'une lignée sans une insertion particulière, il est possible d'évaluer un effet de cette insertion sur le niveau d'expression d'un ou des gènes voisins. Il faut pour cela obtenir des individus issus du croisement des deux lignées, et concevoir des amorces pour PCR quantitative à partir d'un échantillon d'ARN (ou *Reverse Transcription and quantitative PCR* en anglais) (RT-qPCR) dont une paire amplifie spécifiquement l'allèle avec insertion et l'autre l'allèle sans insertion. La RT-qPCR permet ensuite d'évaluer si les deux allèles sont exprimés différemment. Il est à noter que cette approche n'exclut pas un effet d'un SNP/d'une InDel proche.

4.2.3.3 CRISPR-Cas9

Une évaluation de l'effet de l'insertion candidate sur le niveau d'expression d'un ou des gènes voisins, excluant un effet d'un SNP/InDel proche, pourra éventuellement se faire en utilisant l'approche courtes répétitions palindromiques regroupées et régulièrement espacées - protéine 9 associée à CRISPR (ou *Clustered Regularly Interspaced Short Palindromic Repeats-CRISPR associated protein 9* en anglais) (CRISPR-Cas9). Cette technique permet d'éditer le génome, et peut-être utiliser pour « supprimer » un ET (SAIKA et collab., 2019). Il s'agirait ensuite de comparer l'expression du ou des gènes chez la lignée modifiée avec la lignée d'origine. Il devient alors aussi possible de tester un effet phénotypique de l'insertion. Si au cours de ma thèse il n'a pas été possible de tester l'impact fonctionnel des insertions candidates trouvées, le travail de validation expérimentale est en cours, et devrait nous permettre de conclure.

4.3 Variations de dynamique intra-génomiques

On sait depuis longtemps que la répartition des ET n'est pas homogène au sein des génomes (BARTOLOMÉ et collab., 2002; MEDSTRAND et collab., 2002; WRIGHT et collab., 2003). Ceci est illustré dans les travaux présentés ici notamment par le caractère bimodal de la distribution de densité des ET dans le génome de *D. suzukii*. Cette répartition non homogène reflète nécessairement une variation intra-génomique dans la dynamique des ET.

4.3.1 Un « impact » de la densité en gènes

Les résultats du Chapitre 3 indiquent des variations intra-génomiques dans le taux d'accumulation d'ET chez *D. suzukii* au cours des derniers millions d'années. Plus précisément les ET semblent s'accumuler plus lentement dans les régions riches en gènes. L'augmentation du contenu en ET observée dans le Chapitre 2 quant à elle est potentiellement limitée aux régions riches en gènes. En effet, du fait de la très forte densité en ET, l'estimation du polymorphisme d'insertions dont découle notre mesure de l'abondance en ET est pratiquement impossible dans les régions pauvres en gènes. Enfin, toujours dans le deuxième chapitre, nos résultats suggèrent une dynamique des ET différente entre autosomes et gonosomes uniquement pour les régions riches en gènes, avec une densité en ET plus forte sur les gonosomes. On a donc une interaction entre la localisation autosomale ou gonosomale et la présence dans des régions riches ou pauvres en gènes.

4.3.2 Les facteurs clés

Les résultats présentés dans ce manuscrit pointent vers la densité en gènes comme un facteur clé impactant la dynamique des ET localement. Que la densité en gènes puisse jouer un rôle important dans la dynamique locale des ET est vraisemblable. Ceci s'expliquerait notamment par des différences d'impact, entre les insertions proches de gènes et les insertions plus éloignées, sur la valeur sélective de l'hôte. En effet, on attend un effet délétère plus fort des insertions proches des gènes (FINNEGAN, 1992), et donc une plus forte contre sélection de ces insertions. En accord avec

cette hypothèse, on trouve notamment la plus forte accumulation d'ET dans les régions pauvres en gènes au cours des derniers millions d'années observée dans le Chapitre 3.

Toutefois cette dernière observation pourrait aussi s'expliquer par un plus fort taux de recombinaison dans ces régions, donc de recombinaison ectopique, entrainant encore une fois une plus forte contre-sélection. En effet, chez *D. melanogaster*, proche parente de *D. suzukii*, il existe une corrélation positive entre le taux de recombinaison et la densité en gènes. Il est possible que la même relation existe chez *D. suzukii*.

On pourrait aussi imaginer avoir un effet d'autres facteurs sur la dynamique locale des ET, notament la compaction de la chromatine. Des travaux suggèrent que dans les régions euchromatiques, où la chromatine est peu compacte, l'insertion des ET soit plus facile que dans les régions hétérochromatiques, où l'ADN est très compact. Cet effet pourrait être d'autant plus important proche des gènes, dont l'expression nécessite une ouverture de la chromatine (FONTANILLAS et collab., 2007; WALSER et collab., 2006). Ceci n'est toutefois pas cohérent avec l'observation d'une plus forte accumulation d'ET dans les régions pauvres en gènes au cours des derniers millions d'années.

4.3.3 Vers une prise en compte plus systématique des variations intra-génomiques ?

À l'heure actuelle les variations intra-génomiques ne sont pas nécessairement prises en compte dans l'étude de la dynamique des ET. Lorsque l'on compare le contenu en ET entre espèces ou populations par exemple, on considère souvent le contenu total (LERAT et collab., 2019; TALLA et collab., 2017). Il y a au moins une explication d'ordre technique à ceci. En effet, pour bon nombre d'espèces, l'assemblage génomique, quand il existe, est incomplet. Par exemple, une comparaison de deux assemblages, telle que celle effectuée dans le Chapitre 3 entre *D. suzukii* et *D. biarmipes*, n'est pas envisageable avec des assemblages obtenus à partir des NGS. L'arrivée du séquençage longues lectures, en améliorant la qualité des assemblages, devrait faciliter la prise en compte des variations intra-génomiques.

Chez la Drosophile notamment, il devrait maintenant être possible d'évaluer la différence de taux d'accumulation des ET dans les différents compartiments du génome aux échelles intra- et inter-spécifique. Pour ce qui est de l'échelle intraspécifique, il existe un jeu de données de séquençage « poolé » de 48 populations Européennes de *D. melanogaster* (KAPUN et collab., 2018). L'application de méthodes similaires à celle employée dans le Chapitre 2, i.e. une estimation du contenu en ET via l'étude du polymorphisme d'insertion, et l'investigation des variations locales pourrait permettre de déterminer si le contenu en ET varie différemment dans les différentes régions du génome. Il serait ainsi possible de déterminer si la majorité de la variation est liée aux régions euchromatiques ou aux régions hétérochromatiques. En ce qui concerne l'échelle interspécifique, la comparaison des assemblages longues lectures de 15 espèces de Drosophile disponibles (MILLER et collab., 2018), permettrait de la même manière d'identifier une hétérogénéité spatiale dans le taux d'accumulation.

4.4 Deux « oubliés »

Parmi les moteurs potentiels de la dynamiques des ET, deux moteurs potentiellement importants, n'ont pas fait l'objet d'une étude approfondie.

4.4.1 Le rôle des mécanismes de défense

En limitant leur activité, les mécanismes de défense de l'hôte sont supposés jouer un rôle crucial dans la dynamique des ET. L'impact éventuel de la variation génétique au niveau des gènes effecteurs de la voie des piRNA a été étudiée à l'échelle des populations dans le Chapitre 2. Ces travaux n'ont pas permis de mettre en évidence de rôle déterminant. Plus précisement ils montrent une absence d'enrichissiment en gènes effecteurs de la voie des piRNA dans les régions significativement associées à l'abondance des ET. Une étude réalisée chez *A. thaliana* présente des résultats comparables avec seulement un gène potentiellement impliqué dans les mécanismes de défense de l'hôte parmi 230 loci significativement associées à l'abondance des ET (Quadrana et collab., 2016).

Si une récente expérience d'évolution expérimentale chez le nématode suggère un impact du relâchement de la sélection au niveau des gènes effecteurs de la voie des piRNA sur le niveau d'expression des ET (Bergthorsson et collab., 2020), dans le Chapitre 3 aucun relâchement de la sélection sur les gènes effecteurs de la voie des piRNA n'a été trouvée en lien avec la prolifération des ET observée au cours des derniers millions d'années. Il est à noter toutefois que, du fait du caractère partiel des assemblages génomiques, cette dernière étude n'a pu être faite que sur trois gènes. L'application d'une méthodologie similaire, avec des assemblages obtenus par technologies de séquençage longues lectures permettrait d'établir un résultat avec beaucoup plus de certitude. Un tel assemblage existe maintenant pour *D. subpulchrella* (Kim et collab., 2020).

De manière générale, les résultats obtenus dans cette thèse ne mettent pas en évidence un rôle décisif des mécanismes de défense pour expliquer la dynamique des ET. Toutefois, cette question aurait pu faire l'objet d'une étude plus approfondie. Notamment les gènes ne sont pas les seuls acteurs de la voie des piRNA. En effet les loci producteurs de piRNA, ou clusters de piRNA, jouent aussi un rôle clé dans la régulation de l'activité des ET. Ceux-ci n'ont pas pu être étudiés ici. Ainsi parmi les ~5000 régions génomiques associées à des variations d'abondance de certaines familles d'ET dans les populations de *D. suzukii*, certaines pourraient correspondre à des clusters de piRNA. Le séquençage des piRNA, comme par (Roy et collab., 2020) ou comme en Annexe 1, et l'alignement de ces piRNA sur le génome permettrait de déterminer la position des clusters et de valider ou invalider cette hypothèse. Il est aussi envisageable que des réductions de taille des clusters de piRNA aient joué un rôle dans l'accumulation des ET observée dans le génome de *D. suzukii* après sa divergence avec *D. biarmipes*. Le séquençage des piRNA chez *D. biarmipes* et *D. suzukii*, permettrait de comparer la taille des clusters entre ces deux espèces, et offrirait ainsi de premiers éléments de réponse.

4.4.2 L'évolution des ET eux-même

Paradoxalement un grand oublié dans l'étude de la dynamique des ET c'est l'évolution des ET eux-mêmes. Aujourd'hui rares sont les études sur les pressions de sélection qui agissent sur la séquence des ET (MES et DOELEMAN, 2006; THABET et collab., 2015). L'étude du dN/dS chez la cyanobactérie *Crocosphaera watsonii* suggère pourtant une action de la sélection positive au niveau de la séquence de la transposase au préalable. La sélection positive pourrait jouer un rôle dans l'évolution de l'activité des ET, l'évolution des préférences d'insertions, ou tout autre paramètre potentiellement clé de la dynamique des ET.

Théoriquement, l'étude des pressions de sélection semble envisageable, notamment avec des données telles que celles utilisées dans cette thèse. En effet, en ayant potentiellement vérifié au préalable son comportement sur des données simulées, on pourrait imaginer utiliser le ratio dN/dS. À partir de la banque d'ET générée pour *D. suzukii*, et de banques similaires générée pour des espèces proches, on peut aligner les séquences codantes et estimer un dN/dS par famille. On pourrait ainsi voir si les familles les plus abondantes présentent un dN/dS plus important, comme on pourrait l'attendre si la sélection positive favorise la prolifération.

4.5 Considérations « bioinformatiques »

Cette thèse doit beaucoup à l'outil bioinformatique et aux dernières technologies de séquençage, mais elle fait aussi ressortir un certain nombre de limitations relatives à leur utilisation pour l'étude des ET.

4.5.1 Les difficultés liées aux caractéristiques des ET

4.5.1.1 Le caractère répété

Le caractère répété des ET permet certes leur identification dans les assemblages comme dans les lectures génomiques, mais il rend aussi leur étude difficile à partir des données de séquençage. La répétition de séquences complique notamment l'assemblage des génomes. Les régions répétées sont souvent imparfaitement assemblées et difficiles à étudier à partir des assemblages (voir aussi Section 0.3.1).

Les lectures longues permettent un meilleur assemblage des régions répétées que les lectures courtes. Le premier assemblage de la lignée de référence de *D. suzukii* obtenu à partir de lectures courtes comprend 236 millions de paires de bases, dont 4.9% d'ET [1] (CHIU et collab., 2013). L'assemblage de la même lignée obtenu à partir de lectures longues comprend quant à lui 270 millions de paires de bases, dont 35-47% de séquences répétées (PARIS et collab., 2020).

Dans le cas où seules des lectures courtes sont disponibles, le contenu en ET est généralement obtenu en alignant ces lectures sur une banque de séquences d'ET (voir section 0.3.4). Dans le Chapitre 3 par exemple, les comparaisons de contenu en ET entre espèces ont été effectuées à partir des lectures de séquençage et non à partir des assemblages.

[1] Plus précisement il s'agit de 4.9% d'ET pour les segments assemblé de plus de 5 kb uniquement

Le travail présenté en Annexe 1, montre que l'estimation du contenu en ET dans les assemblages longues lectures de *D. melanogaster* et *D. simulans* est corrélée avec l'estimation du contenu en ET à partir de lectures courtes (Annexe 1, fig. 2). Ceci suggère la validité des méthodes consistant à aligner les lectures courtes directement sur une banque de séquences d'ET. Toutefois les résultats présentés en Annexe 5, semblent indiquer l'existence de biais, au moins dans certains jeux de données, liés notamment à une couverture non-homogène le long du génome (voir Section 4.5.3). De plus, l'assemblage à partir de lectures longues permet d'aller plus loin dans l'analyse du contenu, e.g. étude de la distribution relative des gènes et des ET, et sa « démocratisation » sera une plus value importante, notamment pour l'étude des espèces non modèles.

Une autre difficulté engendrée par le caractère répété des ET est la nécessité de regrouper ensemble les séquences, des lectures ou des génomes assemblés, qui sont des copies d'un même ET. La banque d'ET construite dans ce manuscrit est telle que toutes les séquences possédant plus de 80% d'identité sont considérées comme autant de copies d'un même ET. Le seuil de 80% d'identité, bien que conventionnel est arbitraire, et ne correspond pas nécessairement à une réalité évolutive. De fait, certaines analyses réalisées au niveau de la famille perdent potentiellement de leur sens. Lorsque l'on teste l'existence d'une corrélation entre variables climatiques et abondance d'un ET par exemple, on mélange potentiellement deux ET avec des histoires évolutives différentes. Ceci est d'autant plus vrai que certaines lectures s'alignent sur plusieurs ET de la banque rendant impossible la distinction entre eux.

4.5.1.2 La répartition non uniforme

La répartition souvent non uniforme des ET, et notamment l'existence de régions les accumulant, complique aussi leur analyse bioinformatique. Dans ces régions les ET sont très proches, et éventuellement nichés les uns dans les autres. Cela engendre deux problèmes principaux. D'une part, l'étude du polymorphisme d'insertion est presque impossible dans ces régions. En effet pour étudier le polymorphisme d'une insertion il est nécessaire que celle-ci soit flanquée de région(s) non répétée(s) (voir Section 0.3.3). Or dans le génome de *D. suzukii* près de 40% des fenêtres de 200 kb contiennent plus de 75% d'ET (voir Annexe 5). D'autre part, lors de l'annotation des copies dans le génome assemblé, il est difficile de reconstituer des copies complètes lorsqu'un ET s'est inséré dedans. Ceci explique vraisemblablement les presque 40000 copies de la superfamille *Copia* retrouvées dans l'assemblage de *D. suzukii*. En effet, en considérant une taille de copie moyenne pour ces éléments à LTR de 7500 kb, avec 40000 copies, 300 Mb seraient occupés par des éléments de la superfamille. Ce chiffre est supérieur à celui de la taille de l'assemblage (270 mb) ! Il est donc vraisemblable que certaines copies, « coupées en deux » par une insertion ultérieure soient comptabilisées deux fois. Ce phénomène est potentiellement amplifié par la dégradation des copies qui se produit au cours du temps. En effet, avec le temps et sous l'effet des mutations, la séquence des copies évolue et leur identification par homologie peut aboutir à considérer comme copies différentes des portions d'une même copie. On peut aussi noter que, bien que massivement utilisé, le logiciel RepeatMasker n'a jamais fait l'objet d'une publication dans un journal à comité de lecture, et possède potentiellement certains biais.

4.5.2 Des estimations outils spécifiques

Si, contrairement à RepeatMasker, la très grande majorité des outils bioinformatiques a fait l'objet d'une publication, beaucoup d'entre eux donnent des résultats très différents sur un même jeu de données. Par exemple, dans le Chapitre 3, l'estimation de la quantité d'ET dans les génomes de *D. suzukii* et sept espèces proches avec l'outil dnaPipeTE (GOUBERT et collab., 2015) donne des valeurs supérieures à celles obtenues avec l'outil deviaTE (WEILGUNY et KOFLER, 2019). Dans l'Annexe 1, dnaPipeTE estime des contenus en ET plus importants dans des lignées de *D. melanogaster* et *D. simulans* que l'outil TEcount (LERAT et collab., 2016). Parfois, comme c'est le cas pour ce dernier résultat, les différences entre outils ne sont pas facilement interprétables. En effet, on pourrait attendre que l'outil dnaPipeTE, qui construit lui même sa banque d'éléments à partir de l'assemblage des lectures courtes de chaque échantillon (voir Section 0.3.4), estime des contenus en ET chez *D. melanogaster* et *D. simulans* inférieurs à TEcount. Ceci parce que TEcount utilise ici lui une banque issue de plusieurs années de travail sur la Drosophile qu'on pourrait attendre plus complète que celle construite par dnaPipeTE. Afin d'identifier, et potentiellement d'éviter, d'éventuels biais, il est possible de tester les outils utilisés avec des données simulées. La simulation de données de « PoolSeq » dans le Chapitre 3, a permis de vérifier la fiabilité de l'outil PoPoolationTE2 pour estimer le polymorphisme d'insertion avec nos données (KOFLER et collab., 2016).

4.5.3 Profondeur de séquençage non uniforme: un biais sous estimé ?

L'abondance des ET peut-être obtenue de deux manières avec des lectures courtes : l'alignement sur une librairie d'ET et l'utilisation d'outils dédiés à l'étude du polymorphisme d'insertion (voir section 0.3.4). Souvent considérée comme plus fiable, car indépendante de la qualité de l'assemblage ou de l'existence de régions avec de fortes densité d'ET, la deuxième méthode est souvent préférée (LERAT et collab., 2019; QUADRANA et collab., 2016). Toutefois cette méthode fait l'hypothèse que la profondeur de séquençage est uniforme le long du génome. Or l'étude de la couverture dans des fenêtres génomiques de 200 kb dans les échantillons de « PoolSeq » de *D. suzukii* suggèrent d'importantes variations (voir Annexe 5). Les régions fortement répétées ne présentent pas nécessairement la même couverture que les régions faiblement répétées. Pour l'échantillon « FR-Run » par exemple, la couverture médiane des régions fortement répétées est de 59X, contre 85X pour les régions faiblement répétées. Dans cet échantillon le contenu en ET estimé par l'alignement des lectures sur une librairie d'ET est vraisemblablement sous estimé. Le problème est d'autant plus grand que le biais de couverture semble variable entre échantillons, rendant impossible la comparaison. D'après nos analyses une part très importante de la variance dans les estimations obtenues pourraient être expliquées par le ratio de couverture entre les régions faiblement et les régions fortement répétées. Le fait que la couverture soit en très grande majorité inférieure dans les régions fortement répétées pourraient être dû à un protocole d'extraction de l'ADN imparfait. En effet, chez *D. melanogaster* les régions fortement répétées sont hétérochromatiques, c'est-à-dire que l'ADN y est plus compacté. Si la même tendance existe chez *D. suzukii*, il est possible que du fait d'un protocole mal adapté ces régions soient moins bien représentées dans l'ADN extrait. Il s'agit toutefois d'une simple hypothèse.

4.6 Conclusion

Ce travail de thèse a permis d'améliorer notre compréhension des ET et de leur dynamique chez l'espèce invasive *D. suzukii*. Avec 47% d'ET dans son génome, *D. suzukii* présente un des pourcentage d'occupation du génome par les ET les plus importants du genre Drosophila. Si le répétome de *D. suzukii* se distingue par sa taille, il semble toutefois partager avec les autres espèces de Drosophile le caractère actif de ces éléments (plus de 75% des insertions ségregent à faible fréquence). De plus, on retrouve chez *D. suzukii* une caractéristique répandue dans l'arbre de la vie, une distribution non homogène des ET dans le génome. Ces derniers sont concentrés dans les régions pauvres en gènes. Une observation qui reflète vraisemblablement l'effet majoritairement délétère des ET à proximité des gènes, et qui souligne l'importance de la sélection purificatrice dans leur dynamique.

L'étude du polymorphisme d'insertions dans 22 populations de *D. suzukii* à permis de mettre en évidence une large augmentation du contenu en ET associée au processus d'invasion des continents Américain et Européen depuis l'Asie en 2008. La comparaison du contenu en ET entre *D. suzukii* et espèces proches montre que cette augmentation récente du contenu en ET a été précédée d'une autre plus ancienne, plus lente et essentiellement limitée aux régions génomiques pauvres en gènes. C'est cette dernière, s'étalant sur les quatres derniers millions d'années, qui serait largement responsable des 47% d'ET génomique chez *D. suzukii*. L'utilisation du $\widehat{\theta_W}$ comme proxy de la taille efficace dans les 22 populations de *D. suzukii* suggère fortement un rôle d'un relâchement de la sélection purificatrice, via une réduction de la taille des populations, comme moteur de l'accumulation des ET observée lors de l'invasion des continents Américain et Européen. En revanche, la comparaison des pressions de sélection moyennes entre *D. suzukii* et espèces proches, par le moyen du dN/dS, n'a pas permis de mettre en lien l'accumulation d'ET ces quatre derniers millions d'années avec un relâchement de la sélection purificatrice.

Au delà d'un potentiel rôle des variations de taille des populations, les travaux présentés ici apportent de nouveaux éléments quand à la dynamique des ET lors d'une invasion biologique. Si l'impact des changements environnementaux sur l'activité des ET à été largement étudié en laboratoire chez la Drosophile, nos résultats indiquent notamment un faible rôle de ces changements dans les variations de contenu en ET observées lors de l'invasion de *D. suzukii*. Ils soulignent aussi un rôle complexe des variations génétiques aux niveaux des séquences des ET, mais aussi des séquences génomiques de leur hôtes. Enfin, bien que cette thèse souligne la prépondérance d'un impact négatif des ET sur leur hôte, elle suggère aussi un potentiel rôle adaptatif lors de l'invasion pour un petit nombre d'insertions.

À très court terme la validation expérimentale de nos insertions candidates commencée en laboratoire permettra de confirmer ou d'infirmer leur impact fonctionnel. À plus long terme, la démocratisation des séquençages longues lectures et PoolSeq, qui ont grandement bénéficié à cette thèse, devrait permettre de valider la généralité de nos observations. La combinaison de ce type de données avec des données de séquençage de piRNA, permettra de localiser et de déterminer la taille des clusters de piRNA, et ainsi de mieux comprendre leur rôle dans la dynamique des ET à l'échelle des populations, mais aussi à l'échelles des espèces.

« Stop, ça y est,
J'arrête de penser
Je vais courir,
Je vais marcher»

Ben Mazué

A Annexes

Sommaire

A.1 A Transposon Story: From TE Content to TE Dynamic Invasion of Drosophila Genomes Using the Single-Molecule Sequencing Technology from Oxford Nanopore

cells

MDPI

Article

A Transposon Story: From TE Content to TE Dynamic Invasion of *Drosophila* Genomes Using the Single-Molecule Sequencing Technology from Oxford Nanopore

Mourdas Mohamed [1,†], Nguyet Thi-Minh Dang [2,†], Yuki Ogyama [1], Nelly Burlet [3], Bruno Mugat [1], Matthieu Boulesteix [3], Vincent Mérel [3], Philippe Veber [3], Judit Salces-Ortiz [3,4], Dany Severac [5], Alain Pélisson [1], Cristina Vieira [3], François Sabot [2], Marie Fablet [3,*] and Séverine Chambeyron [1,*]

1 Institute of Human Genetics, UMR9002, CNRS and Montpellier University, 34396 Montpellier, France; mourdas.mohamed@igh.cnrs.fr (M.M.); yuki.ogiyama@igh.cnrs.fr (Y.O.); bruno.mugat@igh.cnrs.fr (B.M.); alain.pelisson@igh.cnrs.fr (A.P.)
2 IRD/UM UMR DIADE, 911 avenue Agropolis BP64501, 34394 Montpellier, France; dangminhnguyet09@gmail.com (N.T.-M.D.); francois.sabot@ird.fr (F.S.)
3 Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, 69622 Villeurbanne, France; nelly.burlet@univ-lyon1.fr (N.B.); matthieu.boulesteix@univ-lyon1.fr (M.B.); vincent.merel@etu.univ-lyon1.fr (V.M.); philippe.veber@univ-lyon1.fr (P.V.); judit.salces@ibe.upf-csic.es (J.S.-O.); cristina.heddi@univ-lyon1.fr (C.V.)
4 Institute of Evolutionary Biology (IBE), CSIC-Universitat Pompeu Fabra, 08003 Barcelona, Spain
5 MGX-Montpellier GenomiX, c/o Institut de Génomique Fonctionnelle, CNRS, INSERM, Université de Montpellier, 34094 Montpellier, France; Dany.Severac@mgx.cnrs.fr
* Correspondence: marie.fablet@univ-lyon1.fr (M.F.); severine.chambeyron@igh.cnrs.fr (S.C.); Tel.: +33-47-243-2916 (M.F.); +33-43-435-9949 (S.C.)
† These authors contributed equally to this work.

Received: 27 June 2020; Accepted: 23 July 2020; Published: 25 July 2020

Abstract: Transposable elements (TEs) are the main components of genomes. However, due to their repetitive nature, they are very difficult to study using data obtained with short-read sequencing technologies. Here, we describe an efficient pipeline to accurately recover TE insertion (TEI) sites and sequences from long reads obtained by Oxford Nanopore Technology (ONT) sequencing. With this pipeline, we could precisely describe the landscapes of the most recent TEIs in wild-type strains of *Drosophila melanogaster* and *Drosophila simulans*. Their comparison suggests that this subset of TE sequences is more similar than previously thought in these two species. The chromosome assemblies obtained using this pipeline also allowed recovering piRNA cluster sequences, which was impossible using short-read sequencing. Finally, we used our pipeline to analyze ONT sequencing data from a *D. melanogaster* unstable line in which LTR transposition was derepressed for 73 successive generations. We could rely on single reads to identify new insertions with intact target site duplications. Moreover, the detailed analysis of TEIs in the wild-type strains and the unstable line did not support the trap model claiming that piRNA clusters are hotspots of TE insertions.

Keywords: transposable elements; ONT; *Drosophila melanogaster*; *Drosophila simulans*; piRNA

1. Introduction

Transposable elements (TEs) are major components of almost all eukaryotic genomes [1,2]. They can be separated into three main groups that include several TE superfamilies and families:

DNA transposons, Long-Terminal Repeat (LTR) elements, and Long Interspersed Nuclear Elements (LINEs) [2,3]. Different methods (e.g., Southern blotting [4,5], in-situ hybridization on polytene chromosomes [6,7], and PCR [8,9]) were first used to estimate TE content in *Drosophila* genomes and to understand how TEs invade and shape genomes by affecting genome function and evolution. However, technical problems linked to TE repetitive nature and diversity have not allowed for the reaching of firm conclusions and many questions about TE biology remain unanswered.

Then, next-generation short-read sequencing technologies allowed for characterizing the global TE content within and between related species. Moreover, the high coverage provided by Illumina sequencing led to the identification of consensus sequences for each TE family. Several computational methods were developed, such as RepeatExplorer [10] and dnaPipeTE [11], to analyze Illumina data from different *Drosophila* species, and to study TE biology at the populational level.

In TE biology, it is also important to estimate the TE insertion (TEI) rate to determine the degree of polymorphism within and between populations. This is an indicator of the activity level of each TE family and can help to date transposition events [12,13]. Illumina sequencing of pools of individuals allowed for determining TEI frequency in natural samples (from large numbers of individuals to populations) [14]. To study individual TEI, dedicated software tools were developed (e.g., TIDAL [15], T-Lex/T-Lex2 [16], PopoolTE2 [17]) based on the analyses of: (1) the TEI junction and flanking sequences (split-reads), (2) the paired-end information, (3) the depth of coverage, or (4) a mix of these three criteria. However, these approaches revealed only a portion of the repetitive sequence landscape, and they detect many false positives due to various factors. The first one is linked to the library preparation and PCR amplification that lead to the generation of PCR chimera and thus false positive insertions [18], or to biased sequence representativity (AT- and GC-rich sequences are less represented in Illumina sequencing). The second factor is inherent to the sequencing size (short reads) that does not span more than 400 bp, thus hindering the full sequencing of any repeat or variation larger than this size, especially insertions [19]. The third one is related to the difficulty in detecting TEIs occurring at low frequency in an individual or population. Indeed, these TEIs are usually under-represented in the sequencing data and generally confused with background errors [18]. The comparison of different methods to identify TEIs shows very small levels of overlap [20]. Another weak point of the Illumina sequencing technology is that the insertion size and sequence are not accessible, because this approach generally only gives the global position.

Long-read, or third-generation, sequencing technology might improve the detection of long structural variants and thus of TE variations, and also reduce the detection of false positives/false negatives. This technology should allow for the identification of full copies. Indeed, long-read sequencing methods generate individual reads that are mostly longer (>15 kb) than many of the repeats (TE sequences are generally smaller than 10 kb). Moreover, it solves the problems linked to PCR-based library preparation because it relies on direct DNA sequencing without amplification. However, the main drawback of long-read sequencing, such as the Oxford Nanopore Technology (ONT), is the high rate of single read sequencing errors (3 to 8% for the recent sequencing and base calling) that could introduce bias in data interpretation. This problem is partially solved by increasing the coverage and by improving the final assembly quality by polishing, thus providing an almost perfect genome sequence. Such an approach, based on PacBio sequencing, has already allowed the detection of 38% more TEIs in *Drosophila* chromosome 2 L compared with the available short-read sequencing estimates [21]. Different *Drosophila* genome assemblies using ONT sequencing have also been reported [22,23]. Long-read sequencing methods allow almost complete chromosome-scale genome assemblies, instead of the fragmented draft genomes provided by short reads. Therefore, the assembled individual genomes can be directly compared without the need for any reference genome and their relative structural variants can be scored without biases (or very few). In addition, long-read sequencing of genomes should allow identifying real TEI sites and accurately determining TE copy number at the inter- and intra-population levels. This approach might also help to analyze

repetitive regions like PIWI-interacting RNA (piRNA) clusters that contribute to maintaining genome integrity by repressing TE mobility.

Here, we developed strategies to generate *de novo* assemblies of high quality long-read sequencing data, suitable for genomic analyses of TEs present at high and low frequencies in *Drosophila* populations. We first validated our method by comparing the data (genome size, TE content and TEI site estimation) obtained by short and long-read sequencing in *D. melanogaster* and *D. simulans*, two closely related species, but that may vary in TE content [24,25]. We found that, although the *D. simulans* genome contains a large number of old and degraded TE copies, among the most recent pool of insertions, DNA transposons display higher intra-family sequence divergence than LTR elements, suggesting that elements of this group invaded the genome more recently than DNA transposons. Moreover, we observed that piRNA production correlates with TE genome occupancy. When considering the most recent pool of TE insertions, we could not find convincing evidence supporting the piRNA clusters trap model [26,27]. Finally, we developed and validated an approach to identify TEI that occur at low frequencies in a population.

2. Materials and Methods

2.1. Drosophila Strains

The wild-type *D. melanogaster* and *D. simulans* strains from natural populations were kept at 24 °C in standard laboratory conditions on cornmeal–sugar–yeast–agar medium. The eight samples of *D. melanogaster* and *D. simulans* natural populations were collected using fruit baits in France (Gotheron, 44°56′0″N 04°53′30″E-"goth" lines) and Brazil (Saõ Jose do Rio Preto 20°41′04.3″S 49°21′26.1″W-"sj" lines) in June 2014. Two isofemale strains per species and geographical origin were established directly from gravid females from the field (French *D. melanogaster*: dmgoth63, dmgoth101; Brazilian *D. melanogaster*: dmsj23, dmsj7; French *D. simulans*: dsgoth613, dsgoth31; Brazilian *D. simulans*: dssj27, dssj9). Brothers and sisters were then mated for 30 generations to obtain inbred strains with a very low amount of intra-line genetic variability.

A previously published *D. melanogaster* laboratory line [28] was used for Piwi knockdown (piwi KD) in adult follicle cells. This line carries three components: (i) a GAL4 UAS activator driven by the follicle cell-specific traffic jam (tj)-promoter (tj-GAL4), (ii) an UAS short hairpin(sh)-piwi that induces Piwi RNAi, and (iii) the ubiquitously expressed thermo-sensitive GAL4-inhibitor $GAL80^{ts}$. At 20 °C, $GAL80^{ts}$ sequesters GAL4, preventing sh-piwi expression. At 25 °C, $GAL80^{ts}$ is partially inactive, allowing some GAL4-driven expression of sh-piwi in somatic follicle cells. The resulting partial Piwi depletion allows for the derepression of at least two LTR families (ZAM and gtwin) in follicle cells and their integration as new proviruses in the progeny genome [28]. The polymorphism of this line was partially reduced by isolating a single pair of parents and the line was thereafter stably maintained at 20 °C as a large population (more than 500 progenitors at each generation). The G0 and G0-F100 genomic libraries were prepared shortly after isolation of this line and at the hundredth generation, respectively. Soon after isolation of this isofemale line, a subset of individuals at the pupal to early adult stages was shifted to 25 °C for 5 days, and this was repeated for at least 500 flies for 73 successive generations of partial Piwi KD. Then, after six more generations of stabilization at 20 °C, a third genomic library, called G73, was generated.

2.2. Genome Size Estimations

Flow cytometry: genome size was estimated according to [29] using fresh samples of 4-day-old females heads with 10 replicates (five heads per replicate) for each *Drosophila* wild-type strain.

findGSE: k-mer distribution was established from the Illumina reads using findGSE [30]. Briefly, adaptors were first removed from the reads with Skewer version 0.2.2 (paired-ends) or NxTrim version 0.4.3-6eb8d5e (mate pairs), when necessary. Reads were then treated essentially as previously described [31] to remove duplicates, filter out reads mapping to reference mitochondrial genomes

(GenBank AF200854.1 and AF200828.1 [32]) or microbial contaminants. This allowed for establishing the 21-mer distributions from which genome sizes were estimated using findGSE [30] with default parameters, except for dmsj23 in which the k-mer distribution clearly displayed a peak corresponding to heterozygous regions and was thus treated accordingly.

2.3. Illumina Sequencing

Wild-type strains: DNA was extracted from 3 to 5-day-old females for each strain using the Qiagen DNeasy Blood&Tissue kit (# 69506) and following the manufacturer's instructions. Genomic DNA (1.5 µg) was fragmented for a target insert size of 300 base pairs and sequenced by paired-end Illumina HiSeq (125 bp reads). Library and sequencing were performed by the GeT-PlaGe facility, Génopole Toulouse (France).

2.4. DNA Isolation, Oxford Nanopore MinION Sequencing and Base Calling

DNA was extracted from ~100 males from each wild-type and from thePiwi KD lines using the Qiagen DNeasy Blood&Tissue kit. The genomic DNA quality and quantity were evaluated using a NanoDrop™ One UV-Vis spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and a Qubit® 1.0 Fluorometer (Invitrogen, Carlsbad, CA, USA), respectively. Three micrograms of DNA were repaired using the NEBNext FFPE DNA Repair Mix (NEB M6630). End repair and dA-tailing were performed using the NEBNext End repair/dA-tailing Module (E7546, NEB). Ligation was then performed with the Ligation Sequencing Kit 1D (SQK-LSK108, ONT, for G0, and SQK-LSK109 ONT for wild type strains, G73 and G0-F100 samples). MinION sequencing was performed according to the manufacturer's guidelines using R9.4.1 flow cells (FLO-MIN106, ONT) and a Nanopore MinIon Mk1b sequencer (ONT) controlled by the ONT MinKNOW software (version 18.3.1 for G0, version 19.05.0.0 for isogenic wild-type strains, and version 19.10.1 for the G73 and G0-F100 samples). Base calling was performed after sequencing using Albacore (version 2.3.3) for G0, and the GPU-enabled guppy basecaller in high accuracy mode for isogenic wild-type strains (version 3.1.5), G73 (version 3.3.3) and G0-F100 samples (version 3.4.4).

2.5. TE Content and TEI Site Estimates from Illumina Sequencing

TE abundance was estimated using forward reads and two methods: the TEcount module of TEtools [33] and dnaPipeTE (v1.0.0 and v.1.3.1) [11]. TEcount estimates TE abundance by quantifying reads that map to a set of known TE sequences, here the rosetta fasta file [34]. This tool was run using default parameters and Bowtie2 (v2.2.4) [35,36]. dnaPipeTE assembles repeated sequences from a subsample of reads (<1x) and quantifies reads mapping to these sequences to estimate TE abundance. dnaPipeTE was used with the following parameters: -sample_number 2, -genome_coverage 0.25). Concerning the genome size option, 175 Mb and 147 Mb were used for *D. melanogaster* and *D. simulans* samples, respectively. The rosetta fasta file was used as library [34].

TEIs were detected in Illumina sequencing data using a dedicated mapping-based algorithm similar to that implemented in PoPoolationTE2 [17] with paired-end reads as input, FlyBase reference genomes (ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.16_FB2017_03/fasta/dmel-all-chromosome-r6.16.fasta.gz and ftp://ftp.flybase.net/genomes/Drosophila_simulans/dsim_r2.02_FB2017_04/gtf/dsim-all-r2.02.gtf.gz), and the TE sequence library at https://github.com/bergmanlab/transposons/raw/e2a12ff708c42dcce5b15d6af290506d78021212/releases/D_mel_transposon_sequence_set_v10.1.fa. Sequencing reads are mapped to the reference genome and TE sequences using Bowtie2 (version 2.3.3) [36]. Then, the algorithm scans the resulting Binary Alignment/Map (BAM) files for pairs in which one end matches to the reference genome, the other end to a TE sequence, and the pair cannot be mapped concordantly to the genome. For each pair, the position of the genome-mappable read is noted, and positions are clustered in order to have no read further apart than 100 bp in that cluster. Each cluster is then interpreted as an insertion, the position of which is the mean of the position of the reads it contains, and the strength of which is evaluated on

the basis of the number of reads it contains. For the purpose of this study, only insertions that were supported by at least 50 reads were retained. Unlike PoPoolationTE2, the insertions detected with this procedure correspond to occurrences absent from the reference genome.

2.6. Small RNA Extraction and Sequencing

For small RNA sequencing, two replicates per strain were prepared. Small RNA was isolated from 50 pairs of ovaries using HiTrap Q HP anion exchange columns (Cytiva, Velizy-Villacoublay, France) as described in [37], and the eluate was run on a 10% TBE urea gel (Thermo Fisher Scientific). Small RNA size selection (18–50 bp) was performed on gel at the sequencing facility. Quality was checked with the Bioanalyzer small RNA kit (Agilent, Santa Clara, CA, USA). Library construction was performed using the TruSeq Small RNA Library kit (Illumina, San Diego, CA, USA) and sequenced (1 × 50 single reads) on an Illumina HiSeq 4000 at the IGBMC Microarray and Sequencing facility. Adapter sequences were removed using cutadatp [38]. Size selection was then performed using PRINSEQ lite version 0.20.4 [39]. All subsequent analyses were built upon small RNA counts after normalization according to the miRNA amounts, as described in [34].

2.7. Genome Assembly

Raw nanopore reads were QC checked using Nanoplot v1.10.2 (https://github.com/wdecoster/NanoPlot) for sequencing run statistics. Reads with QC < 7 were removed by the sequence provider (Montpellier Genomix, Montpellier, France) before QC. For each dataset, mean length, N50 reads, total reads and bases are listed in Table S1 and Table 1. Reads were submitted to Flye v2.6 [40] with standard options, except –plasmids and -threads 16. Raw contigs were polished using four rounds of RACON v1.3.2 [41] with standard options and 20 threads (-t option; the required mapping was performed using minimap2 [42] v2.16 and -x map-ont -t 20 options). At each step, basic assembly metrics (N50, length, L50) were recorded using Assembly-Stats v1.0.1 (https://github.com/sanger-pathogens/assembly-stats). Once polished, assemblies were visually inspected using D-genies v1.2.0 [43], and incongruencies manually corrected using samtools v1.9.0 [44], faidx command for sequence extraction, and Gepard [45] v1.4.0 for visual determination of breaking points. The corrected assemblies underwent super scaffolding using RaGOO v1.1 [46] with -s (structural variants (SV)) and -t 4, using the specific reference genome (from FlyBase): Dmel_R6.23 for G0 and *D. melanogaster* samples, Dsim_r2.02 for *D. simulans*, and the previously assembled G0 for G73 and G0_F100 samples. Once the assembly was finalized at the chromosome scale, a Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis [47] using the gVolante web service [48] was performed using the BUSCO v2/v3 option and the *Arthropoda* reference set (Figure 1). TE content was estimated in the corresponding chromosome assemblies using RepeatMasker (http://www.repeatmasker.org) and the Dfam database [49].

Table 1. Statistics for the de novo assemblies before scaffolding. All lengths are expressed in bases. The Benchmarking Universal Single-Copy Orthologs (BUSCO) score indicates the "complete hit" level.

Name	Size	Nb contig	Mean Length	Longest	N50	L50	BUSCO Score, %
dmgoth101	130,483,042	1213	107,571	20,963,225	14,899,963	4	c: 98.6
dmgoth63	134,481,426	1005	133,812	22,615,553	16,996,519	4	c:98.03
dmsj23	131,331,777	1094	120,047	22,945,221	10,553,205	5	c:98.5
dmsj7	131,360,683	1197	109,742	18,094,419	6,212,683	7	c:98.7
dsgoth31	135,039,133	822	164,281	27,577,085	17,530,992	4	c: 98.3
dsgoth613	132,908,190	918	144,78	22,559,698	16,120,890	4	c:98.6
dssj27	134,309,820	866	155,092	27,370,717	20,976,825	3	c:98.6
dssj9	142,009,588	508	279,546	27,589,620	19,611,840	4	c:99
G0	127,415,251	642	198,466	5,037,957	1,208,862	33	c:93.7
G0-F100	139,374,117	836	166,715	17,781,420	9,085,947	6	c:98.97
G73	144,335,962	584	247,15	24,539,270	12,530,957	4	c:98.7

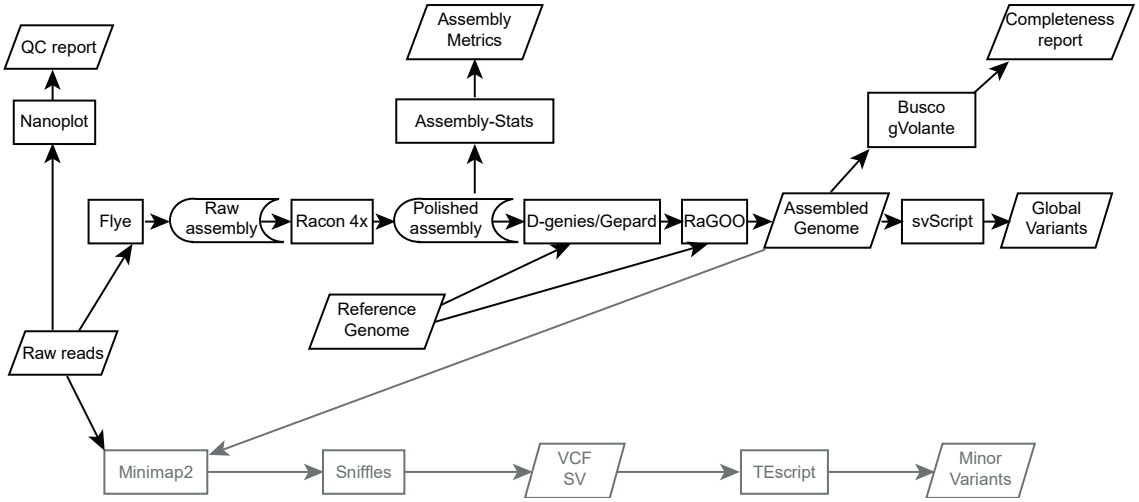


Figure 1. Schematic of the method used for genome assembly and for transposable element insertion (TEI) detection. Global variants (black) were detected from genome assemblies, and minor variants (gray) by remapping reads in these assemblies. The reference genomes used for RaGOO scaffolding were Dmel_R6.23 for G0 and for wild-type *D. melanogaster* strains, Dsim_R2.02 for wild-type *D. simulans* strains, and the G0 assembly for G73 and G0-F100.

2.8. Global Structural Variant Detection

Global variant detection (i.e., variants common to most genomes of a considered sample compared with the reference genome, see below) was performed using the svTEidentification.py tool (available at https://github.com/DrosophilaGenomeEvolution/TrEMOLO). Briefly, this tool recovers the insertion and deletion positions and creates the associated fasta sequence, based on the Assemblytics report from the RaGOO scaffolding (the deletions are extracted from the reference and the insertions from the new assembly). Once the fasta file corresponding to the SVs was recovered, these sequences were matched with the Basic Local Alignment Search Tool, nucleotide to nucleotide (BLASTN)+ v2.4.0 to a specified TE database. Hits larger than 80% of the TE sequence and identical to more than 80% at the nucleotide level were considered as candidate for new TE insertions/deletions (TEI/TED) in the G0, G0-F100 and G73 samples. For wild-type strains, new insertions/deletions were detected without any filter. The potential candidates were then listed in a tabular format that included their position, size and percentage of size or similarity compared with the reference TEs. The used TE database was a collection of the reference TEs from Bergman's laboratory (https://github.com/bergmanlab/transposons) and from previously published data [50].

2.9. LTR Minor Insertional Variant (LTR MIV) Detection

Each raw long read was mapped using minimap2 v2.16 (-ax map-ont -t 16 as options) to the assembly corresponding to that set of long reads. After recovering the sam file, samtools v1.10.0 was used to compress and sort the sam file in BAM with samtools view and samtools sort (basic options, but with 16 threads), and the MD tag was added using samtools calmd. Then, SV were detected in the resulting sorted BAM file using Sniffles v1.0.10 with at least 1 read and –report_seq -s 1 -n -1 as parameters [51]. These sequences longer than 1000 bp were aligned with BLASTN v2.4.0+ (-outfmt 6) to the LTR subset (60 families) of the database used before. A nucleotide alignment of more than 94% identity and a minimum of 90% of the total length of the TE consensus sequence were then considered as criteria to validate a putative LTR minor insertion variant (LTR MIV), if the length of the variant did not exceed the total size of the TE by more than 18 nt. This corresponds to the largest target site duplication (TSD) ever reported to flank any LTR TE [52]. All codes are available in a snakemake file at https://github.com/DrosophilaGenomeEvolution/TrEMOLO.

2.10. Fluorescent In Situ Hybridization on Polytene Chromosomes

Polytene chromosomes were squashed from salivary glands of third instar male larvae. *Not*I and *Pst*I restriction enzymes were used to extract a fragment of the ZAM *pol* gene from a previously published plasmid [53]. The probe was labeled with digoxigenin-11-dUTP using the Nick Translation Mix (Roche #11 745 816 910), and signals were detected with anti-digoxigenin-rhodamine Fab fragments (Roche). The fluorescent in situ hybridization method was adapted from a previously described protocol [54].

2.11. Automatic Identification of the Target Site Duplication for LTR MIV

The putative LTR MIVs matching to six LTR families (blood, gtwin, mdg3, ZAM, roo, and copia) were studied. One read supporting each MIV, previously extracted in a fasta file, was compared by BLASTN v2.4.0+ with the corresponding consensus sequence. To automatically check for the presence of a TSD, the positions of the 5′ and 3′ end of the TE alignment were determined within the read. 30nt-long sequences upstream and downstream the putative insertion site were extracted and were aligned to detect the presence, on both sides of the insertion, of a short duplication, the size of which was previously reported by [55] for ZAM and by [52] for the other TEs. The resulting TSD sequences were then extracted and used to create sequence logos with WebLogo (https://github.com/WebLogo/weblogo). All scripts and codes for this automatic extraction are available at the project GitHub.

2.12. piRNA Cluster Identification in the Assembled Genomes

To determine the piRNA cluster localization in genome assemblies, a previous annotation of piRNA clusters in the *D. melanogaster* Dmel_R6.04 genome release was used [56]. The flanking genes for each of the 153 major piRNA clusters were identified, their sequence was extracted and mapped to the new reference using BLASTN to locate the limits of the corresponding piRNA clusters in the corresponding assemblies. When only a single gene could be used as border, the piRNA cluster length described in [56] was used to define the other border. Bona fide piRNAs were extracted from the previously published G0 small RNA-seq library [28], and from each of the small RNA-seq libraries presented here, as reads longer than 23 nt that do not map (bowtie –best) to sequences of other known small RNAs (downloaded from FlyBase [57] and MirBase [58]). These selected small RNA reads were then mapped to the corresponding assemblies using Bowtie 1.2.2 [59]. Bowtie parameters were selected to keep only reads that display unique alignments and <2 mismatches (–best -v 2 -m 1). The positions of uniquely mapped reads were determined in the assembly, and sequences with more than 500 reads were conserved and compared to the piRNA cluster coordinates determined in the assembly of that line. Table S4 shows the list of the 42 piRNA clusters corresponding to the best piRNA producers in the G0 line. The coordinates of these 42 regions were then determined in the G73 and G0-F100 assemblies. For wild-type strains, the piRNA abundance was computed within 1 kb windows.

2.13. Comparison of ZAM Sequences

After obtaining the corresponding region of the ZAM insertions the fasta sequence was extracted (using bedtools getfasta) and compared with the ZAM sequence at a global level using redotable v1.1.

3. Results and Discussion

3.1. Using Oxford Nanopore Technologies (ONT) to De Novo Assemble the Highly Contiguous Genomes of Several Isogenic Wild-Type Strains and of one Unstable Line

The ONT-based single-molecule long-read sequencing data provided between 5 and 24 million reads, with a depth of coverage ranging from 40x to 196x (mean = 130x), and a N50 ranging from 3.7 to almost 20 kb (mean = 11 kb) (QC 7 reads only; Table S1). The N50 large range was explained by the different methods used for genomic DNA extraction and ligation (Materials and Methods). Our assembled genome procedure is summarized in Figure 1. To compare our data with the reference

D. melanogaster and *D. simulans* genomes, whole genome alignments and local dot plots were performed using D-genies and Gepard, respectively (Figure S1).

A strong correspondence was observed between most de novo assemblies and the corresponding reference genome, except for the G73 and dsgoth31 assemblies in which incongruent contigs were detected. These incongruent contigs were manually broken at the discrepancy points (Figure S1) and the final statistics for the de novo assemblies were obtained using Assembly-Stats (Table 1).

Using our approach based only on ONT data, the N50 ranged from 1.2 Mb (L50 of 33 contigs) to 21 Mb (L50 of 3 contigs). The previously described de novo *D. melanogaster* hybrid assembly obtained using BioNano and assembly merging [23] reported a N50 of 9 Mb (L50 of 6 contigs) for the raw data, and a N50 of 21.3 Mb (L50 of 3 contigs) after merging. Moreover, the BUSCO score of their hybrid assembly was 97.2% after Illumina polishing, while the BUSCO score of our assemblies ranged from 93.7% to almost 99% (98.5% for the reference Dmel_R6.23 assembly [23]) only with RACON polishing. This comparison indicates that our assemblies are of high quality, and that RaGOO use as scaffolder allowed obtaining high-quality assemblies at the chromosome scale.

3.2. Estimation of Genome size Using Different Methods

To determine the quality of the ONT-based assemblies of the isogenic wild-type *D. melanogaster* and *D. simulans* genomes, their sizes were compared to the genome sizes estimated with two other approaches: findGSE (based on k-mer estimation) and flow cytometry (Table S2).

Genome size estimates varied between 142 and 144 Mb (flow cytometry) and 129 and 132 Mb (findGSE) for the *D. simulans* strains and between 162 and 163 Mb (flow cytometry) and 133 and 137 Mb (findGSE) for the *D. melanogaster* strains after excluding dmsj7. The k-mer distribution obtained for this strain was much more scattered than the others, and resulted in a k-mer-based genome size estimate of 147 Mb, most probably an artefact. The size estimated obtained using the ONT data ranged between 131 and 142 Mb for the wild-type *D. simulans* strains and between 130 and 134 Mb for the *D. melanogaster* strains, with similar values for the final assemblies. The correlation coefficients were significant only between the ONT-based and the flow cytometry estimates for *D. melanogaster* ($r = 0.9675$, $p = 0.0325$), but not *D. simulans* (flow cytometry: $r = 0.7564$, $p = 0.2436$; findGSE: $r = 0.1237$, $p = 0.8763$). The correlation only with the flow cytometry estimate indicates that the different genome compositions, and probably the different amounts of heterochromatin affect the estimations obtained by findGSE. The genome size estimates obtained with findGSE were globally more similar than those obtained using the de novo assembly approach, but no correlation was observed between these values, probably due to the different amounts of repeats present in the various strains. In conclusion, genome size estimations present several biases in function of the used method, and ONT assemblies seem to give values close to those obtained by flow cytometry, which is a more global method.

3.3. Comparison of TE Abundance in the Isogenic Wild-Type Strains Measured by Illumina and ONT Sequencing

To validate the ONT approach, the TE abundance in the isogenic wild-type *D. melanogaster* and *D. simulans* strains was evaluated using dnaPipeTE [11] and TEcount [33] for Illumina sequencing data, and RepeatMasker for ONT assembled chromosomes (Figure 2). Overall, TE content (expressed as genome percentage) was often higher when estimated using dnaPipeTE (Illumina data) (Wilcoxon matched-pairs signed rank test; $p = 0.0234$) than with RepeatMasker (ONT assemblies) (Wilcoxon matched-pairs signed rank test; $p = 0.0156$). This might be explained by the fact that unlike the RepeatMasker TE database, dnaPipeTE is based on the de novo detection of TEs and the local assembly of TE families, independently of a previously annotated reference genome, thus recovering the maximum number of reads that correspond to known and unknown TEs. In agreement, the correlation was higher between the results obtained with RepeatMasker (ONT data) and the results obtained with TEcount, which is based on the read similarity against a curated database of known TEs [34] ($r = 0.8921$, $p < 0.0001$), than with dnaPipeTE ($r = 0.8504$, $p < 0.0001$) (Figure 2, right panel). As previously reported,

the LTR group was more abundant than the LINE and DNA transposon groups in all *Drosophila* genomes (see [60] for a review).

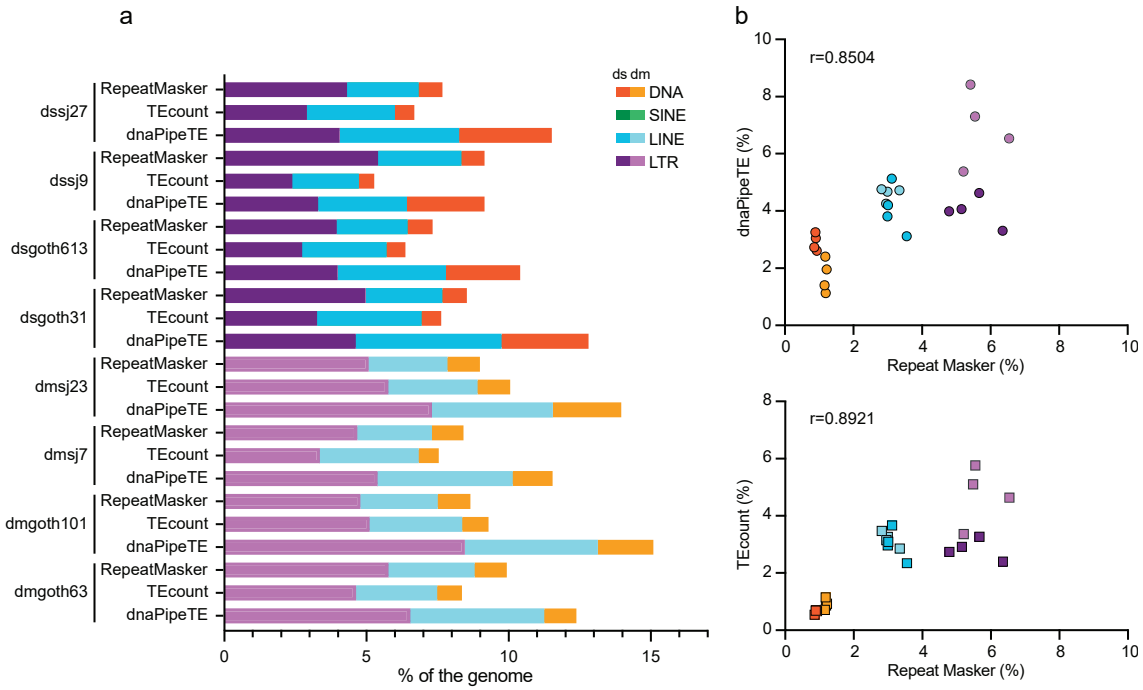


Figure 2. Estimation of the TE percentage in the *D. melanogaster* and *D. simulans* genomes (isogenic wild-type strains). (**a**) Estimation of the TE percentage using RepeatMasker (ONT chromosome assemblies), and dnaPipeTE or TEcount (Illumina reads). (**b**) Correlations between the estimates obtained with the indicated methods.

3.4. Comparison of the TEI Sites Identified in the Isogenic Wild-Type Strains Using the Illumina and ONT Data

Before focusing on the results provided by the ONT approach, we first compared these data to the classically used Illumina results based on discordant pairs of reads (method developed in the laboratory, see Material and Methods). The number of TEI sites tended to be higher when using the Illumina data than ONT data (Wilcoxon paired test, p-value = 0.023). This could be due to the presence of false positives caused by PCR artefacts during the Illumina library preparation [18], and/or to the fact that some TEIs might have been too short (fragmented or partially deleted) to be identified using the assembled ONT data. Using the Illumina approach, TEI numbers were significantly lower in the *D. simulans* than in the *D. melanogaster* strains (Wilcoxon test, p-value = 0.029), but not when using the ONT data (Wilcoxon test, p-value = 0.343) (Figure 3). This may reflect a bias towards *D. melanogaster* sequences in our TE reference file, and/or a long-term difference in TE dynamics between these species [25,61]. Comparisons (chi-square tests) of TEI distributions across TE groups (DNA, LINE, LTR) (see Table S3) showed that in *D. simulans*, the distributions obtained using both approaches were similar. Conversely, in *D. melanogaster*, the TEI number for retrotransposons was significantly higher relative to the other groups, when using the Illumina approach. This may be due to the higher propensity of *D. melanogaster* retrotransposons to be involved in Illumina PCR chimeras [18] because of their higher genome occupancy (Figure 2), and this difference may be amplified by the exponential behavior of the PCR reaction.

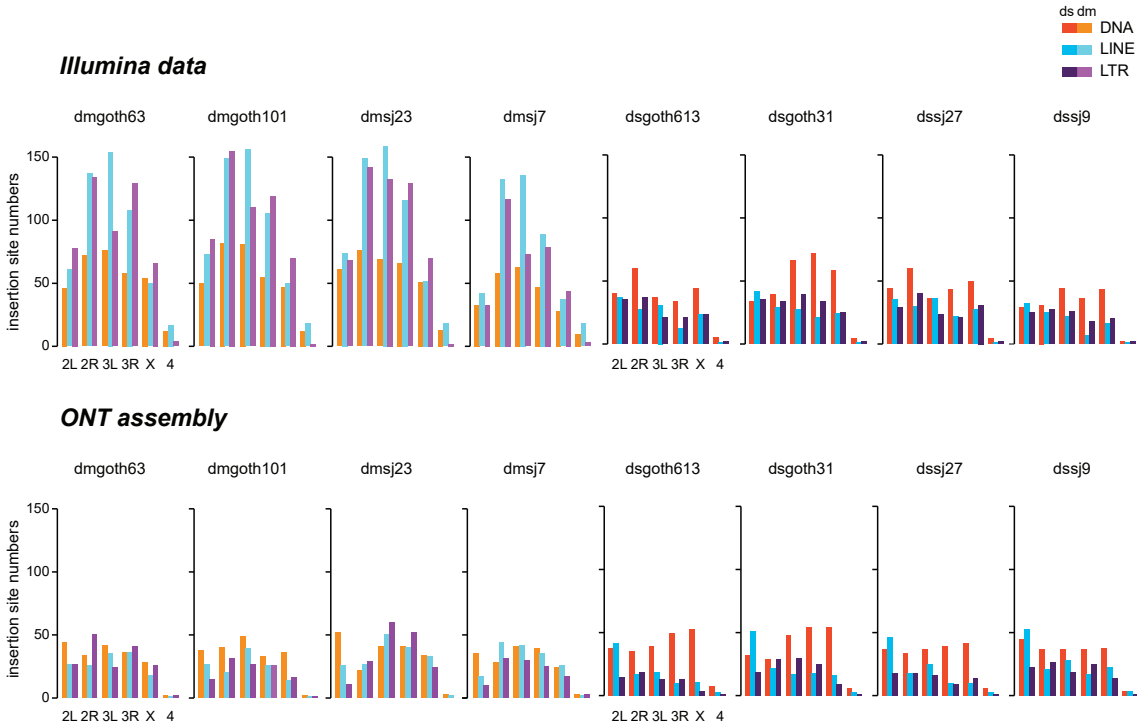


Figure 3. Insertion site numbers for each TE group and per chromosome, determined using Illumina data (upper panels) or Oxford Nanopore Technology (ONT) chromosome assemblies (lower panels).

In the subsequent analyses, only TEIs identified using the ONT approach (i.e., the most reliable set of recent insertions) were considered.

3.5. TEI Landscape in the Isogenic Wild-Type Strains

Using the ONT approach, the de novo genome assembly of each wild-type strain was compared with the reference genome and the detected insertional structural variants were called global variants (see Figure 1). These global variants correspond to the most recent TEIs. On average, there were 492 and 456 global variants in *D. melanogaster* and *D. simulans*, respectively (Table 2).

Table 2. Number of transposable elements insertions (TEIs) identified as global variants in the Oxford Nanopore Technology (ONT) chromosome assemblies.

	dmgoth63	**dmgoth101**	**dmsj23**	**dmsj7**	**dsgoth613**	**dsgoth31**	**dssj27**	**dssj9**
Total Insertion Number	515	448	550	456	434	496	420	474

DNA transposons were the most abundant group in both species (188 and 215 copies, on average, in *D. melanogaster* and *D. simulans*, respectively), and LTR retrotransposons the least abundant (147 and 117 copies, on average, in *D. melanogaster* and *D. simulans*, respectively). These results may seem in contradiction with the previous data on genome occupancy. However, in this analysis only recent insertions were considered. Moreover, as DNA transposons are in general smaller than LTR retrotransposons, similar levels of genome occupancy correspond to higher copy numbers for DNA transposons than for LTR retrotransposons.

Comparison of the locations of the insertions identified in the chromosome assemblies showed that 22 global variants were present in all four *D. melanogaster* strains, and 23 in all four *D. simulans* strains. These were mainly DNA transposons (n = 9 and n = 10, respectively). The number of shared pairwise global variants was rather low, roughly 10% of all insertions in most comparisons (Figure 4a).

D. simulans strains appeared equally distant in terms of insertion sites. Conversely, a geographical structuring could be observed in the *D melanogaster* comparisons: strains from the same population shared more insertion sites than strains from distinct populations.

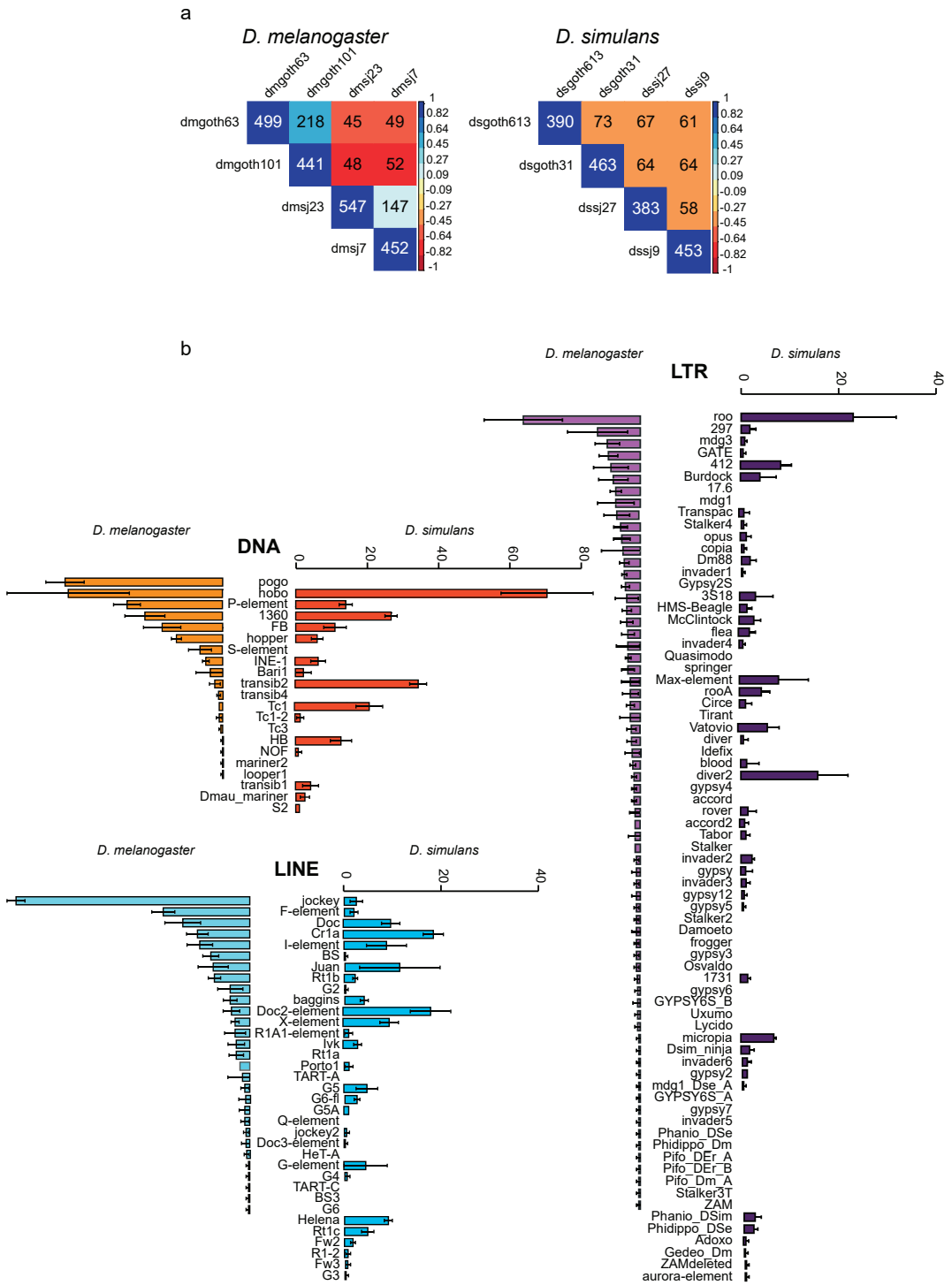


Figure 4. Global variant copy numbers in wild-type *D melanogaster* and *D. simulans* strains. (**a**) Number of shared global variants among strains. The color scale (on the right of each panel) shows the distance based on the number of pairwise shared insertions (indicated in black in the figure). Values in white correspond to the total numbers of the identified insertions for the considered strains. (**b**) Mean TEI numbers for the indicated TE groups computed in the wild-type *D melanogaster* and *D. simulans* strains based on the ONT chromosome assemblies.

The mean copy numbers for the different TE families were weakly correlated between *D. melanogaster* and *D. simulans* (Figure 4b) (Spearman rho = 0.33, *p*-value = 1e-4, across 129 TE families). Few families were found in the *D. simulans* strains but not in the *D. melanogaster* strains, and vice versa. In *D.* melanogaster strains, the most abundant families were roo (mean copy number: 24.00), jockey (mean copy number: 48.00), and pogo (mean copy number: 44.25), for LTR retrotransposons, LINEs, and DNA elements, respectively. In *D. simulans*, they were roo (mean copy number: 23.00), Cr1a (mean copy number: 18.50), and hobo (mean copy number: 70.50). In addition, some TE families displayed different copy numbers across strains. For instance, the 297 family had 18 copies in dmgoth63, 6 in dmgoth101, 6 in dmsj23, and 5 in dmsj7. Such patterns are suggestive of recent, independent activations, or even bursts of some families in specific strains, as suggested by in situ hybridization studies in a large number of samples [62]. Kofler et al. (2015) studied TE patterns in *D. melanogaster* and *D. simulans* field samples using Illumina pool-seq data [63]. By computing the insertion frequencies for each family of a subset of 121 TE families, they established that LTR elements were more frequent in *D. melanogaster* than in *D. simulans* populations, whereas DNA transposons were more frequent in *D. simulans* samples. A similar trend was observed in the present work: 147 LTR retrotransposon insertions in *D. melanogaster* and 117 in *D. simulans* (Wilcoxon test *p*-value = 0.343); 188 DNA transposon insertions in *D. melanogaster* and 215 in *D. simulans* (Wilcoxon test *p*-value = 0.029).

3.6. Comparison of TE Dynamics in Isogenic Wild-Type D. Melanogaster and D. Simulans Strains by Studying TEI Sequences in ONT Assemblies

The major advantage of the ONT approach is its ability to retrieve whole TEI sequences, while short read-based approaches only give access to TE insertion sites. First, the TEI sizes across strains were compared by parsing the BLAST results at the insertion level and by computing the insertion lengths (Figure 5a). The mean insertion lengths (i.e., fragment sizes) significantly varied among TE groups (2-way ANOVA, *p*-value = 2e-81), but not between species (2-way ANOVA, *p*-value = 0.22). LTR retrotransposons were the largest (mean size = 2692 bp), followed by LINEs (mean size = 1290 bp), and DNA transposons (mean size = 1210 bp). The observed absence of difference between species in these global variants differs from what was previously described. Indeed, for a subset of 15 families, Lerat et al. found that TE copies were more internally deleted (i.e., shorter) in *D. simulans* than in *D. melanogaster* [24]. However, analysis of these 15 families using our ONT data indicated that they displayed, on average, longer fragment sizes compared with the other TE families in *D. melanogaster* (Wilcoxon test, *p*-value = 8e-19), but not in *D. simulans* (Wilcoxon test, *p*-value = 0.34) [24]. This suggests that Lerat et al. 2011 focused on TE families that have particularly large copies in *D. melanogaster* [24], probably because they have been more studied in the past due to their easier analysis by in situ hybridization on polytene chromosomes [7,25,64].

Then, the Refiner module of RepeatModeler (http://www.repeatmasker.org/RepeatModeler) was used to compute the intra-family sequence divergence (average Kimura distance) (Figure 5b). This measure is a proxy of the time passed since the last transposition wave(s). Overall, these distributions were not significantly different between *D. melanogaster* and *D. simulans* and among TE groups (2-way ANOVA; species effect, *p*-value = 0.151; group effect, *p*-value = 0.701), showing that the TE recent dynamics are similar in these two species. However, in *D. simulans,* DNA transposons displayed significantly higher intra-family divergence compared with LTR retrotransposons (Wilcoxon test, *p*-value = 0.023). This suggests that among the most recent transposition events, DNA transposon insertions occurred slightly less recently in *D. simulans*.

Kofler et al. 2015 assumed that population frequencies of TE insertions provide an estimator for the insertion age. However, we find that their population frequencies were not correlated with our measures of intra-family sequence divergence (Spearman correlation coefficients: −0.714 (*p*-value = 0.136) and 0.116 (*p*-value = 0.827) for *D. melanogaster* and *D. simulans*, respectively). We think that intra-family sequence divergence is a more direct estimate of the age of transposition events; however, this discrepancy may also reflect differences in the origins of the sampled flies [61,64]. Alternatively,

it may suggest that other factors influence insertion frequencies, besides the age since the initial transposition burst. In addition, our analysis only included TEIs that are not found in the reference genome, i.e., TEIs that result from transposition events more recent than the set-up of the actual populations. Altogether, while the TE ancient dynamics are different between *D. melanogaster* and *D. simulans* [60], the present results suggest that *D. melanogaster* and *D. simulans* TE landscapes are rather similar when comparing only global variants (i.e., the subset of the most recent insertions). As already proposed [25], this may reveal that the colonization of *D. simulans* genome by TEs has now reached a state similar to that of *D. melanogaster*, although it started more recently.

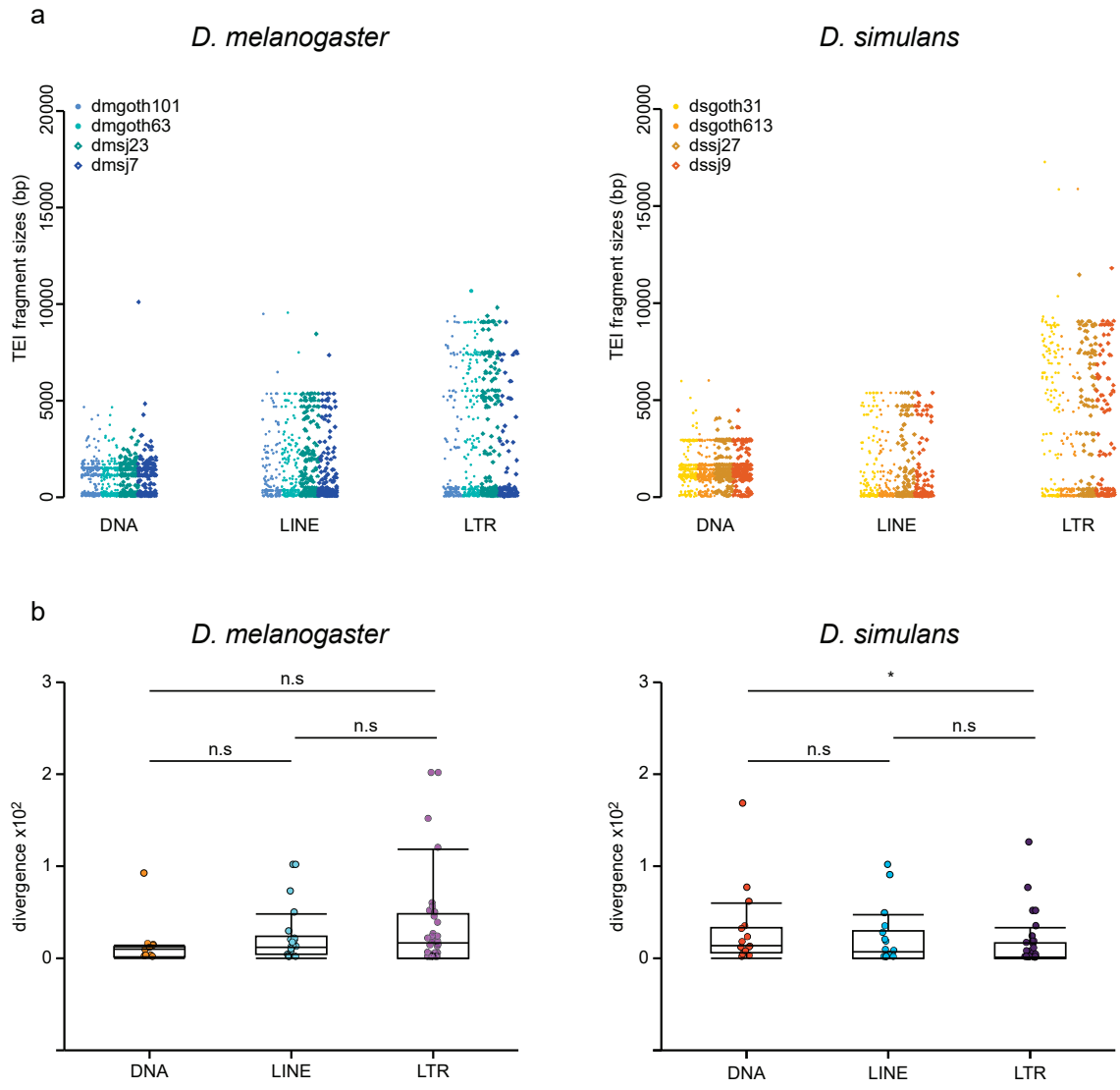


Figure 5. Global variant sequence analysis in wild-type *D. melanogaster* and *D. simulans* strains. (**a**) Distributions of TE copy lengths (i.e., fragment size) in bp for all global variants across strains and TE groups. (**b**) Intra-family sequence divergence (average Kimura distance) computed per strain and per TE family.

3.7. piRNAs, piRNA Clusters and TEIs in Isogenic Wild-Type Strains

Another way to study TE dynamics is to understand the way the production of piRNAs is linked to the TEI type and structure. Indeed, some relationships might exist between piRNA abundance and the recent activity of TEs, estimated by the intra-family sequence divergence. Therefore, piRNA production, TE length and intra-family sequence divergence were analyzed for each TE

group and strain. This analysis highlighted a significant TE group effect: piRNA counts were higher in retrotransposon families (LTR elements and LINEs) than in DNA transposon families (p-value = 2e-9). Moreover, piRNA counts were significantly and positively correlated with genome occupancy (p-value = 5e-7), which strongly depends on TE copy number (Figure 6a). The hypothesis that TE copy numbers determine piRNA abundance was previously suggested in *D. melanogaster* [65,66] and is confirmed here also for *D. simulans*. However, it should be noted that genome occupancy accounts only for 6.2% of the total variation of piRNA counts, indicating that many other factors are involved in TE control.

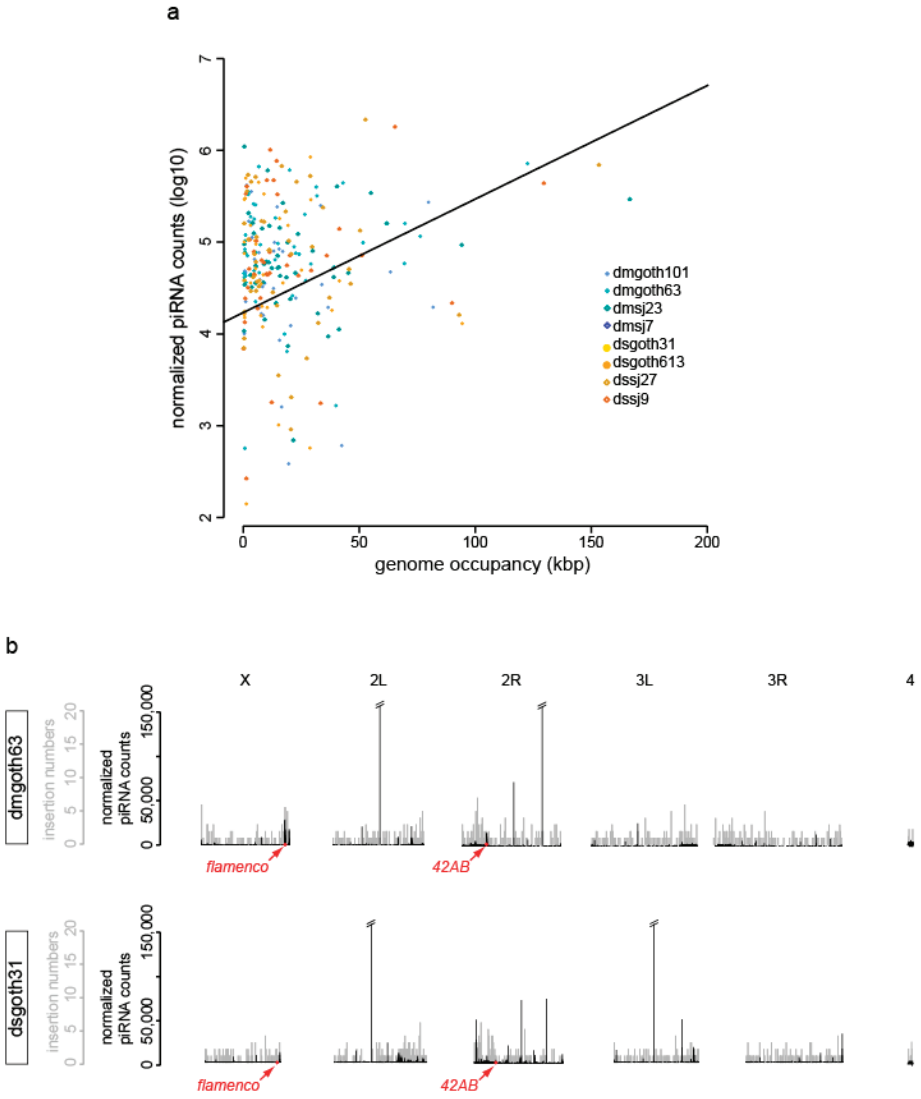


Figure 6. piRNA analyses in wild-type *D. melanogaster* and *D. simulans* strains. (**a**) Normalized piRNA counts (log10) relative to genome occupancy for all strains and the two species and linear regression curve. Each dot is a TE family. (**b**) Results for the dmgoth63 and dsgoth31 strains are shown as examples. Uniquely mapping piRNAs along ONT chromosome assemblies (black, normalized piRNA counts). Global variants identified along ONT chromosome assemblies (gray). Red arrows indicate flamenco (X chromosome) and 42AB (2R chromosome). Data for the other strains are provided in Figure S2. The off-scale peaks might correspond to microRNAs that are absent from miRBase.

These observations are also in agreement with the idea that newly integrated copies become piRNA producers [67], and that longer copies produce more piRNAs. It should be noted that retrotransposons are on average longer than DNA transposons.

As ONT assemblies also include piRNA cluster sequences, 42AB and flamenco (the two major piRNA cluster producers in *D. melanogaster*) could be retrieved using their flanking genes (see Material and Methods) [68] from each assembly. Alignment of the uniquely mapped piRNA sequences against the assembly of each wild-type isogenic strain (Figure 6b and Figure S2, black lines) indicated that the regions corresponding to 42AB and flamenco did not display any enrichment in global variant insertion numbers (Figure 6b, gray lines). This indicates that recent TEIs are not specifically enriched in the two major piRNA cluster producers in *D. melanogaster* and *D. simulans* strains. Therefore, the analysis of the de novo assembled genomes to follow the piRNA cluster dynamics in these isogenic wild-type strains did not highlight the previously reported high TEI insertion rate within piRNA clusters [26,50,69,70]. Our data suggests the number of recent TEIs fixed in these piRNA clusters is not different compared with anywhere else in the genome. This discrepancy could be explained by the high frequency of deletions (from several base pairs up to several kilobases) that seems to occur in these regions and that affect ancient TEs, which remain as vestiges in these loci, and also recently inserted TEs [50].

3.8. Recent TEIs May Not Be Frequent Enough to Be Incorporated in the Assembled Genomes

To challenge the ONT assembly approach, a bioinformatic analysis was performed to identify recent LTR TEIs that occurred during the last 73 generations (G73) in the unstable Piwi KD line (Materials and Methods and [28]). As a control, to estimate the basal transposition rate when TEs are normally repressed by the functional piRNA pathway, the genome of the hundredth generation (called G0-F100) after establishment of the stable G0 isofemale line was also sequenced. Using the pipeline for detection of global variants (Figure 1), no new ZAM insertion could be detected in the G73 assembled genome compared with the G0 reference genome. This is not consistent with previous data obtained by PCR quantification of the ZAM copy number [28]. Therefore, in situ hybridization analysis was performed to determine whether de novo ZAM insertions were present on polytene chromosomes of G73 male larvae (Figure 7a and Figure S3). This analysis confirmed the presence of the two preexisting ZAM insertions identified on chromosome 2R as global variants in the G0 de novo assembly (compared with the Dmel_R6.23 reference genome). These two insertions were also detected in all three G73 larvae analyzed, as well as many other ZAM signals that were not observed in the G0 samples (Figure 7a and Figure S3). As each of these many G73-specific new ZAM insertions was present in a single larva, they were not incorporated in the G73 de novo assembled genome due to their low frequency, and therefore could not be detected as global variants. Based on the G0 assembled genome, the sequences of the two shared ZAM detected by FISH on chromosome 2R could be accessed. One contained the full length canonical ZAM consensus sequence, while the other displayed an internal deletion (Figure 7b).

3.9. A Long Read-Based Pipeline to Detect Low Frequency TEI Polymorphisms

To determine whether ONT can be used to detect TEIs with a frequency not high enough to be recovered in the assembled haplotype, an approach to identify "minor insertional variants" (MIV) was developed (Material and Methods, paragraph 2.9, and Figure 1 (gray)). Minimap2 was used to map each individual long read to the corresponding assembled genome, and Sniffles to obtain the list of variants that had been neglected during the assembling process. Some of the sequences identified as MIVs matched to the 60 canonical LTR TE consensus sequences (Materials and Methods).

As expected, very few LTR MIVs were detected in the G0-F100 "stable line". Only copia and roo, which have high transposition rates [71], exhibited more than four variants (14 and 22, respectively) among the 51 LTR MIVs detected (Figure 7c). Also in the G73 line, copia and roo were among the more active LTR families (35 and 48 LTR MIVs among the 274 LTR MIVs detected) (Figure 7c). However, two other LTR families, ZAM and gtwin (51 and 93 LTR MIVs, respectively), showed a 50-fold increase in G73 compared with G0-F100, which is more than an order of magnitude higher than what observed for any other LTR family.

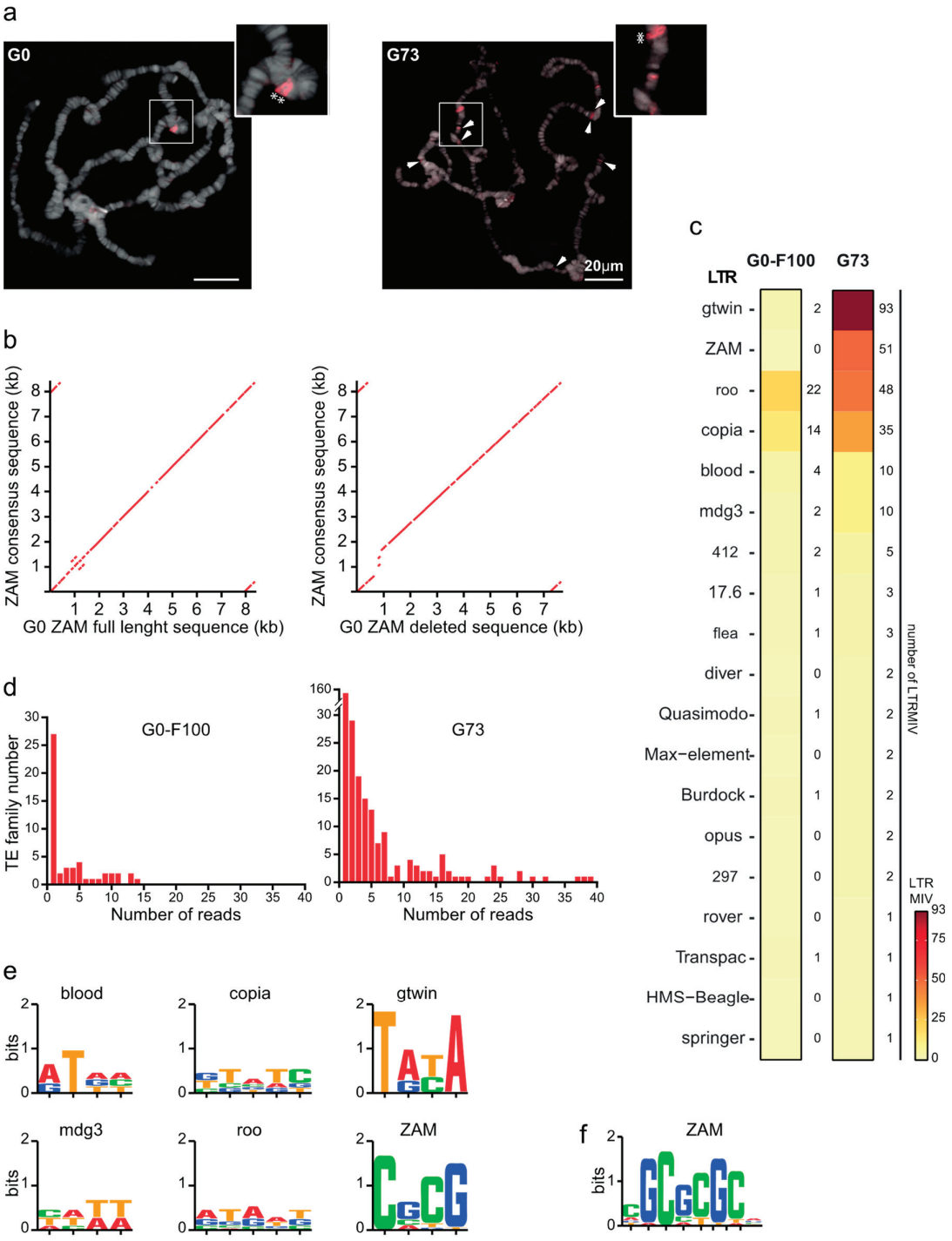


Figure 7. Characterization of the Long-Terminal Repeat minor insertion variant (LTR MIV) in the stable (G0) and unstable (G73) lines. (**a**) ZAM copies visualized by fluorescent in situ hybridization in G0 (left) and G73 (right) polytene chromosomes. The two global variants correspond to non-reference ZAM copies present in G0 and G73 (asterisks in the zoomed images). Arrowheads show the new ZAM insertions in G73. More examples are presented in Figure S3. (**b**) Dot plot of the sequence comparison between the ZAM sequences accessed from the de novo assembled G0 genome and the ZAM consensus sequence. (**c**) Heat map of the LTR MIV detected in the G0-F100 (stable) and G73 (unstable) libraries. (**d**) Histograms showing the number of reads supporting each LTR MIV. (**e**) Sequence logo of TSD defined using the LTR MIV automatic detection procedure. (**f**) the ZAM TSM motif defined using the automatic and manual LTR MIV detection procedures.

The next question was to determine whether the 274 LTR MIVs, present at low frequency in G73, had occurred after the establishment of the isofemale line. Indeed, such insertions could have been already present in G0 at high frequency (and therefore, could have been incorporated in the G0 but not in the G73 assembled genome) or at low frequency (and, therefore, detectable only as MIVs in G0). The first hypothesis was ruled out by comparing global deletions in G73 and G0. Very few G0 insertions were lost in the G73 assembly and they all belonged to five LTR families (mdg3, Transpac, 3S18, blood, and driver) that did not show a large MIV increase in G73 (data not shown). The total absence of LTR MIVs in G0 was not in favor of the second hypothesis.

As a large fraction of the 274 LTR MIVs in G73 were supported by a single read (Figure 7d), the next step was to check whether they were bona fide insertions by looking for insertional hallmarks, such as the target site duplications (TSDs) that occur upon integration as a result of staggered double-strand breaks at this site [72]. Flanking duplications were first detected automatically for each of the top six LTR families (mdg3, blood, copia, roo, ZAM, and gtwin) by aligning the two 30nt-long sequences that flank each putative LTR MIV extracted from the read supporting the variant. This analysis showed that depending on the LTR family, 30–80% of MIVs were flanked by a short duplication of the expected size (4 or 5 nt) (Table 3) [52]. The TSD consensus sequences identified are presented in Figure 7e.

Table 3. Target site duplication (TSD) flanking Long-Terminal Repeat minor insertion variants (LTR MIVs) in the G73 line.

	LTR Family					
	gtwin	**roo**	**ZAM**	**copia**	**Blood**	**mdg3**
Total LTR MIV detected (*n*)	93	48	51	35	10	10
TSD automatic detection (*n*)	66	15	25	11	8	5
TSD automatic detection (%)	71	31	49	31	80	50
Additional TSD manually detected (*n*)	NA	NA	23	NA	NA	NA

The failure to automatically detect a TSD for the other LTR MIVs could be due to the frequent sequencing errors, a known ONT drawback. When located in the genome-LTR junction region, such errors, which may include several nt-long indels, could impair the automatic detection of the expected TSD, as shown in Figure S4 for the manual inspection of the 2R-33863 putative ZAM insertion. Even when junctions are correctly determined, a simple sequencing error in one of the duplicated sequences might prevent their perfect matching. However, it was possible to correct the errors present in these single reads by aligning them with the empty genomic target present on the assembled genome (see, Figure S4). Using this method to manually inspect the sequence of all 51 ZAM variant reads, 48 bona fide insertions were identified, as judged by the presence of the expected 4-nt TSD included in a palindromic GC-rich 6-nt target site motif (TSM) (Figure 7f) [52,55].

Therefore, despite ONT low sequencing accuracy, LTR MIVs could be detected with high sensitivity (insertions present in a population at a frequency <1%, because detected as single reads in a 197x average coverage library) and specificity (FDR of 3/51 = 6%).

3.10. Invading LTR Elements Are Not Preferentially Trapped by piRNA Clusters

It is widely assumed that a TE invasion is stopped when a member of the TE family jumps into a piRNA cluster that then triggers the production of piRNAs to repress this TE family (i.e., trap model) [27]. Long-read sequencing data allowed determining whether new insertions accumulated in major piRNA source loci during the 73 generations of LTR TE derepression. Comparison of the 42 major piRNA clusters after their localization in the G0 and G73 assemblies (Table S4) did not highlight any new TEI into any of these piRNA clusters in the G73 assembled genome. However, new insertions that occurred during the 73 generations of piRNA pathway impairment could still segregate as MIVs in the G73 population. Indeed, among the 274 LTR MIVs present in G73, 6.57% (n = 18) were located within the 42 major piRNA producers (Figure 8). However, this proportion was very similar to that of the

piRNA cluster size relative to the total de novo assembled genome size (7.36%). Therefore, unlike what expected in the trap model, LTR retrotransposons do not seem to have preferentially accumulated in piRNA clusters during the 73 generations of transposition burst. Specifically, assuming a binomial law with $n = 274$ and $p = 0.0736$ and using a one-tailed test, more than 29 insertions (and not the 18 detected) belonging to many different TE families would have been necessary to validate the hypothesis that piRNA clusters are TE trappers (5% probability threshold).

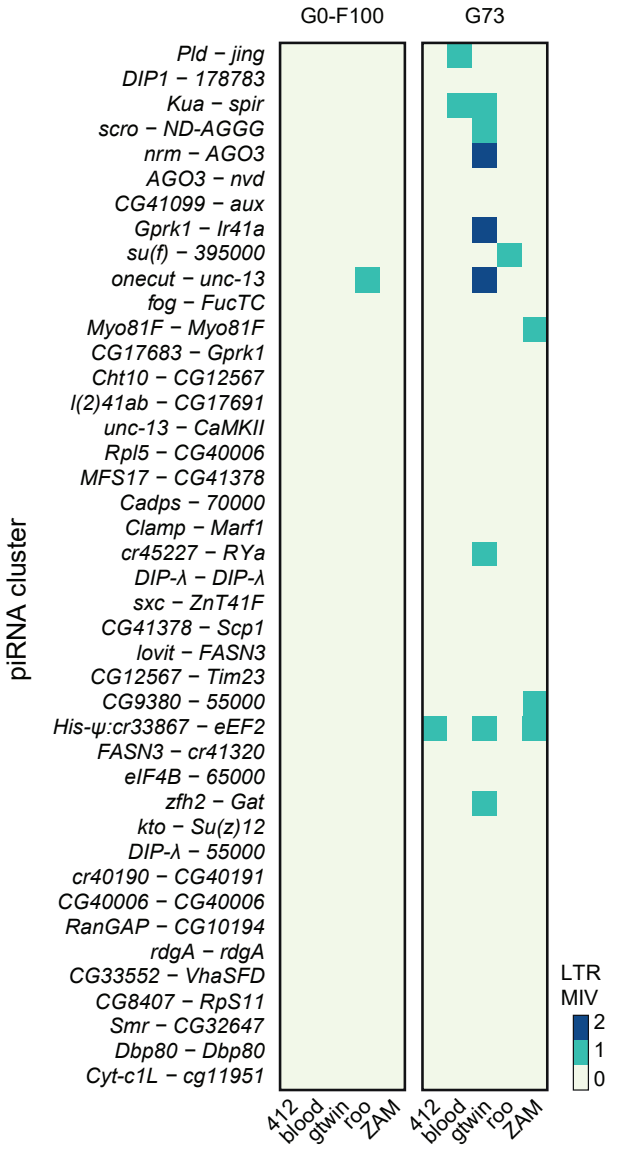


Figure 8. Heat map of the LTR MIVs inserted in piRNA clusters and detected in the G0-F100 and G73 lines.

More than 50% of the LTR MIVs located in piRNA clusters belonged to the gtwin family, suggesting that this family inserts preferentially into piRNA clusters. Indeed, among the 93 gtwin MIVs, 11 (11.8%) were found in piRNA clusters, which is very close to the minimal number ($n = 12$) required to reject the null hypothesis of random insertion in the genome (binomial law with $n = 93$, $p = 0.0736$, and 5% probability threshold). More data on de novo gtwin mobilization are needed to confirm their preferential integration in piRNA clusters during a transposition burst and to support the trap model for this TE family.

4. Conclusions

Our work demonstrates that long reads are crucial in order to finely describe TE landscapes at the intra-genome scale. Using isogenic wild-type strains and an unstable line with a succession of transposition bursts, we could characterize the most common TE variants in different strains and identify TE minor variants observed soon after transposition. The parallel analysis of two close species (*D. melanogaster* and *D. simulans*) and two genetic backgrounds allowed us to show that overall, TE recent dynamics are quite similar between species and among strains. However, there is still some strain specificity concerning the identity of the most recently active TE families. ONT is also a powerful tool to investigate the dynamics of piRNA clusters, which are in general inaccessible using short-read sequencing methods. We show here that recent TEIs are not enriched in piRNA clusters, despite recent bursts of TE transposition. Moreover, ONT allows detecting very recent TEIs that are sequenced as singleton reads.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4409/9/8/1776/s1, Figure S1: D-genies genome-wide dot plot of ONT assembly contigs versus reference genome, Figure S2: piRNA analyses in wild-type strains, Figure S3: ZAM copies were visualized by fluorescent in situ hybridization on G0 and G73, Figure S4: Alignments of the 2R-33863 insertion variant to the ZAM consensus sequence. Table S1: Statistics about sequencing data. All lengths are expressed in bases. Quality is expressed in standard Phred scale, Table S2: Genome size estimations using different methods, Table S3: Comparison of TEI distributions across TE groups using chi-square tests, Table S4: piRNA cluster coordinates based on flanking genes in de novo assembled genomes.

Author Contributions: Conceptualization, S.C.; Data curation, N.T.-M.D., M.M., F.S., Y.O., N.B., J.S.-O., D.S. and A.P.; Formal analysis, M.M., M.B., M.F., N.T.-M.D., F.S., A.P. and V.M.; Funding acquisition, C.V. and S.C.; Investigation, S.C., M.F.; C.V.; Methodology, P.V., A.P., F.S. and M.F.; Software, M.M., N.T.-M.D., P.V., F.S. and M.F.; Supervision, M.F. and S.C.; Visualization, B.M.; Writing—original draft, C.V., F.S. and S.C.; Writing—review & editing, C.V., F.S., M.F. and S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fondation pour la Recherche Médicale, grant number "DEQ20180339167" to S.C, by the ANR Exhyb to C.V., by the CNRS.

Acknowledgments: We thank J. Gonzalez and C. Goubert for the discussion, C. Jourdan and B. Barckmann for G0-F100, G73 and G0 DNA extraction, C. Brun for the polytene mapping, D. Gourion for the modelization and statistics in Section 3.10 and A.S. Fiston-Lavier for the discussions. We thank Ndomassi Tando and the IRD itrop "Plantes Santé" bioinformatic platform for providing HPC resources and support for our research project. T.-M.N.D. was supported by France Excellence. D.S. acknowledges financial support from France Génomique National infrastructure, funded as part of "Investissement d'avenir" program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09)."

Conflicts of Interest: The authors declare no competing interests.

Data Availability: Long reads sequencing data used for this study have been deposited at ENA (https://www.ebi.ac.uk/ena) under the accession numbers PRJEB39340 and ERP122844. The small RNA-seq datasets and the Illumina DNA-seq datasets were deposited in NCBI SRA (https://www.ncbi.nlm.nih.gov/sra) under the accession numbers PRJNA644327 and PRJNA644748, respectively. Release 6.23 of the *D. melanogaster* genome and Release 2.2 of the *D. Simulans* used in this study are available on FlyBase (http://www.flybase.org). Bioinformatic scripts and pipelines used for long reads analyses are available at https://github.com/DrosophilaGenomeEvolution/TrEMOLO and for small reads Illumina insertion at https://gitlab.in2p3.fr/pveber/te-insertion-detector/.

References

1. Biémont, C.; Vieira, C. Genetics: Junk DNA as an evolutionary force. *Nature* **2006**, *443*, 521–524. [CrossRef]
2. Wicker, T.; Sabot, F.; Hua-Van, A.; Bennetzen, J.L.; Capy, P.; Chalhoub, B.; Flavell, A.; Leroy, P.; Morgante, M.; Panaud, O.; et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **2007**, *8*, 973–982. [CrossRef] [PubMed]
3. Kapitonov, V.V.; Jurka, J. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.* **2008**, *9*, 411–412. [CrossRef] [PubMed]
4. Brookfield, J.F.; Montgomery, E.; Langley, C.H. Apparent absence of transposable elements related to the P elements of D. melanogaster in other species of Drosophila. *Nature* **1984**, *310*, 330–332. [CrossRef] [PubMed]
5. Black, D.M.; Jackson, M.S.; Kidwell, M.G.; Dover, G.A. KP elements repress P-induced hybrid dysgenesis in Drosophila melanogaster. *EMBO J.* **1987**, *6*, 4125–4135. [CrossRef] [PubMed]

6. Biémont, C.; Ronsseray, S.; Anxolabéhère, D.; Izaabel, H.; Gautier, C. Localization of P elements, copy number regulation, and cytotype determination in Drosophila melanogaster. *Genet. Res.* **1990**, *56*, 3–14. [CrossRef]
7. Biémont, C.; Monti-Dedieu, L.; Lemeunier, F. Detection of Transposable Elements in Drosophila Salivary Gland Polytene Chromosomes by In Situ Hybridization. In *Mobile Genetic Elements: Protocols and Genomic Applications*; Miller, W.J., Capy, P., Eds.; Methods in Molecular Biology; Humana Press: Totowa, NJ, USA, 2004; pp. 21–28, ISBN 978-1-59259-755-0.
8. Ignatenko, O.M.; Zakharenko, L.P.; Dorogova, N.V.; Fedorova, S.A. P elements and the determinants of hybrid dysgenesis have different dynamics of propagation in Drosophila melanogaster populations. *Genetica* **2015**, *143*, 751–759. [CrossRef]
9. Onder, B.S.; Kasap, O.E. P element activity and molecular structure in Drosophila melanogaster populations from Firtina Valley, Turkey. *J. Insect Sci. Online* **2014**, *14*, 16. [CrossRef]
10. Novák, P.; Neumann, P.; Macas, J. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinform.* **2010**, *11*, 378. [CrossRef]
11. Goubert, C.; Modolo, L.; Vieira, C.; ValienteMoro, C.; Mavingui, P.; Boulesteix, M. De Novo Assembly and Annotation of the Asian Tiger Mosquito (*Aedes albopictus*) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (*Aedes aegypti*). *Genome Biol. Evol.* **2015**, *7*, 1192–1205. [CrossRef]
12. Granzotto, A.; Lopes, F.R.; Lerat, E.; Vieira, C.; Carareto, C.M.A. The evolutionary dynamics of the Helena retrotransposon revealed by sequenced Drosophila genomes. *BMC Evol. Biol.* **2009**, *9*, 174. [CrossRef] [PubMed]
13. Rebollo, R.; Lerat, E.; Kleine, L.L.; Biémont, C.; Vieira, C. Losing helena: The extinction of a drosophila line-like element. *BMC Genom.* **2008**, *9*, 149. [CrossRef] [PubMed]
14. Schlötterer, C.; Tobler, R.; Kofler, R.; Nolte, V. Sequencing pools of individuals—Mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **2014**, *15*, 749–763. [CrossRef] [PubMed]
15. Rahman, R.; Chirn, G.; Kanodia, A.; Sytnikova, Y.A.; Brembs, B.; Bergman, C.M.; Lau, N.C. Unique transposon landscapes are pervasive across Drosophila melanogaster genomes. *Nucleic Acids Res.* **2015**, *43*, 10655–10672. [CrossRef] [PubMed]
16. Fiston-Lavier, A.-S.; Barrón, M.G.; Petrov, D.A.; González, J. T-lex2: Genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. *Nucleic Acids Res.* **2015**, *43*, e22. [CrossRef] [PubMed]
17. Kofler, R.; Gómez-Sánchez, D.; Schlötterer, C. PoPoolationTE2: Comparative population genomics of transposable elements using Pool-Seq. *Mol. Biol. Evol.* **2016**, *33*, 2759–2764. [CrossRef]
18. Treiber, C.D.; Waddell, S. Resolving the prevalence of somatic transposition in Drosophila. *eLife* **2017**, *6*, e28297. [CrossRef]
19. Pollard, M.O.; Gurdasani, D.; Mentzer, A.J.; Porter, T.; Sandhu, M.S. Long reads: Their purpose and place. *Hum. Mol. Genet.* **2018**, *27*, R234–R241. [CrossRef]
20. Lerat, E.; Goubert, C.; Guirao-Rico, S.; Merenciano, M.; Dufour, A.-B.; Vieira, C.; González, J. Population-specific dynamics and selection patterns of transposable element insertions in European natural populations. *Mol. Ecol.* **2019**, *28*, 1506–1522. [CrossRef]
21. Chakraborty, M.; VanKuren, N.W.; Zhao, R.; Zhang, X.; Kalsow, S.; Emerson, J.J. Hidden genetic variation shapes the structure of functional elements in Drosophila. *Nat. Genet.* **2018**, *50*, 20–25. [CrossRef]
22. Miller, D.E.; Staber, C.; Zeitlinger, J.; Hawley, R.S. Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing. *G3 Genes Genomes Genet.* **2018**, *8*, 3131–3141. [CrossRef]
23. Solares, E.A.; Chakraborty, M.; Miller, D.E.; Kalsow, S.; Hall, K.; Perera, A.G.; Emerson, J.J.; Hawley, R.S. Rapid Low-Cost Assembly of the Drosophila melanogaster Reference Genome Using Low-Coverage, Long-Read Sequencing. *G3 Genes Genomes Genet.* **2018**, *8*, 3143–3154. [CrossRef] [PubMed]
24. Lerat, E.; Burlet, N.; Biémont, C.; Vieira, C. Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. *Gene* **2011**, *473*, 100–109. [CrossRef] [PubMed]
25. Vieira, C.; Lepetit, D.; Dumont, S.; Biémont, C. Wake up of transposable elements following Drosophila simulans worldwide colonization. *Mol. Biol. Evol.* **1999**, *16*, 1251–1255. [CrossRef]
26. Bergman, C.M.; Quesneville, H.; Anxolabéhère, D.; Ashburner, M. Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. *Genome Biol.* **2006**, *7*, R112. [CrossRef]

27. Kofler, R. Dynamics of Transposable Element Invasions with piRNA Clusters. *Mol. Biol. Evol.* **2019**, *36*, 1457–1472. [CrossRef]
28. Barckmann, B.; El-Barouk, M.; Pélisson, A.; Mugat, B.; Li, B.; Franckhauser, C.; Fiston Lavier, A.-S.; Mirouze, M.; Fablet, M.; Chambeyron, S. The somatic piRNA pathway controls germline transposition over generations. *Nucleic Acids Res.* **2018**, *46*, 9524–9536. [CrossRef]
29. Romero-Soriano, V.; Burlet, N.; Vela, D.; Fontdevila, A.; Vieira, C.; García Guerreiro, M.P. Drosophila Females Undergo Genome Expansion after Interspecific Hybridization. *Genome Biol. Evol.* **2016**, *8*, 556–561. [CrossRef]
30. Sun, H.; Ding, J.; Piednoël, M.; Schneeberger, K. findGSE: Estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics* **2018**, *34*, 550–557. [CrossRef]
31. Di Giovanni, D.; Lepetit, D.; Guinet, B.; Bennetot, B.; Boulesteix, M.; Couté, Y.; Bouchez, O.; Ravallec, M.; Varaldi, J. A behavior-manipulating virus relative as a source of adaptive genes for Drosophila parasitoids. *Mol. Biol. Evol.* **2020**. [CrossRef]
32. Ballard, J.W.O. Comparative Genomics of Mitochondrial DNA in Drosophila simulans. *J. Mol. Evol.* **2000**, *51*, 64–75. [CrossRef] [PubMed]
33. Lerat, E.; Fablet, M.; Modolo, L.; Lopez-Maestre, H.; Vieira, C. TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids Res.* **2017**, *45*, e17. [CrossRef] [PubMed]
34. Roy, M.; Viginier, B.; Saint-Michel, É.; Arnaud, F.; Ratinier, M.; Fablet, M. Viral infection impacts transposable element transcript amounts in Drosophila. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 12249–12257. [CrossRef] [PubMed]
35. Langmead, B.; Wilks, C.; Antonescu, V.; Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinforma. Oxf. Engl.* **2019**, *35*, 421–432. [CrossRef] [PubMed]
36. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef] [PubMed]
37. Grentzinger, T.; Armenise, C.; Brun, C.; Mugat, B.; Serrano, V.; Pelisson, A.; Chambeyron, S. piRNA-mediated transgenerational inheritance of an acquired trait. *Genome Res.* **2012**, *22*, 1877–1888. [CrossRef]
38. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **2011**, *17*, 10–12. [CrossRef]
39. Schmieder, R.; Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **2011**, *27*, 863–864. [CrossRef]
40. Kolmogorov, M.; Yuan, J.; Lin, Y.; Pevzner, P.A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **2019**, *37*, 540–546. [CrossRef]
41. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **2013**, arXiv:1303.3997 (q-bio).
42. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *34*, 3094–3100. [CrossRef] [PubMed]
43. Cabanettes, F.; Klopp, C. D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **2018**, *6*, e4958. [CrossRef] [PubMed]
44. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
45. Krumsiek, J.; Arnold, R.; Rattei, T. Gepard: A rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **2007**, *23*, 1026–1028. [CrossRef]
46. Alonge, M.; Soyk, S.; Ramakrishnan, S.; Wang, X.; Goodwin, S.; Sedlazeck, F.J.; Lippman, Z.B.; Schatz, M.C. RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **2019**, *20*, 224. [CrossRef] [PubMed]
47. Seppey, M.; Manni, M.; Zdobnov, E.M. BUSCO: Assessing Genome Assembly and Annotation Completeness. In *Gene Prediction: Methods and Protocols*; Kollmar, M., Ed.; Methods in Molecular Biology; Springer: New York, NY, USA, 2019; pp. 227–245, ISBN 978-1-4939-9173-0.
48. Nishimura, O.; Hara, Y.; Kuraku, S. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics* **2017**, *33*, 3635–3637. [CrossRef]
49. Hubley, R.; Finn, R.D.; Clements, J.; Eddy, S.R.; Jones, T.A.; Bao, W.; Smit, A.F.A.; Wheeler, T.J. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **2016**, *44*, D81–D89. [CrossRef]

50. Zanni, V.; Eymery, A.; Coiffet, M.; Zytnicki, M.; Luyten, I.; Quesneville, H.; Vaury, C.; Jensen, S. Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 19842–19847. [CrossRef]
51. Sedlazeck, F.J.; Rescheneder, P.; Smolka, M.; Fang, H.; Nattestad, M.; von Haeseler, A.; Schatz, M.C. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **2018**, *15*, 461–468. [CrossRef]
52. Linheiro, R.S.; Bergman, C.M. Whole Genome Resequencing Reveals Natural Target Site Preferences of Transposable Elements in Drosophila melanogaster. *PLoS ONE* **2012**, *7*, e30008. [CrossRef]
53. Arnaud, F.; Peyretaillade, E.; Dastugue, B.; Vaury, C. Functional characteristics of a reverse transcriptase encoded by an endogenous retrovirus from Drosophila melanogaster. *Insect Biochem. Mol. Biol.* **2005**, *35*, 323–331. [CrossRef] [PubMed]
54. Lavrov, S.; Déjardin, J.; Cavalli, G. Combined immunostaining and FISH analysis of polytene chromosomes. In *Methods in Molecular Biology*; Humana Press: Clifton, NJ, USA, 2004; Volume 247, pp. 289–303. [CrossRef]
55. Leblanc, P.; Dastugue, B.; Vaury, C. The Integration Machinery of ZAM, a Retroelement from Drosophila melanogaster, Acts as a Sequence-Specific Endonuclease. *J. Virol.* **1999**, *73*, 7061–7064. [CrossRef] [PubMed]
56. George, P.; Jensen, S.; Pogorelcnik, R.; Lee, J.; Xing, Y.; Brasset, E.; Vaury, C.; Sharakhov, I.V. Increased production of piRNAs from euchromatic clusters and genes in Anopheles gambiae compared with Drosophila melanogaster. *Epigenetics Chromatin* **2015**, *8*, 50. [CrossRef] [PubMed]
57. Dos Santos, G.; Schroeder, A.J.; Goodman, J.L.; Strelets, V.B.; Crosby, M.A.; Thurmond, J.; Emmert, D.B.; Gelbart, W.M. FlyBase: Introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res.* **2015**, *43*, D690–D697. [CrossRef]
58. Kozomara, A.; Griffiths-Jones, S. miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **2014**, *42*, D68–D73. [CrossRef]
59. Langmead, B. Aligning Short Sequencing Reads with Bowtie. *Curr. Protoc. Bioinforma.* **2010**, *32*, 11.7.1–11.7.14. [CrossRef]
60. Mérel, V.; Boulesteix, M.; Fablet, M.; Vieira, C. Transposable elements in Drosophila. *Mobile DNA*. in press.
61. Vieira, C.; Fablet, M.; Lerat, E.; Boulesteix, M.; Rebollo, R.; Burlet, N.; Akkouche, A.; Hubert, B.; Mortada, H.; Biémont, C. A comparative analysis of the amounts and dynamics of transposable elements in natural populations of Drosophila melanogaster and Drosophila simulans. *J. Environ. Radioact.* **2012**, *113*, 83–86. [CrossRef]
62. Vieira, C.; Biémont, C. Geographical variation in insertion site number of retrotransposon 412 in Drosophila simulans. *J. Mol. Evol.* **1996**, *42*, 443–451. [CrossRef]
63. Kofler, R.; Nolte, V.; Schlötterer, C. Tempo and Mode of Transposable Element Activity in Drosophila. *PLoS Genet.* **2015**, *11*, e1005406. [CrossRef]
64. Biémont, C.; Nardon, C.; Deceliere, G.; Lepetit, D.; Lœvenbruck, C.; Vieira, C. Worldwide Distribution of Transposable Element Copy Number in Natural Populations of Drosophila Simulans. *Evolution* **2003**, *57*, 159–167. [CrossRef] [PubMed]
65. Kelleher, E.S.; Barbash, D.A. Analysis of piRNA-mediated silencing of active TEs in Drosophila melanogaster suggests limits on the evolution of host genome defense. *Mol. Biol. Evol.* **2013**, *30*, 1816–1829. [CrossRef] [PubMed]
66. Song, J.; Liu, J.; Schnakenberg, S.L.; Ha, H.; Xing, J.; Chen, K.C. Variation in piRNA and Transposable Element Content in Strains of Drosophila melanogaster. *Genome Biol. Evol.* **2014**, *6*, 2786–2798. [CrossRef] [PubMed]
67. Shpiz, S.; Ryazansky, S.; Olovnikov, I.; Abramov, Y.; Kalmykova, A. Euchromatic Transposon Insertions Trigger Production of Novel Pi- and Endo-siRNAs at the Target Sites in the Drosophila Germline. *PLoS Genet* **2014**, *10*, e1004138. [CrossRef]
68. Brennecke, J.; Aravin, A.A.; Stark, A.; Dus, M.; Kellis, M.; Sachidanandam, R.; Hannon, G.J. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. *Cell* **2007**, *128*, 1089–1103. [CrossRef]
69. Goriaux, C.; Desset, S.; Renaud, Y.; Vaury, C.; Brasset, E. Transcriptional properties and splicing of the flamenco piRNA cluster. *EMBO Rep.* **2014**, *15*, 411–418. [CrossRef]
70. Duc, C.; Yoth, M.; Jensen, S.; Mouniée, N.; Bergman, C.M.; Vaury, C.; Brasset, E. Trapping a somatic endogenous retrovirus into a germline piRNA cluster immunizes the germline against further invasion. *Genome Biol.* **2019**, *20*, 127. [CrossRef]

71. Díaz-González, J.; Domínguez, A.; Albornoz, J. Genomic distribution of retrotransposons 297, 1731, copia, mdg1 and roo in the Drosophila melanogaster species subgroup. *Genetica* **2010**, *138*, 579–586. [CrossRef]
72. Craig, N.L. *Mobile DNA II*; ASM Press: Washington, DC, USA, 2002.

A.2 Phenotypic and transcriptomic responses to stress differ according to population geography in an invasive species

Pierre Marin[1], Angelo Jaquet[1], Justine Picarle[1], Marie Fablet[1], **Vincent Mérel**[1], Marie-Laure Delignette-Muller[1], Mariana Galvão Ferrarini[1,2], Patricia Gibert[1] & Cristina Vieira[1].

Co-corresponding authors: Patricia Gibert - patricia.gibert@univ-lyon1.fr & Cristina Vieira - cristina.vieira@univ-lyon1.fr

Affiliations:

1: Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, F-69622 Villeurbanne, France

2: UMR0203, Biologie Fonctionnelle, Insectes et Interactions (BF2i), Institut National des Sciences Appliquées de Lyon (INSA-Lyon), Institut National de la Recherche Agronomique (INRA), University of Lyon (Univ Lyon), F-69621 Villeurbanne, France

A.2.1 Abstract

Background: Adaptation to rapid environmental changes must occur within a short time scale. In this context, studies of invasive species may provide insights into the underlying mechanisms of rapid adaptation as these species have repeatedly encountered and successfully adapted to novel environmental conditions. Here we investigated how invasive and non-invasive populations of *D. suzukii* deal with an oxidative stress at both the phenotypic and molecular level. We also investigated the impact of transposable element insertions on the differential gene expression between genotypes in response to oxidative stress.

Results: Invasive populations lived longer in the untreated condition than non-invasive Japanese populations. As expected, lifespan was greatly reduced following exposure to paraquat, but this reduction varied among genotypes (a genotype by environment interaction, GEI) with invasive genotypes appearing more affected by exposure than non-invasive genotypes. We also performed transcriptomic sequencing of selected genotypes upon and without paraquat and detected a large number of genes differentially expressed, distinguishing the genotypes in the untreated environment. While a small core set of genes were differentially expressed by all genotypes following paraquat exposure, much of the response of each population was unique. Interestingly, we identified a set of genes presenting genotype by environment interaction (GEI). Many of these differences may reflect signatures of history of past adaptation. Transposable elements (TEs) were not activated after oxidative stress and differentially expressed (DE) genes were significantly depleted of TEs.

Conclusion: In the decade since the invasion from the south of Asia, invasive populations of *D. suzukii* have diverged from populations in the native area regarding their genetic response to oxidative stress. This suggests that such transcriptomic changes could be involved in the rapid adaptation to local environments.

A.2.2 Full text

The full text is available at: https://www.biorxiv.org/content/10.1101/2020.09.02.279315v1/

A.3 Comparative transcriptomics between *Drosophila mojavensis* and *D. arizonae* reveal transgressive gene expression and underexpression of spermatogenesis-related genes and hybrid testes

Cecilia A. Banho[1], **Vincent Mérel**[2], Thiago Y. K. Oliveira[3], Claudia M. A. Carareto[1] & Cristina Vieira[2].

Corresponding author: Cristina Vieira - cristina.vieira@univ-lyon1.fr

Affiliations:

1: Department of Biology, UNESP - São Paulo State University, São José do Rio Preto, São Paulo State (SP), Brazil

2: Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, F-69622 Villeurbanne, France

3: Laboratory of Molecular Immunology, The Rockefeller University, New York, NY, USA

A.3.1 Abstract

Interspecific hybridization is a stressful condition that can lead to sterility and/or inviability through improper gene regulation in Drosophila species with a high divergence time. However, the extent of these abnormalities in hybrids of recently diverging species is not well known. Some studies have shown that in Drosophila, the mechanisms of postzygotic isolation may evolve more rapidly in males than in females, and that the degree of viability and sterility is associated with the genetic distance between species. Here, by testing the hypothesis that the severity of the hybrid phenotype is associated with the degree of gene misregulation, we observed that hybrids with different phenotypes (female fertility and male sterility with and without sperm motility) showed different degrees of gene misregulation. Through the use of gene expression analyses in hybrids between four *Drosophila mojavensis* subspecies and *D. arizonae* (repleta group, Drosophila), we have shown that the degree of gene differential expression is greater in the testes than in ovaries. Moreover, the degree of gene misregulation was higher and presented a bias for underexpression in hybrids without motile sperm. In addition, for these hybrids, we identified candidate genes that were mostly associated with spermatogenesis dysfunction.

A.4 Watterson's theta evolution in *D. suzukii* populations

A.4.1 Materials & Methods

In order to assess if $\widehat{\theta_W}$ achieved equilibrium in our invasive populations we performed forward simulations using SLiM (v3.5) (HALLER et MESSER, 2019). Using a chromosome length of 5 kb, a recombination rate of 2.32 * 10^{-8} (COMERON et collab., 2012) and a mutation rate of 2.8 * 10^{-9} (KEIGHTLEY et collab., 2014), we followed $\widehat{\theta_W}$ in a population that undergo a bottleneck. The initial population size was chosen to represent the population size of our native populations. Assuming that $\widehat{\theta_W}$ achieved equilibrium in these populations, we can use $\widehat{\theta_W}$=4$N_e\mu$ to estimate the population size. Given a mean $\widehat{\theta_W}$ of 0.019 in these populations and a mutation rate of 2.8 * 10^{-9} (KEIGHTLEY et collab., 2014), population size should be 1.7 * 10^7 individuals. After a burnin period of 7.5*N_e the population size was divided by a factor ranging from to 10 to 1000. This bottleneck was followed by a period of expansion with a growth rate comprised between 1 and 5. Note that to improve computing time a 0.05 downscaling was performed.

A.4.2 Results

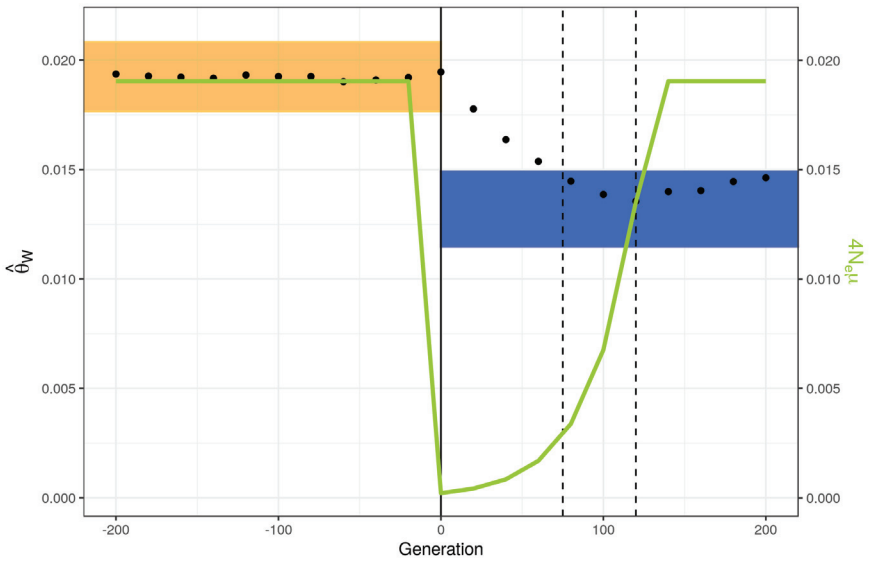


Figure A.1: **Evolution of $\widehat{\theta_W}$ in a simulated population undergoing a bottleneck.** Mutation and recombination rates come from D. melanogaster studies (see Materials Methods). The initial population size mimics the expected population size in native populations. At generation 0, population size is divided by 10, and then population size is multiplied by two every generation. The vertical solid line represents the bottleneck. For comparison, in our PoolSeq dataset sampling was done 75-120 generations after the bottleneck (considering a bottleneck occuring in 2008, a sampling between 2013 and 2015, and 15 generation per year). The dashed lines define the sampling period. The green line represents 4$N_e\mu$, i.e. the expected value of $\widehat{\theta_W}$ at the equilibrium. The orange rectangle represents the range of $\widehat{\theta_W}$ in native populations. The blue rectangle represents the range of $\widehat{\theta_W}$ in invasive populations.

A.5 Estimation du contenu en ET génomique avec des lectures courtes: alignement sur une banque d'ET

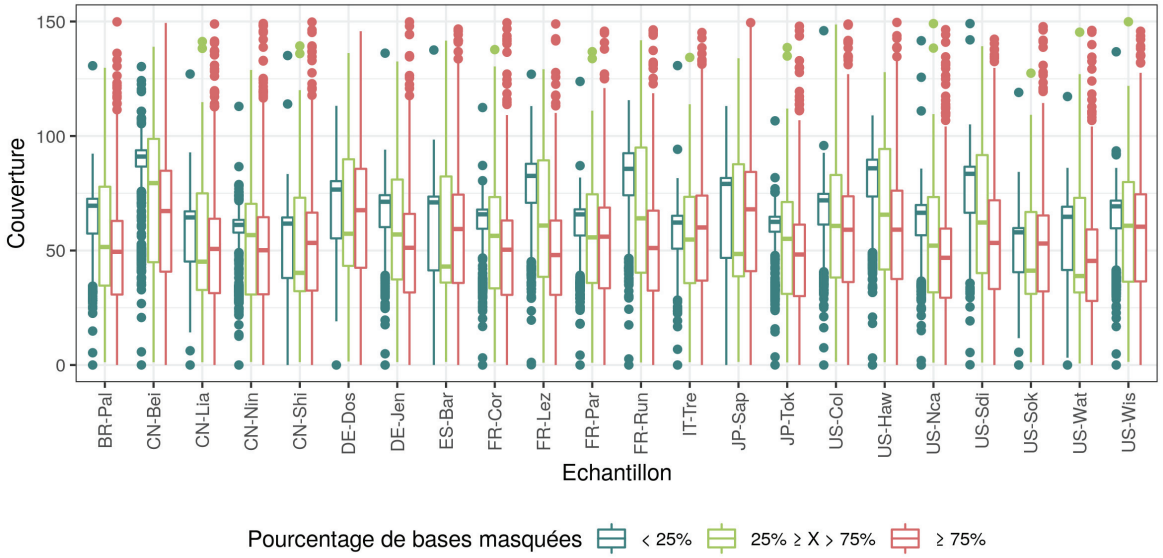


Figure A.2: **Couverture dans les échantillons de PoolSeq de *D. suzukii* en fonction du pourcentage de bases masquées localement.** Couverture et pourcentage de bases masquées sont estimées pour des fenêtres non chevauchantes de 200 Kb.

Il est théoriquement possible d'utiliser deux types de méthodes pour estimer l'abondance des ET avec des lectures courtes : l'alignement sur une banque d'ET et l'utilisation d'outils dédiés à l'étude du polymorphisme d'insertion (voir Section 0.3.4). Les deux méthodes ont été utilisées pour calculer l'abondance des ET dans les populations de *D. suzukii*. Les résultats obtenus diffèrent. Les résultats obtenus par alignement sur une banque d'ET indiquent une absence de structure géographique et des variations très importantes d'abondance en ET entre échantillons. Plus précisement deux « sous méthodes » ont été utilisées pour estimer l'abondance par alignement sur une banque d'ET. Dans le premier cas, un sous jeu de données de lectures (tel que la couverture=6.25X) est aligné avec RepeatMasker sur une banque d'ET [1]. Le pourcentage de paires de bases alignées est alors considéré comme le pourcentage d'ET dans l'échantillon (comme dans LERAT et collab. (2019)). Dans le deuxième cas, les lectures sont alignées sur la banque d'ET et trois régions non répétées, i.e. extraites de l'assemblage masqué, à l'aide de l'outil deviaTE (WEILGUNY et KOFLER, 2019). Le nombre d'insertions par génome haploïde est obtenu en faisant le ratio du nombre de lectures qui s'alignent sur l'ET et du nombre de lectures qui s'alignent sur des régions non répétées. Nos résultats suggèrent que les deux sous méthodes sont biaisées par une profondeur de séquençage variable entre les régions très répétées et les régions peu répétées, en interaction avec le facteur « échantillon » . Le génome de référence a été « découpé » en 1695 fenêtres non chevauchantes (dont 1149 de 200 kb et 451 plus petites). Au total, 42% des fenêtres présentaient moins de 25% de pb masquées par RepeatMasker, 6% entre 25% et 75%, et 52% plus de 75% (en ne considérant que les fenêtre de 200 kb, 56% des fenêtres présentaient moins de 25% de pb masquées par RepeatMasker, 4% entre 25% et 75%, et 39% plus de 75%). Sauf pour un échantillon, les régions fortement répetées

[1] La version de la banque d'ET utilisée ici ne correspond pas exactement à la version du Chapitre 2 mais à une version précedente

présentent une couverture médiane plus faibles que les régions faiblement répétées (fig.A.2). Pour l'échantillon « FR-Run » par exemple, la couverture médiane des régions fortement répétées est de 59X, contre 85X pour les régions faiblement repetées. Le ratio de couverture médiane entre régions fortement et faiblement répétées explique 50% de la variance dans les résultats obtenus avec la première « sous-méthode » ($R^2 = 50\%$, $p = 0.0001981$), et 97% pour la deuxième « sous-méthode » ($R^2 = 97\%$, $p =< 2.2e-16$)[1] (fig.A.3). Plus simplement, plus le séquençage est biasé vers les régions contenant beaucoup d'ET, plus l'abondance estimée pour l'échantillon est importante.

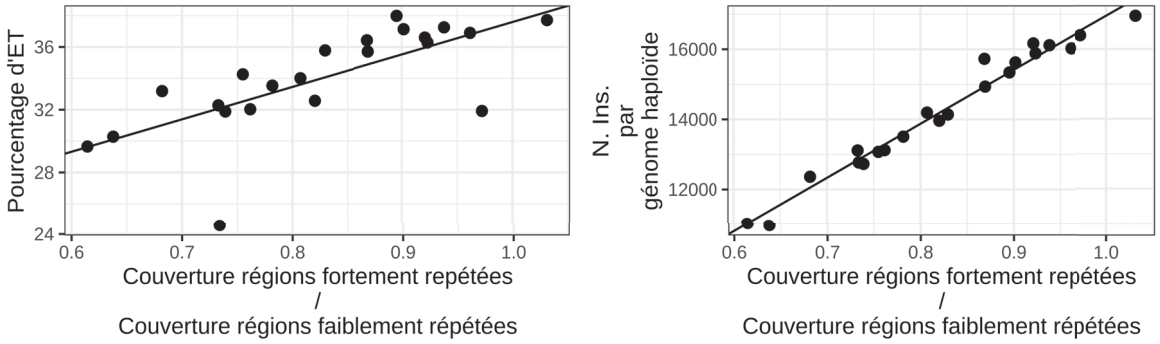


Figure A.3: **Estimation du contenu en ET avec des lectures courtes par alignement sur une banque d'ET dans les échantillons de PoolSeq de *D. suzukii*.** Le contenu en ET est estimé comme le pourcentage de bases alignées sur la banque d'ET dans un échantillon de lectures sur le graphique de gauche. Pour le graphique de droite le nombre d'insertions par génome haploïde correspond à un ratio entre le nombre de lectures alignées sur les ET, et le nombre de lectures qui s'alignent sur des régions non répétées.

[1] « Je ne prendrai pas ce chiffre pour dire que c'est un échec : ça n'a pas marché », Emmanuel Macron

Liste des acronymes

CRISPR-Cas9 courtes répétitions palindromiques regroupées et régulièrement espacées - protéine 9 associée à CRISPR (ou *Clustered Regularly Interspaced Short Palindromic Repeats-CRISPR associated protein 9* en anglais).

EAS Expression Allèle Spécifique (ou *Allele Specific Expression* en anglais).

ET Élément Transposable.

F_ST Indice de fixation de Wright (ou *Fixation Index-Statistics* en anglais).

GO Gene Ontology.

InDel Insertion/Déletion.

LINE Longs élément nucléaire intercalé (ou *Long Interspersed Nuclear Element* en anglais).

LTR Longues Terminaisons Répétées (ou *Long Terminal Repeats* en anglais).

MAF Fréquence de l'allèle minoritaire (ou *Minor Allelic Frequency* en Anglais).

NCBI Centre Américain pour les Informations Biotechnologiques (ou *National Center For Biotechnology Information* en anglais).

NGS Séquençage de nouvelle génération (ou *Next Generation Sequencing* en anglais).

PCR Réaction en chaîne par polymérase (ou *Polymerase Chain Reaction* en anglais).

piRNA ARN interagissants avec PIWI (ou *PIWI-interacting RNA* en anglais).

RC Cercles roulants (ou *Rolling Circle* en anglais).

RT-qPCR PCR quantitative à partir d'un échantillon d'ARN (ou *Reverse Transcription and quantitative PCR* en anglais).

SINE Petit élément nucléaire intercalé (ou *Short Interspersed Nuclear Element* en anglais).

SNP Polymorphisme d'un seul nucléotide (ou *Single Nucleotide Polymorphism* en anglais).

Liste des symboles

C_2 Statistique contrastant les fréquences alléliques entre groupes de populations, avec une correction pour la structure des populations.

D_{Taj} D de Tajima.

dN/dS Ratio divergence non-synonyme/divergence synonyme.

eBp_{is} p-value associée au test d'association entre différence de fréquences alléliques et covariables population spécifiques.

Kr/Kc Ratio substitutions radicales/substitutions conservatives.

N_e Taille efficace.

ω Ratio divergence non-synonyme/divergence synonyme.

s Coefficient de sélection.

μ Taux de mutation par paire de base.

$\widehat{\theta_W}$ Theta de Watterson.

XtX Statistique assimilable à un F_ST SNP spécifique corrigé pour la structure des populations.

Liste complète des références

a, «FlyBase Gene Report: Dmel\papi», URL https://flybase.org/reports/FBgn0031401. 96

b, «FlyBase Gene Report: Dmel\tapas», URL http://flybase.org/reports/FBgn0027529. 96

c, «Genome List - Genome - NCBI», URL https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/. 14

ADAMS, M. D., S. E. CELNIKER, R. A. HOLT, C. A. EVANS, J. D. GOCAYNE, P. G. AMANATIDES, S. E. SCHERER, P. W. LI, R. A. HOSKINS, R. F. GALLE, R. A. GEORGE, S. E. LEWIS, S. RICHARDS, M. ASHBURNER, S. N. HENDERSON, G. G. SUTTON, J. R. WORTMAN, M. D. YANDELL, Q. ZHANG, L. X. CHEN, R. C. BRANDON, Y.-H. C. ROGERS, R. G. BLAZEJ, M. CHAMPE, B. D. PFEIFFER, K. H. WAN, C. DOYLE, E. G. BAXTER, G. HELT, C. R. NELSON, G. L. GABOR, MIKLOS, J. F. ABRIL, A. AGBAYANI, H.-J. AN, C. ANDREWS-PFANNKOCH, D. BALDWIN, R. M. BALLEW, A. BASU, J. BAXENDALE, L. BAYRAKTAROGLU, E. M. BEASLEY, K. Y. BEESON, P. V. BENOS, B. P. BERMAN, D. BHANDARI, S. BOLSHAKOV, D. BORKOVA, M. R. BOTCHAN, J. BOUCK, P. BROKSTEIN, P. BROTTIER, K. C. BURTIS, D. A. BUSAM, H. BUTLER, E. CADIEU, A. CENTER, I. CHANDRA, J. M. CHERRY, S. CAWLEY, C. DAHLKE, L. B. DAVENPORT, P. DAVIES, B. D. PABLOS, A. DELCHER, Z. DENG, A. D. MAYS, I. DEW, S. M. DIETZ, K. DODSON, L. E. DOUP, M. DOWNES, S. DUGAN-ROCHA, B. C. DUNKOV, P. DUNN, K. J. DURBIN, C. C. EVANGELISTA, C. FERRAZ, S. FERRIERA, W. FLEISCHMANN, C. FOSLER, A. E. GABRIELIAN, N. S. GARG, W. M. GELBART, K. GLASSER, A. GLODEK, F. GONG, J. H. GORRELL, Z. GU, P. GUAN, M. HARRIS, N. L. HARRIS, D. HARVEY, T. J. HEIMAN, J. R. HERNANDEZ, J. HOUCK, D. HOSTIN, K. A. HOUSTON, T. J. HOWLAND, M.-H. WEI, C. IBEGWAM, M. JALALI, F. KALUSH, G. H. KARPEN, Z. KE, J. A. KENNISON, K. A. KETCHUM, B. E. KIMMEL, C. D. KODIRA, C. KRAFT, S. KRAVITZ, D. KULP, Z. LAI, P. LASKO, Y. LEI, A. A. LEVITSKY, J. LI, Z. LI, Y. LIANG, X. LIN, X. LIU, B. MATTEI, T. C. MCINTOSH, M. P. MCLEOD, D. MCPHERSON, G. MERKULOV, N. V. MILSHINA, C. MOBARRY, J. MORRIS, A. MOSHREFI, S. M. MOUNT, M. MOY, B. MURPHY, L. MURPHY, D. M. MUZNY, D. L. NELSON, D. R. NELSON, K. A. NELSON, K. NIXON, D. R. NUSSKERN, J. M. PACLEB, M. PALAZZOLO, G. S. PITTMAN, S. PAN, J. POLLARD, V. PURI, M. G. REESE, K. REINERT, K. REMINGTON, R. D. C. SAUNDERS, F. SCHEELER, H. SHEN, B. C. SHUE, I. SIDÉN-KIAMOS, M. SIMPSON, M. P. SKUPSKI, T. SMITH, E. SPIER, A. C. SPRADLING, M. STAPLETON, R. STRONG, E. SUN, R. SVIRSKAS, C. TECTOR, R. TURNER, E. VENTER, A. H. WANG, X. WANG, Z.-Y. WANG, D. A. WASSARMAN, G. M. WEINSTOCK, J. WEISSENBACH, S. M. WILLIAMS, T. WOODAGE, K. C. WORLEY, D. WU, S. YANG, Q. A. YAO, J. YE, R.-F. YEH, J. S. ZAVERI, M. ZHAN, G. ZHANG, Q. ZHAO, L. ZHENG, X. H. ZHENG, F. N. ZHONG, W. ZHONG, X. ZHOU, S. ZHU, X. ZHU, H. O. SMITH, R. A. GIBBS, E. W. MYERS, G. M. RUBIN et J. C. VENTER. 2000, «The Genome Sequence of

Drosophila melanogaster», *Science*, vol. 287, 5461, doi:10.1126/science.287.5461.2185, p. 2185–2195, ISSN 0036-8075, 1095-9203. URL https://science-sciencemag-org.inee.bib.cnrs.fr/content/287/5461/2185. 13, 65, 97

ADRION, J. R., M. J. SONG, D. R. SCHRIDER, M. W. HAHN et S. SCHAACK. 2017, «Genome-Wide Estimates of Transposable Element Insertion and Deletion Rates in Drosophila Melanogaster», *Genome Biology and Evolution*, vol. 9, 5, doi:10.1093/gbe/evx050, p. 1329–1340, ISSN 1759-6653. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5447328/. 67

ALEXA, A. et J. RAHNENFUHRER. «Gene set enrichment analysis with topGO», , p. 26. 102

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS et D. J. LIPMAN. 1990, «Basic local alignment search tool», *Journal of Molecular Biology*, vol. 215, 3, doi:10.1016/S0022-2836(05)80360-2, p. 403–410, ISSN 0022-2836. URL http://www.sciencedirect.com/science/article/pii/S0022283605803602. 17, 69

ARKHIPOVA, I. R. 2017, «Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories», *Mobile DNA*, vol. 8, doi:10.1186/s13100-017-0103-2, ISSN 1759-8753. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5718144/. 70

ATALLAH, J., L. TEIXEIRA, R. SALAZAR, G. ZARAGOZA et A. KOPP. 2014, «The making of a pest: the evolution of a fruit-penetrating ovipositor in Drosophila suzukii and related species», *Proceedings of the Royal Society B: Biological Sciences*, vol. 281, 1781, doi:10.1098/rspb.2013.2840, ISSN 0962-8452. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3953835/. 21, 25

BADUEL, P., L. QUADRANA, B. HUNTER, K. BOMBLIES et V. COLOT. 2019, «Relaxed purifying selection in autopolyploids drives transposable element over-accumulation which provides variants for local adaptation», *Nature Communications*, vol. 10, 1, doi:10.1038/s41467-019-13730-0, ISSN 2041-1723. URL http://www.nature.com/articles/s41467-019-13730-0. 11, 88

BAILLY-BECHET, M., A. HAUDRY et E. LERAT. 2014, «"One code to find them all": a perl tool to conveniently parse RepeatMasker output files», *Mobile DNA*, vol. 5, 1, doi:10.1186/1759-8753-5-13, p. 13, ISSN 1759-8753. URL https://doi.org/10.1186/1759-8753-5-13. 17, 70

BAO, Z. et S. R. EDDY. 2002, «Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes», *Genome Research*, vol. 12, 8, doi:10.1101/gr.88502, p. 1269–1276, ISSN 1088-9051. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC186642/. 16, 69

BARTOLOMÉ, C., X. MASIDE et B. CHARLESWORTH. 2002, «On the Abundance and Distribution of Transposable Elements in the Genome of Drosophila melanogaster», *Molecular Biology and Evolution*, vol. 19, 6, doi:10.1093/oxfordjournals.molbev.a004150, p. 926–937, ISSN 0737-4038. URL https://academic.oup.com/mbe/article/19/6/926/1094868. 12, 13, 55, 65, 118

BENJAMINI, Y. et Y. HOCHBERG. 1995, «Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing», *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, 1, p. 289–300, ISSN 0035-9246. URL https://www.jstor.org/stable/2346101. 73

BERGMAN, C. M. et D. BENSASSON. 2007, «Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in Drosophila melanogaster», *Proceedings of the National Academy of Sciences*, vol. 104, 27, doi:10.1073/pnas.0702552104, p. 11 340–11 345, ISSN 0027-8424, 1091-6490. URL https://www.pnas.org/content/104/27/11340. 87

BERGTHORSSON, U., C. J. SHEEBA, A. KONRAD, T. BELICARD, T. BELTRAN, V. KATJU et P. SARKIES. 2020, «Long-term experimental evolution reveals purifying selection on piRNA-mediated control of transposable element expression», *BMC Biology*, vol. 18, 1, doi:10.1186/s12915-020-00897-y, p. 162, ISSN 1741-7007. URL https://doi.org/10.1186/s12915-020-00897-y. 120

BIÉMONT, C., A. AOUAR et C. ARNAULT. 1987, «Genome reshuffling of the copia element in an inbred line of Drosophila melanogaster», *Nature*, vol. 329, 6141, doi:10.1038/329742a0, p. 742–744, ISSN 1476-4687. URL https://www-nature-com.inee.bib.cnrs.fr/articles/329742a0. 8, 56, 67

BIÉMONT, C., F. LEMEUNIER, P. GARCIA GUERREIRO, J. BROOKFIELD, C. GAUTIER, S. AULARD et E. PASYUKOVA. 1994, «Population dynamics of the copia, mdg1, mdg3, gypsy, and P transposable elements in a natural population of Drosophila melanogaster», *Genetical research*, vol. 63, doi:10.1017/S0016672300032353, p. 197–212. 9

BLUMENSTIEL, J. P., X. CHEN, M. HE et C. M. BERGMAN. 2014, «An Age of Allele Test of Neutrality for Transposable Element Insertions», *Genetics*, vol. 196, 2, doi:10.1534/genetics.113.158147, p. 523–538, ISSN 0016-6731, 1943-2631. URL https://www.genetics.org/content/196/2/523. 12, 56

BOGAERTS-MÁRQUEZ, M., M. G. BARRÓN, A.-S. FISTON-LAVIER, P. VENDRELL-MIR, R. CASTANERA, J. M. CASACUBERTA et J. GONZÁLEZ. 2020, «T-lex3: an accurate tool to genotype and estimate population frequencies of transposable elements using the latest short-read whole genome sequencing data», *Bioinformatics*, vol. 36, 4, doi:10.1093/bioinformatics/btz727, p. 1191–1197, ISSN 1367-4803. URL https://academic.oup.com/bioinformatics/article/36/4/1191/5580493. 18

BOISSINOT, S., J. DAVIS, A. ENTEZAM, D. PETROV et A. V. FURANO. 2006, «Fitness cost of LINE-1 (L1) activity in humans», *Proceedings of the National Academy of Sciences*, vol. 103, 25, doi:10.1073/pnas.0603334103, p. 9590–9594, ISSN 0027-8424, 1091-6490. URL https://www.pnas.org/content/103/25/9590. 5, 7, 13

BOISSINOT, S., A. ENTEZAM et A. V. FURANO. 2001, «Selection against deleterious LINE-1-containing loci in the human lineage», *Molecular Biology and Evolution*, vol. 18, 6, doi:10.1093/oxfordjournals.molbev.a003893, p. 926–935, ISSN 0737-4038. 55

BOLDA, M., R. GOODHUE et F. ZALOM. 2010, «Spotted wing drosophila: potential economic impact of a newly established pest», *Giannini Foundation Agric. Econ.*, vol. 13, p. 5–8. 24

BOURGEOIS, Y., R. P. RUGGIERO, I. HARIYANI et S. BOISSINOT. 2020, «Disentangling the determinants of transposable elements dynamics in vertebrate genomes using empirical evidences and simulations», *PLOS Genetics*, vol. 16, 10, doi:10.1371/journal.pgen.1009082, p. e1009 082, ISSN 1553-7404. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009082. 5, 7, 9, 10, 12, 115

BROUHA, B., J. SCHUSTAK, R. M. BADGE, S. LUTZ-

PRIGGE, A. H. FARLEY, J. V. MORAN et H. H. KAZAZIAN. 2003, «Hot L1s account for the bulk of retrotransposition in the human population», *Proceedings of the National Academy of Sciences*, vol. 100, 9, doi:10.1073/pnas.0831042100, p. 5280–5285, ISSN 0027-8424, 1091-6490. URL https://www.pnas.org/content/100/9/5280. 9

BROWN, A. J. L. et J. E. MOSS. 1987, «Transposition of the I element and copia in a natural population of Drosophila melanogaster», *Genetics Research*, vol. 49, 2, doi:10.1017/S0016672300026914, p. 121–128, ISSN 1469-5073, 0016-6723. URL https://www.cambridge.org/core/journals/genetics-research/article/transposition-of-the-i-element-and-copia-in-a-natural-population-of-drosophila-melanogaster/956ECA10001D0C8C6FBE9742DCBA6289. 9

BURT, A. et R. TRIVERS. 2006, *Genes in conflict: the biology of selfish genetic elements*, Belknap Press of Harvard University Press, Cambridge, Mass, ISBN 978-0-674-01713-9. OCLC: ocm60882032. 10, 87

C. ELEGANS SEQUENCING CONSORTIUM. 1998, «Genome sequence of the nematode C. elegans: a platform for investigating biology», *Science (New York, N.Y.)*, vol. 282, 5396, doi:10.1126/science.282.5396.2012, p. 2012–2018, ISSN 0036-8075. 7, 13, 55, 87

CALOS, M. P. et J. H. MILLER. 1980, «Transposable elements», *Cell*, vol. 20, 3, doi:10.1016/0092-8674(80)90305-0, p. 579–595, ISSN 0092-8674. URL http://www.sciencedirect.com/science/article/pii/0092867480903050. 3

CASTILLO, D. M., J. C. MELL, K. S. BOX et J. P. BLUMENSTIEL. 2011, «Molecular evolution under increasing transposable element burden in Drosophila: A speed limit on the evolutionary arms race», *BMC Evolutionary Biology*, vol. 11, 1, doi:10.1186/1471-2148-11-258, p. 258, ISSN 1471-2148. URL https://doi.org/10.1186/1471-2148-11-258. 8, 88, 98

CATANIA, F., M. O. KAUER, P. J. DABORN, J. L. YEN, R. H. FFRENCH-CONSTANT et C. SCHLOTTERER. 2004, «World-wide survey of an Accord insertion and its association with DDT resistance in Drosophila melanogaster», *Molecular Ecology*, vol. 13, 8, doi:10.1111/j.1365-294X.2004.02263.x, p. 2491–2504, ISSN 0962-1083. 9, 11

CHAKRABORTY, M., C.-H. CHANG, D. E. KHOST, J. VEDANAYAGAM, J. R. ADRION, Y. LIAO, K. MONTOOTH, C. D. MEIKLEJOHN, A. M. LARRACUENTE et J. J. EMERSON. 2020, «Evolution of genome structure in the Drosophila simulans species complex», *bioRxiv*, doi:10.1101/2020.02.27.968743, p. 2020.02.27.968743. URL https://www.biorxiv.org/content/10.1101/2020.02.27.968743v1, publisher: Cold Spring Harbor Laboratory Section: New Results. 96

CHARLESWORTH, B. et D. CHARLESWORTH. 1983, «The population dynamics of transposable elements», *Genetical Research*, vol. 42, 1, doi:10.1017/S0016672300021455, p. 1–27, ISSN 0016-6723, 1469-5073. URL https://www.cambridge.org/core/product/identifier/S0016672300021455/type/journal_article. 4, 7, 55, 87

CHARLESWORTH, B., A. LAPID et D. CANADA. 1992, «The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster* . I. Element frequencies and distribution», *Genetical Research*, vol. 60, 2, doi:10.1017/S0016672300030792, p. 103–114, ISSN 0016-6723, 1469-5073. URL https://www.cambridge.org/core/pro

duct/identifier/S0016672300030792/type/journal_article. 9

CHIU, J. C., X. JIANG, L. ZHAO, C. A. HAMM, J. M. CRIDLAND, P. SAELAO, K. A. HAMBY, E. K. LEE, R. S. KWOK, G. ZHANG, F. G. ZALOM, V. M. WALTON et D. J. BEGUN. 2013, «Genome of Drosophila suzukii, the Spotted Wing Drosophila», *G3: Genes|Genomes|Genetics*, vol. 3, 12, doi:10.1534/g3.113.008185, p. 2257–2271, ISSN 2160-1836. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3852387/. 64, 121

CINI, A., C. IORIATTI et G. ANFORA. 2012, «A review of the invasion of Drosophila suzukii in Europe and a draft research agenda for integrated pest management», *Bulletin of Insectology*, vol. 65, p. 149–160. 22

COMERON, J. M., R. RATNAPPAN et S. BAILIN. 2012, «The Many Landscapes of Recombination in Drosophila melanogaster», *PLOS Genetics*, vol. 8, 10, doi:10.1371/journal.pgen.1002905, p. e1002 905, ISSN 1553-7404. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002905. 158

CRIDLAND, J. M., S. J. MACDONALD, A. D. LONG et K. R. THORNTON. 2013, «Abundance and Distribution of Transposable Elements in Two Drosophila QTL Mapping Resources», *Molecular Biology and Evolution*, vol. 30, 10, doi:10.1093/molbev/mst129, p. 2311–2327, ISSN 0737-4038. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3773372/. 13, 55

DABORN, P. J., J. L. YEN, M. R. BOGWITZ, G. L. GOFF, E. FEIL, S. JEFFERS, N. TIJET, T. PERRY, D. HECKEL, P. BATTERHAM, R. FEYEREISEN, T. G. WILSON et R. H. FFRENCH CONSTANT. 2002, «A Single P450 Allele Associated with Insecticide Resistance in Drosophila», *Science*, vol. 297, 5590, doi:10.1126/science.1074170, p. 2253–2256, ISSN 0036-8075, 1095-9203. URL https://science.sciencemag.org/content/297/5590/2253. 9, 11, 55, 56, 64

DEININGER, P. 2011, «Alu elements: know the SINEs», *Genome Biology*, vol. 12, 12, doi:10.1186/gb-2011-12-12-236, p. 236, ISSN 1474-760X. URL https://doi.org/10.1186/gb-2011-12-12-236. 3

DENG, Q., Q. ZENG, Y. QIAN, C. LI et Y. YANG. 2007, «Research on the karyotype and evolution of Drosophila melanogaster species group», *Journal of Genetics and Genomics = Yi Chuan Xue Bao*, vol. 34, 3, doi:10.1016/S1673-8527(07)60021-6, p. 196–213, ISSN 1673-8527. 22

DIGIACOMO, G., J. HADRICH, W. D. HUTCHISON, H. PETERSON et M. ROGERS. 2019, «Economic Impact of Spotted Wing Drosophila (Diptera: Drosophilidae) Yield Loss on Minnesota Raspberry Farms: A Grower Survey», *Journal of Integrated Pest Management*, vol. 10, 1, doi:10.1093/jipm/pmz006. URL https://academic.oup.com/jipm/article/10/1/11/5476556. 24

DINIZ-FILHO, J. A. F., T. N. SOARES, J. S. LIMA, R. DOBROVOLSKI, V. L. LANDEIRO, M. P. DE CAMPOS TELLES, T. F. RANGEL et L. M. BINI. 2013, «Mantel test in population genetics», *Genetics and Molecular Biology*, vol. 36, 4, doi:10.1590/S1415-47572013000400002, p. 475–485, ISSN 1415-4757. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3873175/. 67

DOOLITTLE, W. F. et C. SAPIENZA. 1980, «Selfish genes, the phenotype paradigm and genome evolution», *Nature*, vol. 284, 5757, doi:10.1038/284601a0, p. 601–603, ISSN 1476-4687. URL https://www.nature.com/articles/284601a0. 4, 55, 87

DURET, L., G. MARAIS et C. BIÉMONT. 2000, «Transposons but Not Retrotransposons Are Located Preferentially in Regions of High Recombination Rate in Caenorhabditis elegans», *Genetics*, vol. 156, 4, p. 1661–1669, ISSN 0016-6731, 1943-2631. URL https://www.genetics.org/content/156/4/1661. 13

DWIVEDI, Y. N. et J. R. GUPTA. 1981, «Phylogenetic relationship between Drosophila takahashii Sturtevant and Drosophila pseudotakahashii Mather», *Genetica*, vol. 57, 2, doi:10.1007/BF00131232, p. 87–92, ISSN 0016-6707, 1573-6857. URL http://link.springer.com/10.1007/BF00131232. 90

DÍAZ-GONZÁLEZ, J., J. F. VÁZQUEZ, J. ALBORNOZ et A. DOMÍNGUEZ. 2011, «Long-term evolution of the roo transposable element copy number in mutation accumulation lines of Drosophila melanogaster», *Genetics Research*, vol. 93, 3, doi:10.1017/S0016672311000103, p. 181–187, ISSN 1469-5073, 0016-6723. URL https://www.cambridge.org/core/journals/genetics-research/article/longterm-evolution-of-the-roo-transposable-element-copy-number-in-mutation-accumulation-lines-of-drosophila-melanogaster/A01B400F24D556DA6C0C949A8BA2C832. 56, 67

EDGAR, R. et E. MYERS. 2005, «PILER: Identification and classification of genomic repeats», *Bioinformatics (Oxford, England)*, vol. 21 Suppl 1, doi:10.1093/bioinformatics/bti1003, p. i152–8. 16, 69

EDGAR, R. C. 2010, «Search and clustering orders of magnitude faster than BLAST», *Bioinformatics*, vol. 26, 19, doi:10.1093/bioinformatics/btq461, p. 2460–2461, ISSN 1367-4803. URL https://academic.oup.com/bioinformatics/article/26/19/2460/230188. 70, 100

ELLINGHAUS, D., S. KURTZ et U. WILLHOEFT. 2008, «LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons», *BMC Bioinformatics*, vol. 9, 1, doi:10.1186/1471-2105-9-18, p. 18, ISSN 1471-2105. URL https://doi.org/10.1186/1471-2105-9-18. 16, 69, 100

ELLIOTT, T. A. et T. R. GREGORY. 2015, «What's in a genome? The C-value enigma and the evolution of eukaryotic genome content», *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, 1678, doi:10.1098/rstb.2014.0331, p. 20140331. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2014.0331, publisher: Royal Society. 87

EMILJANOWICZ, L. M., G. D. RYAN, A. LANGILLE et J. NEWMAN. 2014, «Development, reproductive output and population growth of the fruit fly pest Drosophila suzukii (Diptera: Drosophilidae) on artificial diet», *Journal of Economic Entomology*, vol. 107, 4, doi:10.1603/ec13504, p. 1392–1398, ISSN 0022-0493. 21

ESTOUP, A., V. RAVIGNÉ, R. HUFBAUER, R. VITALIS, M. GAUTIER et B. FACON. 2016, «Is There a Genetic Paradox of Biological Invasion?», *Annual Review of Ecology, Evolution, and Systematics*, vol. 47, 1, doi:10.1146/annurev-ecolsys-121415-032116, p. 51–72. URL https://doi.org/10.1146/annurev-ecolsys-121415-032116. 56

FICK, S. E. et R. J. HIJMANS. 2017, «WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas», *International Journal of Climatology*, vol. 37, 12, doi:10.1002/joc.5086, p. 4302–4315, ISSN 1097-0088. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.5086. 73

FINNEGAN, D. J. 1992, «Transposable elements», *Current Opinion in Genetics & Development*,

vol. 2, 6, doi:10.1016/s0959-437x(05)80108-x, p. 861–867, ISSN 0959-437X. 5, 118

FLUTRE, T., E. DUPRAT, C. FEUILLET et H. QUESNEVILLE. 2011, «Considering transposable element diversification in de novo annotation approaches», *PloS One*, vol. 6, 1, doi:10.1371/journal.pone.0016526, p. e16 526, ISSN 1932-6203. 16, 17, 26, 69, 100

FLYNN, J. M., R. HUBLEY, C. GOUBERT, J. ROSEN, A. G. CLARK, C. FESCHOTTE et A. F. SMIT. 2020, «RepeatModeler2 for automated genomic discovery of transposable element families», *Proceedings of the National Academy of Sciences*, vol. 117, 17, doi:10.1073/pnas.1921046117, p. 9451–9457, ISSN 0027-8424, 1091-6490. URL https://www.pnas.org/content/117/17/9451. 16

FONTANILLAS, P., D. L. HARTL et M. REUTER. 2007, «Genome Organization and Gene Expression Shape the Transposable Element Distribution in the Drosophila melanogaster Euchromatin», *PLOS Genetics*, vol. 3, 11, doi:10.1371/journal.pgen.0030210, p. e210, ISSN 1553-7404. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0030210. 88, 119

FRAIMOUT, A., V. DEBAT, S. FELLOUS, R. A. HUFBAUER, J. FOUCAUD, P. PUDLO, J.-M. MARIN, D. K. PRICE, J. CATTEL, X. CHEN, M. DEPRÁ, P. FRANÇOIS DUYCK, C. GUEDOT, M. KENIS, M. T. KIMURA, G. LOEB, A. LOISEAU, I. MARTINEZ-SAÑUDO, M. PASCUAL, M. POLIHRONAKIS RICHMOND, P. SHEARER, N. SINGH, K. TAMURA, A. XUÉREB, J. ZHANG et A. ESTOUP. 2017, «Deciphering the Routes of invasion of Drosophila suzukii by Means of ABC Random Forest», *Molecular Biology and Evolution*, vol. 34, 4, doi:10.1093/molbev/msx050, p. 980–996, ISSN 0737-4038. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5400373/. 23, 24, 26, 57, 60, 65, 66, 68, 71, 74, 116

FRAIMOUT, A., P. JACQUEMART, B. VILLARROEL, D. J. APONTE, T. DECAMPS, A. HERREL, R. CORNETTE et V. DEBAT. 2018, «Phenotypic plasticity of Drosophila suzukii wing to developmental temperature: implications for flight», *Journal of Experimental Biology*, vol. 221, 13, doi:10.1242/jeb.166868, ISSN 0022-0949, 1477-9145. URL https://jeb.biologists.org/content/221/13/jeb166868. 25

GARCÍA GUERREIRO, M. P., B. E. CHÁVEZ-SANDOVAL, J. BALANYÀ, L. SERRA et A. FONTDEVILA. 2008, «Distribution of the transposable elements bilbo and gypsy in original and colonizing populations of Drosophila subobscura», *BMC Evolutionary Biology*, vol. 8, 1, doi:10.1186/1471-2148-8-234, p. 234, ISSN 1471-2148. URL http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-8-234. 7, 55, 66

GARCÍA GUERREIRO, M. P. et A. FONTDEVILA. 2011, «Osvaldo and Isis retrotransposons as markers of the Drosophila buzzatii colonisation in Australia», *BMC Evolutionary Biology*, vol. 11, 1, doi:10.1186/1471-2148-11-111, ISSN 1471-2148. URL http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-11-111. 7, 55, 66

GARCÍA GUERREIRO, M. P. G. 2012, «What makes transposable elements move in the Drosophila genome?», *Heredity*, vol. 108, 5, doi:10.1038/hdy.2011.89, p. 461–468, ISSN 0018-067X, 1365-2540. URL http://www.nature.com/articles/hdy201189. 56, 66

GAUTIER, M. 2015, «Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates», *Genetics*, vol. 201, 4, doi:10.1534/genetics.115.181453, p. 1555–1579, ISSN 0016-6731, 1943-2631. URL https://www.genetics.org/content/201/4/1555. 10, 68, 72, 74

GAUTIER, M., J. YAMAGUCHI, J. FOUCAUD, A. LOISEAU, A. AUSSET, B. FACON, B. GSCHLOESSL, J. LAGNEL, E. LOIRE, H. PARRINELLO, D. SEVERAC, C. LOPEZ-ROQUES, C. DONNADIEU, M. MANNO, H. BERGES, K. GHARBI, L. LAWSON-HANDLEY, L.-S. ZANG, H. VOGEL, A. ESTOUP et B. PRUD'HOMME. 2018, «The Genomic Basis of Color Pattern Polymorphism in the Harlequin Ladybird», *Current Biology*, vol. 28, 20, doi:10.1016/j.cub.2018.08.023, p. 3296–3302.e7, ISSN 0960-9822. URL http://www.sciencedirect.com/science/article/pii/S0960982218310686. 72, 74

GIPPET, J. M., A. M. LIEBHOLD, G. FENN-MOLTU et C. BERTELSMEIER. 2019, «Human-mediated dispersal in insects», *Current Opinion in Insect Science*, vol. 35, doi:10.1016/j.cois.2019.07.005, p. 96–102, ISSN 22145745. URL https://linkinghub.elsevier.com/retrieve/pii/S2214574518301883. 25

GONZALEZ, J., J. M. MACPHERSON et D. A. PETROV. 2009, «A Recent Adaptive Transposable Element Insertion Near Highly Conserved Developmental Loci in Drosophila melanogaster», *Molecular Biology and Evolution*, vol. 26, 9, doi:10.1093/molbev/msp107, p. 1949–1961, ISSN 0737-4038, 1537-1719. URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msp107. 11, 68

GONZÁLEZ, J., T. L. KARASOV, P. W. MESSER et D. A. PETROV. 2010, «Genome-Wide Patterns of Adaptation to Temperate Environments Associated with Transposable Elements in Drosophila», *PLOS Genetics*, vol. 6, 4, doi:10.1371/journal.pgen.1000905, p. e1000905, ISSN 1553-7404. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000905. 12, 56, 67

GONZÁLEZ, J., K. LENKOV, M. LIPATOV, J. M. MACPHERSON et D. A. PETROV. 2008, «High Rate of Recent Transposable Element–Induced Adaptation in Drosophila melanogaster», *PLOS Biology*, vol. 6, 10, doi:10.1371/journal.pbio.0060251, p. e251, ISSN 1545-7885. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0060251. 12, 56, 67, 115, 116

GOSLEE, S. C. et D. L. URBAN. 2007, «The ecodist Package for Dissimilarity-based Analysis of Ecological Data», *Journal of Statistical Software*, vol. 22, 1, doi:10.18637/jss.v022.i07, p. 1–19, ISSN 1548-7660. URL https://www.jstatsoft.org/index.php/jss/article/view/v022i07. 73

GOUBERT, C., L. MODOLO, C. VIEIRA, C. VALIENTEMORO, P. MAVINGUI et M. BOULESTEIX. 2015, «De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes aegypti)», *Genome Biology and Evolution*, vol. 7, 4, doi:10.1093/gbe/evv050, p. 1192–1205. URL https://academic.oup.com/gbe/article/7/4/1192/533768. 16, 19, 91, 101, 123

GRASSI, A., L. GIONGO et L. PALMIERI. 2011, «Drosophila (Sophophora) suzukii (Matsumura), new pest of soft fruits in Trentino (North-Italy) and in Europe», *IOBC/WPRS BULLETIN*, vol. 70, p. 121–128. 22

GREMME, G., S. STEINBISS et S. KURTZ. 2013, «GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations», *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, 3, doi:10.1109/TCBB.2013.68, p. 645–656, ISSN 1545-5963. URL https://doi.org/10.1109/TCBB.2013.68. 69

GUIO, L., M. G. BARRÓN et J. GONZÁLEZ. 2014, «The transposable element Bari-Jheh mediates oxidative stress response in Drosophila», *Molecular Ecology*, vol. 23, 8, doi:https://doi.org/10.1111/mec.12711, p. 2020–2030, ISSN 1365-294X. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.12711. 10

GUÉGUEN, L. et L. DURET. 2018, «Unbiased Estimate of Synonymous and Nonsynonymous Substitution Rates with Nonstationary Base Composition», *Molecular Biology and Evolution*, vol. 35, 3, doi:10.1093/molbev/msx308, p. 734–742, ISSN 1537-1719. 97, 113

HALLER, B. C. et P. W. MESSER. 2019, «SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model», *Molecular Biology and Evolution*, vol. 36, 3, doi:10.1093/molbev/msy228, p. 632–637, ISSN 0737-4038. URL https://doi.org/10.1093/molbev/msy228. 158

HAMBY, K. A., D. E. BELLAMY, J. C. CHIU, J. C. LEE, V. M. WALTON, N. G. WIMAN, R. M. YORK et A. BIONDI. 2016, «Biotic and abiotic factors impacting development, behavior, phenology, and reproductive biology of Drosophila suzukii», *Journal of Pest Science*, vol. 89, 3, doi:10.1007/s10340-016-0756-5, p. 605–619, ISSN 1612-4766. URL https://doi.org/10.1007/s10340-016-0756-5. 21, 25

HAMM, C. A., D. J. BEGUN, A. VO, C. C. R. SMITH, P. SAELAO, A. O. SHAVER, J. JAENIKE et M. TURELLI. 2014, «Wolbachia do not live by reproductive manipulation alone: infection polymorphism in Drosophila suzukii and D. subpulchrella», *Molecular ecology*, vol. 23, 19, doi:10.1111/mec.12901, p. 4871–4885, ISSN 0962-1083. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4180775/. 20

HILL, T. 2019, «Transposable element dynamics are consistent across the Drosophila phylogeny, despite drastically differing content», *bioRxiv*, doi:10.1101/651059, p. 651 059. URL https://www.biorxiv.org/content/10.1101/651059v2. 55, 65, 91

HJELMEN, C. E., H. BLACKMON, V. R. HOLMES, C. G. BURRUS et J. S. JOHNSTON. 2019, «Genome Size Evolution Differs Between Drosophila Subgenera with Striking Differences in Male and Female Genome Size in Sophophora», *G3: Genes|Genomes|Genetics*, vol. 9, 10, doi:10.1534/g3.119.400560, p. 3167–3179, ISSN 2160-1836. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6778784/. 22, 88, 113

HOEDE, C., S. ARNOUX, M. MOISSET, T. CHAUMIER, O. INIZAN, V. JAMILLOUX et H. QUESNEVILLE. 2014, «PASTEC: an automatic transposable element classification tool», *PloS One*, vol. 9, 5, doi:10.1371/journal.pone.0091929, p. e91 929, ISSN 1932-6203. 17, 69

HORVÁTH, V., M. MERENCIANO et J. GONZÁLEZ. 2017, «Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response», *Trends in Genetics*, vol. 33, 11, doi:10.1016/j.tig.2017.08.007, p. 832–841, ISSN 0168-9525. URL https://www.cell.com/trends/genetics/abstract/S0168-9525(17)30145-2. 56, 66

HUANG, X. 1994, «On global sequence alignment», *Computer applications in the biosciences: CABIOS*, vol. 10, 3, doi:10.1093/bioinformatics/10.3.227, p. 227–235, ISSN 0266-7061. 69

HUBLEY, R., R. D. FINN, J. CLEMENTS, S. R. EDDY, T. A. JONES, W. BAO, A. F. SMIT et T. J. WHEELER. 2016, «The Dfam database of repetitive DNA families», *Nucleic Acids Research*, vol. 44, D1, doi:10.1093/nar/gkv1272, p. D81–D89, ISSN 0305-1048, 1362-4962. URL https://academic.oup.com/nar/artic

le-lookup/doi/10.1093/nar/gkv1272. 69

HUERTA-CEPAS, J., F. SERRA et P. BORK. 2016, «ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data», *Molecular Biology and Evolution*, vol. 33, 6, doi:10.1093/molbev/msw046, p. 1635–1638, ISSN 0737-4038. URL https://academic.oup.com/mbe/article/33/6/1635/2579822. 102

JAKŠIĆ, A. M., R. KOFLER et C. SCHLÖTTERER. 2017, «Regulation of transposable elements: Interplay between TE-encoded regulatory sequences and host-specific trans-acting factors in Drosophila melanogaster», *Molecular Ecology*, vol. 26, 19, doi:10.1111/mec.14259, p. 5149–5159, ISSN 1365-294X. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.14259. 3, 8

JEFFREYS, H. 1961, *Theory of Probability, Ed. 3 Oxford University Press*, Oxford. 73, 74

KAPITONOV, V. V. et J. JURKA. 2008, «A universal classification of eukaryotic transposable elements implemented in Repbase», *Nature Reviews. Genetics*, vol. 9, 5, doi:10.1038/nrg2165-c1, p. 411–412; author reply 414, ISSN 1471-0064. 4

KAPUN, M., M. G. BARRÓN, F. STAUBACH, J. VIEIRA, D. J. OBBARD, C. GOUBERT, O. ROTA-STABELLI, M. KANKARE, A. HAUDRY, R. A. W. WIBERG, L. WAIDELE, I. KOZERETSKA, E. G. PASYUKOVA, V. LOESCHCKE, M. PASCUAL, C. P. VIEIRA, S. SERGA, C. MONTCHAMP-MOREAU, J. ABBOTT, P. GIBERT, D. PORCELLI, N. POSNIEN, S. GRATH, SUCENA, A. O. BERGLAND, M. P. G. GUERREIRO, B. S. ONDER, E. ARGYRIDOU, L. GUIO, M. F. SCHOU, B. DEPLANCKE, C. VIEIRA, M. G. RITCHIE, B. J. ZWAAN, E. TAUBER, D. J. ORENGO, E. PUERMA, M. AGUADÉ, P. S. SCHMIDT, J. PARSCH, A. J. BETANCOURT, T. FLATT et J. GONZÁLEZ. 2018, «Genomic analysis of European Drosophila melanogaster populations on a dense spatial scale reveals longitudinal population structure and continent-wide selection», *bioRxiv*, doi:10.1101/313759, p. 313 759. URL https://www.biorxiv.org/content/10.1101/313759v1. 7, 67, 119

KAPUSTA, A. et A. SUH. 2017, «Evolution of bird genomes—a transposon's-eye view», *Annals of the New York Academy of Sciences*, vol. 1389, 1, doi:10.1111/nyas.13295, p. 164–185, ISSN 1749-6632. URL https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/nyas.13295. 101

KEIGHTLEY, P. D., R. W. NESS, D. L. HALLIGAN et P. R. HADDRILL. 2014, «Estimation of the Spontaneous Mutation Rate per Nucleotide Site in a Drosophila melanogaster Full-Sib Family», *Genetics*, vol. 196, 1, doi:10.1534/genetics.113.158758, p. 313–320, ISSN 0016-6731. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3872194/. 158

KIDWELL, M. G., J. F. KIDWELL et J. A. SVED. 1977, «Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: A Syndrome of Aberrant Traits Including Mutation, Sterility and Male Recombination», *Genetics*, vol. 86, 4, p. 813–833, ISSN 0016-6731. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1213713/. 5

KIM, B. Y., J. R. WANG, D. E. MILLER, O. BARMINA, E. DELANEY, A. THOMPSON, A. A. COMEAULT, D. PEEDE, E. R. R. D'AGOSTINO, J. PELAEZ, J. M. AGUILAR, D. HAJI, T. MATSUNAGA, E. E. ARMSTRONG, M. ZYCH, Y. OGAWA, M. STAMENKOVIĆ-RADAK, M. JELIĆ, M. S. VESELINOVIĆ, M. TANASKOVIĆ, P. ERIĆ, J.-J. GAO, T. K. KATOH, M. J. TODA, H. WATABE, M. WATADA, J. S. DAVIS, L. C. MOYLE, G. MANOLI, E. BERTOLINI, V. KOŠŤÁL, R. S. HAWLEY, A. TAKAHASHI,

C. D. JONES, D. K. PRICE, N. WHITEMAN, A. KOPP, D. R. MATUTE et D. A. PETROV. 2020, «Highly contiguous assemblies of 101 drosophilid genomes», *bioRxiv*, doi:10.1101/2020.12.14.422775, p. 2020.12.14.422 775. URL https://www.biorxiv.org/content/10.1101/2020.12.14.422775v1, publisher: Cold Spring Harbor Laboratory Section: New Results. 89, 97, 120

KOCH, P., M. PLATZER et B. R. DOWNIE. 2014, «RepARK–de novo creation of repeat libraries from whole-genome NGS reads», *Nucleic Acids Research*, vol. 42, 9, doi:10.1093/nar/gku210, p. e80, ISSN 1362-4962. 16

KOFLER, R. 2018, «SimulaTE: simulating complex landscapes of transposable elements of populations», *Bioinformatics*, vol. 34, 8, doi:10.1093/bioinformatics/btx772, p. 1419–1420, ISSN 1367-4803. URL https://academic.oup.com/bioinformatics/article/34/8/1419/4665422. 18, 82

KOFLER, R. 2020, «piRNA Clusters Need a Minimum Size to Control Transposable Element Invasions», *Genome Biology and Evolution*, vol. 12, 5, doi:10.1093/gbe/evaa064, p. 736–749. URL https://academic.oup.com/gbe/article/12/5/736/5812784. 9, 98

KOFLER, R., A. J. BETANCOURT et C. SCHLÖTTERER. 2012, «Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in Drosophila melanogaster», *PLoS genetics*, vol. 8, 1, doi:10.1371/journal.pgen.1002487, p. e1002 487, ISSN 1553-7404. 13, 55, 115

KOFLER, R., D. GÓMEZ-SÁNCHEZ et C. SCHLÖTTERER. 2016, «PoPoolationTE2: Comparative Population Genomics of Transposable Elements Using Pool-Seq», *Molecular Biology and Evolution*, vol. 33, 10, doi:10.1093/molbev/msw137, p. 2759–2764, ISSN 0737-4038. URL https://academic.oup.com/mbe/article/33/10/2759/2925581. 18, 26, 71, 74, 123

KOFLER, R., T. HILL, V. NOLTE, A. J. BETANCOURT et C. SCHLÖTTERER. 2015a, «The recent invasion of natural Drosophila simulans populations by the P-element», *Proceedings of the National Academy of Sciences*, vol. 112, 21, doi:10.1073/pnas.1500758112, p. 6659–6663, ISSN 0027-8424, 1091-6490. URL https://www.pnas.org/content/112/21/6659. 64

KOFLER, R., V. NOLTE et C. SCHLÖTTERER. 2015b, «Tempo and Mode of Transposable Element Activity in Drosophila», *PLOS Genetics*, vol. 11, 7, doi:10.1371/journal.pgen.1005406, p. e1005 406, ISSN 1553-7404. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005406. 55, 56, 65, 114

KOFLER, R., P. OROZCO-TERWENGEL, N. D. MAIO, R. V. PANDEY, V. NOLTE, A. FUTSCHIK, C. KOSIOL et C. SCHLÖTTERER. 2011, «PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals», *PLOS ONE*, vol. 6, 1, doi:10.1371/journal.pone.0015925, p. e15 925, ISSN 1932-6203. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0015925. 71, 72

KOFLER, R., K.-A. SENTI, V. NOLTE, R. TOBLER et C. SCHLÖTTERER. 2018, «Molecular dissection of a natural transposable element invasion», *Genome Research*, vol. 28, 6, doi:10.1101/gr.228627.117, p. 824–835, ISSN 1088-9051. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5991514/. 8, 98

KOPP, A. et J. R. TRUE. 2002, «Evolution of male sexual characters in the Oriental Drosophila melanogaster species group»,

Evolution & Development, vol. 4, 4, doi:https://doi.org/10.1046/j.1525-142X.2002.02017.x, p. 278–291, ISSN 1525-142X. URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1525-142X.2002.02017.x. 21

KRZYWINSKI, M. I., J. E. SCHEIN, I. BIROL, J. CONNORS, R. GASCOYNE, D. HORSMAN, S. J. JONES et M. A. MARRA. 2009, «Circos: An information aesthetic for comparative genomics», *Genome Research*, doi:10.1101/gr.092759.109, ISSN 1088-9051, 1549-5469. URL http://genome.cshlp.org/content/early/2009/06/15/gr.092759.109. 70, 101

KURTZ, S., A. PHILLIPPY, A. L. DELCHER, M. SMOOT, M. SHUMWAY, C. ANTONESCU et S. L. SALZBERG. 2004, «Versatile and open software for comparing large genomes», *Genome Biology*, vol. 5, 2, doi:10.1186/gb-2004-5-2-r12, p. R12, ISSN 1474-760X. 70, 92, 101

LANDER, E. S., L. M. LINTON, B. BIRREN, C. NUSBAUM, M. C. ZODY, J. BALDWIN, K. DEVON, K. DEWAR, M. DOYLE, W. FITZHUGH, R. FUNKE, D. GAGE, K. HARRIS, A. HEAFORD, J. HOWLAND, L. KANN, J. LEHOCZKY, R. LEVINE, P. MCEWAN, K. MCKERNAN, J. MELDRIM, J. P. MESIROV, C. MIRANDA, W. MORRIS, J. NAYLOR, C. RAYMOND, M. ROSETTI, R. SANTOS, A. SHERIDAN, C. SOUGNEZ, Y. STANGE-THOMANN, N. STOJANOVIC, A. SUBRAMANIAN, D. WYMAN, J. ROGERS, J. SULSTON, R. AINSCOUGH, S. BECK, D. BENTLEY, J. BURTON, C. CLEE, N. CARTER, A. COULSON, R. DEADMAN, P. DELOUKAS, A. DUNHAM, I. DUNHAM, R. DURBIN, L. FRENCH, D. GRAFHAM, S. GREGORY, T. HUBBARD, S. HUMPHRAY, A. HUNT, M. JONES, C. LLOYD, A. MCMURRAY, L. MATTHEWS, S. MERCER, S. MILNE, J. C. MULLIKIN, A. MUNGALL, R. PLUMB, M. ROSS, R. SHOWNKEEN, S. SIMS, R. H. WATERSTON, R. K. WILSON, L. W. HILLIER, J. D. MCPHERSON, M. A. MARRA, E. R. MARDIS, L. A. FULTON, A. T. CHINWALLA, K. H. PEPIN, W. R. GISH, S. L. CHISSOE, M. C. WENDL, K. D. DELEHAUNTY, T. L. MINER, A. DELEHAUNTY, J. B. KRAMER, L. L. COOK, R. S. FULTON, D. L. JOHNSON, P. J. MINX, S. W. CLIFTON, T. HAWKINS, E. BRANSCOMB, P. PREDKI, P. RICHARDSON, S. WENNING, T. SLEZAK, N. DOGGETT, J. F. CHENG, A. OLSEN, S. LUCAS, C. ELKIN, E. UBERBACHER, M. FRAZIER, R. A. GIBBS, D. M. MUZNY, S. E. SCHERER, J. B. BOUCK, E. J. SODERGREN, K. C. WORLEY, C. M. RIVES, J. H. GORRELL, M. L. METZKER, S. L. NAYLOR, R. S. KUCHERLAPATI, D. L. NELSON, G. M. WEINSTOCK, Y. SAKAKI, A. FUJIYAMA, M. HATTORI, T. YADA, A. TOYODA, T. ITOH, C. KAWAGOE, H. WATANABE, Y. TOTOKI, T. TAYLOR, J. WEISSENBACH, R. HEILIG, W. SAURIN, F. ARTIGUENAVE, P. BROTTIER, T. BRULS, E. PELLETIER, C. ROBERT, P. WINCKER, D. R. SMITH, L. DOUCETTE-STAMM, M. RUBENFIELD, K. WEINSTOCK, H. M. LEE, J. DUBOIS, A. ROSENTHAL, M. PLATZER, G. NYAKATURA, S. TAUDIEN, A. RUMP, H. YANG, J. YU, J. WANG, G. HUANG, J. GU, L. HOOD, L. ROWEN, A. MADAN, S. QIN, R. W. DAVIS, N. A. FEDERSPIEL, A. P. ABOLA, M. J. PROCTOR, R. M. MYERS, J. SCHMUTZ, M. DICKSON, J. GRIMWOOD, D. R. COX, M. V. OLSON, R. KAUL, C. RAYMOND, N. SHIMIZU, K. KAWASAKI, S. MINOSHIMA, G. A. EVANS, M. ATHANASIOU, R. SCHULTZ, B. A. ROE, F. CHEN, H. PAN, J. RAMSER, H. LEHRACH, R. REINHARDT, W. R. MCCOMBIE, M. DE LA BASTIDE, N. DEDHIA, H. BLÖCKER, K. HORNISCHER, G. NORDSIEK, R. AGARWALA, L. ARAVIND, J. A. BAILEY, A. BATEMAN, S. BATZOGLOU, E. BIRNEY, P. BORK, D. G. BROWN, C. B. BURGE, L. CERUTTI, H. C. CHEN, D. CHURCH, M. CLAMP, R. R. COPLEY, T. DOERKS, S. R. EDDY, E. E. EICHLER, T. S. FUREY, J. GALAGAN, J. G. GILBERT, C. HARMON, Y. HAYASHIZAKI, D. HAUSSLER,

H. HERMJAKOB, K. HOKAMP, W. JANG, L. S. JOHNSON, T. A. JONES, S. KASIF, A. KASPRYZK, S. KENNEDY, W. J. KENT, P. KITTS, E. V. KOONIN, I. KORF, D. KULP, D. LANCET, T. M. LOWE, A. MCLYSAGHT, T. MIKKELSEN, J. V. MORAN, N. MULDER, V. J. POLLARA, C. P. PONTING, G. SCHULER, J. SCHULTZ, G. SLATER, A. F. SMIT, E. STUPKA, J. SZUSTAKOWKI, D. THIERRY-MIEG, J. THIERRY-MIEG, L. WAGNER, J. WALLIS, R. WHEELER, A. WILLIAMS, Y. I. WOLF, K. H. WOLFE, S. P. YANG, R. F. YEH, F. COLLINS, M. S. GUYER, J. PETERSON, A. FELSENFELD, K. A. WETTERSTRAND, A. PATRINOS, M. J. MORGAN, P. DE JONG, J. J. CATANESE, K. OSOEGAWA, H. SHIZUYA, S. CHOI, Y. J. CHEN, J. SZUSTAKOWKI et INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. 2001, «Initial sequencing and analysis of the human genome», *Nature*, vol. 409, 6822, doi:10.1038/35057062, p. 860–921, ISSN 0028-0836. 13, 55

LANGE, J. D. et J. E. POOL. 2016, «A haplotype method detects diverse scenarios of local adaptation from genomic sequence variation», *Molecular Ecology*, vol. 25, 13, doi:10.1111/mec.13671, p. 3081–3100, ISSN 1365-294X. 67

LAVERGNE, S. et J. MOLOFSKY. 2007, «Increased genetic variation and evolutionary potential drive the success of an invasive grass», *Proceedings of the National Academy of Sciences*, vol. 104, 10, doi:10.1073/pnas.0607324104, p. 3883–3888, ISSN 0027-8424, 1091-6490. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0607324104. 56

LEE, Y. C. G. et G. H. KARPEN. 2017, «Pervasive epigenetic effects of Drosophila euchromatic transposable elements impact their evolution», *eLife*, vol. 6, doi:10.7554/eLife.25762, ISSN 2050-084X. 5, 6, 65, 98

LEFÉBURE, T., C. MORVAN, F. MALARD, C. FRANÇOIS, L. KONECNY-DUPRÉ, L. GUÉGUEN, M. WEISS-GAYET, A. SEGUIN-ORLANDO, L. ERMINI, C. D. SARKISSIAN, N. P. CHARRIER, D. EME, F. MERMILLOD-BLONDIN, L. DURET, C. VIEIRA, L. ORLANDO et C. J. DOUADY. 2017, «Less effective selection leads to larger genomes», *Genome Research*, vol. 27, 6, doi:10.1101/gr.212589.116, p. 1016–1028, ISSN 1549-5469. 7, 88

LERAT, E., M. FABLET, L. MODOLO, H. LOPEZ-MAESTRE et C. VIEIRA. 2016, «TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes», *Nucleic Acids Research*, doi:10.1093/nar/gkw953, p. gkw953, ISSN 0305-1048, 1362-4962. URL http://nar.oxfordjournals.org/content/early/2016/10/19/nar.gkw953. 123

LERAT, E., C. GOUBERT, S. GUIRAO-RICO, M. MERENCIANO, A.-B. DUFOUR, C. VIEIRA et J. GONZÁLEZ. 2019, «Population-specific dynamics and selection patterns of transposable element insertions in European natural populations», *Molecular Ecology*, vol. 28, 6, doi:10.1111/mec.14963, p. 1506–1522, ISSN 1365-294X. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.14963. 8, 18, 19, 56, 119, 123, 159

LI, H. et R. DURBIN. 2010, «Fast and accurate long-read alignment with Burrows-Wheeler transform», *Bioinformatics (Oxford, England)*, vol. 26, 5, doi:10.1093/bioinformatics/btp698, p. 589–595, ISSN 1367-4811. 71

LI, H., B. HANDSAKER, A. WYSOKER, T. FENNELL, J. RUAN, N. HOMER, G. MARTH, G. ABECASIS, R. DURBIN et 1000 GENOME PROJECT DATA PROCESSING SUBGROUP. 2009, «The Sequence Alignment/Map format and SAMtools», *Bioinformatics (Oxford, England)*, vol. 25, 16, doi:10.1093/bioinformatics/btp352, p. 2078–2079, ISSN 1367-4811. 72

LI, Z.-W., X.-H. HOU, J.-F. CHEN, Y.-C. XU, Q. WU, J. GONZÁLEZ et Y.-L. GUO. 2018, «Transposable Elements Contribute to the Adaptation of Arabidopsis thaliana», *Genome Biology and Evolution*, vol. 10, 8, doi:10.1093/gbe/evy171, p. 2140–2150, ISSN 1759-6653. URL https://academic.oup.com/gbe/article/10/8/2140/5068481. 9, 12, 56, 115

LIN, X., L. LONG, X. SHAN, S. ZHANG, S. SHEN et B. LIU. 2006, «In planta mobilization of mPing and its putative autonomous element Pong in rice by hydrostatic pressurization», *Journal of Experimental Botany*, vol. 57, 10, doi:10.1093/jxb/erj203, p. 2313–2323, ISSN 0022-0957. URL https://academic.oup.com/jxb/article/57/10/2313/473500. 8

LITTLE, C. M., T. W. CHAPMAN et N. K. HILLIER. 2020, «Plasticity Is Key to Success of Drosophila suzukii (Diptera: Drosophilidae) Invasion», *Journal of Insect Science*, vol. 20, 3, doi:10.1093/jisesa/ieaa034. URL https://academic.oup.com/jinsectscience/article/20/3/5/5837529. 25

LOCKTON, S., J. ROSS-IBARRA et B. S. GAUT. 2008, «Demography and weak selection drive patterns of transposable element diversity in natural populations of Arabidopsis lyrata», *Proceedings of the National Academy of Sciences*, vol. 105, 37, doi:10.1073/pnas.0804671105, p. 13 965–13 970, ISSN 0027-8424, 1091-6490. URL https://www.pnas.org/content/105/37/13965. 7, 115

LYNCH, M. et J. S. CONERY. 2003, «The Origins of Genome Complexity», *Science*, vol. 302, 5649, doi:10.1126/science.1089370, p. 1401–1404, ISSN 0036-8075, 1095-9203. URL https://science.sciencemag.org/content/302/5649/1401. 7, 55, 66, 88

MACPHERSON, J. M., G. SELLA, J. C. DAVIS et D. A. PETROV. 2007, «Genomewide Spatial Correspondence Between Nonsynonymous Divergence and Neutral Polymorphism Reveals Extensive Adaptation in Drosophila», *Genetics*, vol. 177, 4, doi:10.1534/genetics.107.080226, p. 2083–2099, ISSN 0016-6731, 1943-2631. URL https://www.genetics.org/content/177/4/2083, publisher: Genetics Section: Investigations. 116

MARIN, P., J. GENITONI, D. BARLOY, S. MAURY, P. GIBERT, C. K. GHALAMBOR et C. VIEIRA. 2020, «Biological invasion: The influence of the hidden side of the (epi)genome», *Functional Ecology*, vol. 34, 2, doi:10.1111/1365-2435.13317, p. 385–400, ISSN 1365-2435. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2435.13317. 56

MARÇAIS, G. et C. KINGSFORD. 2011, «A fast, lock-free approach for efficient parallel counting of occurrences of k-mers», *Bioinformatics*, vol. 27, 6, doi:10.1093/bioinformatics/btr011, p. 764–770, ISSN 1367-4803. URL https://academic.oup.com/bioinformatics/article/27/6/764/234905. 100

MATSUMURA, S. 1931, *6000 Illustrated Insects of Japan-empire*, The Toko-Shoin. Google-Books-ID: EoYqmgEACAAJ. 20

MCCLINTOCK, B. 1950, «The origin and behavior of mutable loci in maize», *Proceedings of the National Academy of Sciences*, vol. 36, 6, doi:10.1073/pnas.36.6.344, p. 344–355, ISSN 0027-8424, 1091-6490. URL https://www.pnas.org/content/36/6/344. 3, 4

MEDSTRAND, P., L. N. VAN DE LAGEMAAT et D. L. MAGER. 2002, «Retroelement distributions in the human genome: variations associated with age and proximity to genes», *Genome Research*, vol. 12, 10, doi:10.1101/gr.388902, p. 1483–1495, ISSN 1088-9051. 12, 13, 65, 118

MES, T. H. M. et M. DOELEMAN. 2006, «Positive Selection on Transposase Genes of Insertion Sequences in the Crocosphaera watsonii Genome», *Journal of Bacteriology*, vol. 188, 20, doi:10.1128/JB.01021-06, p. 7176–7185, ISSN 0021-9193. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1636226/. 121

MI, S., X. LEE, X. LI, G. M. VELDMAN, H. FINNERTY, L. RACIE, E. LAVALLIE, X. Y. TANG, P. EDOUARD, S. HOWES, J. C. KEITH et J. M. MCCOY. 2000, «Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis», *Nature*, vol. 403, 6771, doi:10.1038/35001608, p. 785–789, ISSN 0028-0836. 55

MILLER, D. E., C. STABER, J. ZEITLINGER et R. S. HAWLEY. 2018, «Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing», *G3: Genes|Genomes|Genetics*, vol. 8, 10, doi:10.1534/g3.118.200160, p. 3131–3141, ISSN 2160-1836. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6169393/. 65, 100, 101, 119

MITSUI, H., K. BEPPU et M. T. KIMURA. 2010, «Seasonal life cycles and resource uses of flower- and fruit-feeding drosophilid flies (Diptera: Drosophilidae) in central Japan», *Entomological Science*, vol. 13, 1, doi:https://doi.org/10.1111/j.1479-8298.2010.00372.x, p. 60–67, ISSN 1479-8298. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1479-8298.2010.00372.x. 22

MODOLO, L. et E. LERAT. 2015, «UrQt: an efficient software for the Unsupervised Quality trimming of NGS data», *BMC Bioinformatics*, vol. 16, 1, doi:10.1186/s12859-015-0546-8, p. 137, ISSN 1471-2105. URL https://doi.org/10.1186/s12859-015-0546-8. 100

MOHAMED, M., N. T.-M. DANG, Y. OGYAMA, N. BURLET, B. MUGAT, M. BOULESTEIX, V. MÉREL, P. VEBER, J. SALCES-ORTIZ, D. SEVERAC, A. PÉLISSON, C. VIEIRA, F. SABOT, M. FABLET et S. CHAMBEYRON. 2020, «A Transposon Story: From TE Content to TE Dynamic Invasion of Drosophila Genomes Using the Single-Molecule Sequencing Technology from Oxford Nanopore», *Cells*, vol. 9, 8, doi:10.3390/cells9081776, p. 1776. URL https://www.mdpi.com/2073-4409/9/8/1776, number: 8 Publisher: Multidisciplinary Digital Publishing Institute. 9

MONTGOMERY, E., B. CHARLESWORTH et C. H. LANGLEY. 1987, «A test for the role of natural selection in the stabilization of transposable element copy number in a population of Drosophila melanogaster», *Genetical Research*, vol. 89, 5-6, doi:10.1017/S0016672308009634, p. 435–445. 5, 6

MUGAL, C. F., J. B. WOLF et I. KAJ. 2014, «Why Time Matters: Codon Evolution and the Temporal Dynamics of dN/dS», *Molecular Biology and Evolution*, vol. 31, 1, doi:10.1093/molbev/mst192, p. 212–231, ISSN 0737-4038. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3879453/. 97, 113

MUSE, S. V. et B. S. GAUT. 1994, «A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome», *Molecular Biology and Evolution*, vol. 11, 5, doi:10.1093/oxfordjournals.molbev.a040152, p. 715–724, ISSN 0737-4038. 102

MUÑOZ-LÓPEZ, M. et J. L. GARCÍA-PÉREZ. 2010, «DNA Transposons: Nature and Applications in Genomics», *Current Genomics*, vol. 11, 2, doi:10.2174/138920210790886871, p. 115–128, ISSN 1389-2029. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2874221/. 3

MÉREL, V., M. BOULESTEIX, M. FABLET et C. VIEIRA. 2020a, «Transposable elements in

Drosophila», *Mobile DNA*, vol. 11, 1, doi:10.1186/s13100-020-00213-z, p. 23, ISSN 1759-8753. URL https://doi.org/10.1186/s13100-020-00213-z. 55, 88, 96

MÉREL, V., P. GIBERT, I. BUCH, V. R. RADA, A. ESTOUP, M. GAUTIER, M. FABLET, M. BOULESTEIX et C. VIEIRA. 2020b, «The worldwide invasion of Drosophila suzukii is accompanied by a large increase of transposable element load and a small number of putatively adaptive insertions», *bioRxiv*, doi:10.1101/2020.11.06.370932, p. 2020.11.06.370932. URL https://www.biorxiv.org/content/10.1101/2020.11.06.370932v1, publisher: Cold Spring Harbor Laboratory Section: New Results. 88, 92, 96, 100, 101

NARANJO-LÁZARO, J., M. MELLÍN-ROSAS, V. GONZÁLEZ-PADILLA, J. SÁNCHEZ-GONZÁLEZ, G. MORENO-CARRILLO et H. ARREDONDO-BERNAL. 2014, «Susceptibility of Drosophila suzukii Matsumura (Diptera: Drosophilidae) to Entomophatogenic Fungi», *Southwestern Entomologist*, vol. 39, doi:10.3958/059.039.0119, p. 201–203. 22

NARDON, C., G. DECELIERE, C. LŒVENBRUCK, M. WEISS, C. VIEIRA et C. BIÉMONT. 2005, «Is genome size influenced by colonization of new environments in dipteran species?», *Molecular Ecology*, vol. 14, 3, doi:10.1111/j.1365-294X.2005.02457.x, p. 869–878, ISSN 1365-294X. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-294X.2005.02457.x. 66

NAVILLE, M., S. HENRIET, I. WARREN, S. SUMIC, M. REEVE, J.-N. VOLFF et D. CHOURROUT. 2019, «Massive Changes of Genome Size Driven by Expansions of Non-autonomous Transposable Elements», *Current Biology*, vol. 29, 7, doi:10.1016/j.cub.2019.01.080, p. 1161–1168.e6, ISSN 0960-9822. URL https://www.cell.com/current-biology/abstract/S0960-9822(19)30139-3, publisher: Elsevier. 87

NGUYEN, L.-T., H. A. SCHMIDT, A. VON HAESELER et B. Q. MINH. 2015, «IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies», *Molecular Biology and Evolution*, vol. 32, 1, doi:10.1093/molbev/msu300, p. 268–274, ISSN 0737-4038. URL https://academic.oup.com/mbe/article/32/1/268/2925592. 99

NIKITIN, A. G. et R. C. WOODRUFF. 1995, «Somatic movement of the mariner transposable element and lifespan of Drosophila species», *Mutation Research/DNAging*, vol. 338, 1, doi:10.1016/0921-8734(95)00010-4, p. 43–49, ISSN 0921-8734. URL http://www.sciencedirect.com/science/article/pii/0921873495000104. 5, 64, 87

NIU, X.-M., Y.-C. XU, Z.-W. LI, Y.-T. BIAN, X.-H. HOU, J.-F. CHEN, Y.-P. ZOU, J. JIANG, Q. WU, S. GE, S. BALASUBRAMANIAN et Y.-L. GUO. 2019, «Transposable elements drive rapid phenotypic variation in Capsella rubella», *Proceedings of the National Academy of Sciences*, vol. 116, 14, doi:10.1073/pnas.1811498116, p. 6908–6913, ISSN 0027-8424, 1091-6490. URL https://www.pnas.org/content/116/14/6908. 10, 11, 55, 56

NOVÁK, P., P. NEUMANN et J. MACAS. 2010, «Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data», *BMC bioinformatics*, vol. 11, doi:10.1186/1471-2105-11-378, p. 378, ISSN 1471-2105. 16

NUZHDIN, S. V. 1999, «Sure facts, speculations, and open questions about the evolution of transposable element copy number», *Genetica*, vol. 107, 1, doi:10.1023/A:1003957323876,

p. 129, ISSN 1573-6857. URL https://doi.org/10.1023/A:1003957323876. 5

OLAZCUAGA, L., A. LOISEAU, H. PARRINELLO, M. PARIS, A. FRAIMOUT, C. GUEDOT, L. M. DIEPENBROCK, M. KENIS, J. ZHANG, X. CHEN, N. BOROWIEC, B. FACON, H. VOGT, D. K. PRICE, H. VOGEL, B. PRUD'HOMME, A. ESTOUP et M. GAUTIER. 2020, «A Whole-Genome Scan for Association with Invasion Success in the Fruit Fly Drosophila suzukii Using Contrasts of Allele Frequencies Corrected for Population Structure», *Molecular Biology and Evolution*, vol. 37, 8, doi:10.1093/molbev/msaa098, p. 2369–2385, ISSN 0737-4038. URL https://academic.oup.com/mbe/article/37/8/2369/5821433. 25, 61, 62, 63, 64, 67, 68, 71, 72, 73, 74, 82, 116

OMETTO, L., A. CESTARO, S. RAMASAMY, A. GRASSI, S. REVADI, S. SIOZIOS, M. MORETTO, P. FONTANA, C. VAROTTO, D. PISANI, T. DEKKER, N. WROBEL, R. VIOLA, I. PERTOT, D. CAVALIERI, M. BLAXTER, G. ANFORA et O. ROTA-STABELLI. 2013, «Linking Genomics and Ecology to Investigate the Complex Evolution of an Invasive Drosophila Pest», *Genome Biology and Evolution*, vol. 5, 4, doi:10.1093/gbe/evt034, p. 745–757, ISSN 1759-6653. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3641628/. 64, 86, 88

ORGEL, L. E. et F. H. C. CRICK. 1980, «Selfish DNA: the ultimate parasite», *Nature*, vol. 284, 5757, doi:10.1038/284604a0, p. 604–607, ISSN 1476-4687. URL https://www.nature.com/articles/284604a0. 4, 55, 87

OU, S., W. SU, Y. LIAO, K. CHOUGULE, J. R. A. AGDA, A. J. HELLINGA, C. S. B. LUGO, T. A. ELLIOTT, D. WARE, T. PETERSON, N. JIANG, C. N. HIRSCH et M. B. HUFFORD. 2019, «Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline», *Genome Biology*, vol. 20, 1, doi:10.1186/s13059-019-1905-y, p. 275, ISSN 1474-760X. URL https://doi.org/10.1186/s13059-019-1905-y. 3

OZATA, D. M., I. GAINETDINOV, A. ZOCH, D. OCARROLL et P. D. ZAMORE. 2019, «PIWI-interacting RNAs: small RNAs with big functions», *Nature Reviews Genetics*, vol. 20, 2, doi:10.1038/s41576-018-0073-3, p. 89–108, ISSN 1471-0064. URL https://www-nature-com.inee.bib.cnrs.fr/articles/s41576-018-0073-3. 6, 73, 88, 97, 102

PARIS, M., R. BOYER, R. JAENICHEN, J. WOLF, M. KARAGEORGI, J. GREEN, M. CAGNON, H. PARINELLO, A. ESTOUP, M. GAUTIER, N. GOMPEL et B. PRUD'HOMME. 2020, «Near-chromosome level genome assembly of the fruit pest Drosophila suzukii using long-read sequencing», *Scientific Reports*, vol. 10, 1, doi:10.1038/s41598-020-67373-z, p. 11 227, ISSN 2045-2322. URL https://www.nature.com/articles/s41598-020-67373-z. 19, 54, 57, 58, 59, 64, 69, 74, 100, 121

PASYUKOVA, E. G. et S. V. NUZHDIN. 1993, «Doc and copia instability in an isogenic Drosophila melanogaster stock», *Molecular & general genetics: MGG*, vol. 240, 2, doi:10.1007/bf00277071, p. 302–306, ISSN 0026-8925. 8, 56, 67

PENG, Y., H. C. M. LEUNG, S. M. YIU et F. Y. L. CHIN. 2012, «IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth», *Bioinformatics*, vol. 28, 11, doi:10.1093/bioinformatics/bts174, p. 1420–1428, ISSN 1367-4803. URL https://doi.org/10.1093/bioinformatics/bts174. 99

PENNINGS, P. S. et J. HERMISSON. 2006, «Soft Sweeps III: The Signature of Positive Selection from Recurrent Mutation», *PLoS Genetics*, vol. 2, 12, doi:10.1371/journal.pgen.0020186, ISSN 1553-7390. URL https://www.ncbi.n

lm.nih.gov/pmc/articles/PMC1698945/. 67

PETROV, D. A., Y. T. AMINETZACH, J. C. DAVIS, D. BENSASSON et A. E. HIRSH. 2003, «Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila», *Molecular Biology and Evolution*, vol. 20, 6, doi:10.1093/molbev/msg102, p. 880–892, ISSN 0737-4038. 5, 6, 7, 65, 88, 98

PETROV, D. A., A.-S. FISTON-LAVIER, M. LIPATOV, K. LENKOV et J. GONZÁLEZ. 2011, «Population genomics of transposable elements in Drosophila melanogaster», *Molecular Biology and Evolution*, vol. 28, 5, doi:10.1093/molbev/msq337, p. 1633–1644, ISSN 1537-1719. 5

PICARD, G. 1976, «Non-Mendelian Female Sterility in DROSOPHILA MELANOGASTER : Hereditary Transmission of I Factor», *Genetics*, vol. 83, 1, p. 107–123, ISSN 0016-6731. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1213493/. 5

PRENTIS, P., D. SIGG, S. RAGHU, K. DHILEEPAN, A. PAVASOVIC et A. LOWE. 2009, «Understanding invasion history: Genetic structure and diversity of two globally invasive plants and implications for their management», *Diversity and Distributions*, vol. 15, doi:10.1111/j.1472-4642.2009.00592.x. 68

PRICE, A. L., N. C. JONES et P. A. PEVZNER. 2005, «De novo identification of repeat families in large genomes», *Bioinformatics (Oxford, England)*, vol. 21 Suppl 1, doi:10.1093/bioinformatics/bti1018, p. i351–358, ISSN 1367-4803. 16, 69, 100

PRUD'HOMME, B., N. GOMPEL, A. ROKAS, V. A. KASSNER, T. M. WILLIAMS, S.-D. YEH, J. R. TRUE et S. B. CARROLL. 2006, «Repeated morphological evolution through cis -regulatory changes in a pleiotropic gene», *Nature*, vol. 440, 7087, doi:10.1038/nature04597, p. 1050–1053, ISSN 1476-4687. URL https://www.nature.com/articles/nature04597. 21

QUADRANA, L., A. BORTOLINI SILVEIRA, G. F. MAYHEW, C. LEBLANC, R. A. MARTIENSSEN, J. A. JEDDELOH et V. COLOT. 2016, «The Arabidopsis thaliana mobilome and its impact at the species level», *eLife*, vol. 5, doi:10.7554/eLife.15716, p. e15 716, ISSN 2050-084X. URL https://doi.org/10.7554/eLife.15716. 8, 11, 67, 73, 120, 123

QUESNEVILLE, H., D. NOUAUD et D. ANXOLABÉHÈRE. 2003, «Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes», *Journal of Molecular Evolution*, vol. 57 Suppl 1, doi:10.1007/s00239-003-0007-2, p. S50–59, ISSN 0022-2844. 16

RAHMAN, R., G.-W. CHIRN, A. KANODIA, Y. A. SYTNIKOVA, B. BREMBS, C. M. BERGMAN et N. C. LAU. 2015, «Unique transposon landscapes are pervasive across Drosophila melanogaster genomes», *Nucleic Acids Research*, vol. 43, 22, doi:10.1093/nar/gkv1193, p. 10 655–10 672, ISSN 0305-1048. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4678822/. 18

RANWEZ, V., E. J. P. DOUZERY, C. CAMBON, N. CHANTRET et F. DELSUC. 2018, «MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons», *Molecular Biology and Evolution*, vol. 35, 10, doi:10.1093/molbev/msy159, p. 2582–2584, ISSN 0737-4038. URL https://academic.oup.com/mbe/article/35/10/2582/5079334. 99

REBOLLO, R., M. T. ROMANISH et D. L. MAGER. 2012, «Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes», *Annual Review of Genetics*, vol. 46, 1, doi:10.1146/annurev-genet-110711-155621, p.

21–42, ISSN 0066-4197, 1545-2948. URL http://www.annualreviews.org/doi/10.1146/annurev-genet-110711-155621. 5, 6

RECH, G. E., M. BOGAERTS-MÁRQUEZ, M. G. BARRÓN, M. MERENCIANO, J. L. VILLANUEVA-CAÑAS, V. HORVÁTH, A.-S. FISTON-LAVIER, I. LUYTEN, S. VENKATARAM, H. QUESNEVILLE, D. A. PETROV et J. GONZÁLEZ. 2019, «Stress response, behavior, and development are shaped by transposable element-induced mutations in Drosophila», *PLoS genetics*, vol. 15, 2, doi:10.1371/journal.pgen.1007900, p. e1007 900, ISSN 1553-7404. 12, 56, 67, 68, 115

REVADI, S., S. LEBRETON, P. WITZGALL, G. ANFORA, T. DEKKER et P. G. BECHER. 2015, «Sexual Behavior of Drosophila suzukii», *Insects*, vol. 6, 1, doi:10.3390/insects6010183, p. 183–196, ISSN 2075-4450. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4553537/. 21

RHOADS, A. et K. F. AU. 2015, «PacBio Sequencing and Its Applications», *Genomics, Proteomics & Bioinformatics*, vol. 13, 5, doi:10.1016/j.gpb.2015.08.002, p. 278–289, ISSN 1672-0229. URL http://www.sciencedirect.com/science/article/pii/S1672022915001345. 14

RISHISHWAR, L., L. WANG, J. WANG, S. V. YI, J. LACHANCE et I. K. JORDAN. 2018, «Evidence for positive selection on recent human transposable element insertions», *Gene*, vol. 675, doi:10.1016/j.gene.2018.06.077, p. 69–79, ISSN 1879-0038. 12, 56

RIUS, N., Y. GUILLÉN, A. DELPRAT, A. KAPUSTA, C. FESCHOTTE et A. RUIZ. 2016, «Exploration of the Drosophila buzzatii transposable element content suggests underestimation of repeats in Drosophila genomes», *BMC Genomics*, vol. 17, doi:10.1186/s12864-016-2648-8, ISSN 1471-2164. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4862133/. 14, 64

ROLLINS, L. A., M. F. RICHARDSON et R. SHINE. 2015, «A genetic perspective on rapid evolution in cane toads (Rhinella marina)», *Molecular Ecology*, vol. 24, 9, doi:10.1111/mec.13184, p. 2264–2276, ISSN 1365-294X. 56

ROTA-STABELLI, O., L. OMETTO, G. TAIT, S. GHIROTTO, R. KAUR, F. DRAGO, J. GONZÁLEZ, V. M. WALTON, G. ANFORA et M. V. ROSSI-STACCONI. 2020, «Distinct genotypes and phenotypes in European and American strains of Drosophila suzukii: implications for biology and management of an invasive organism», *Journal of Pest Science*, vol. 93, 1, doi:10.1007/s10340-019-01172-y, p. 77–89, ISSN 1612-4766. URL https://doi.org/10.1007/s10340-019-01172-y. 21, 89, 90

ROUX, J. J. L., G. K. BROWN, M. BYRNE, J. NDLOVU, D. M. RICHARDSON, G. D. THOMPSON et J. R. U. WILSON. 2011, «Phylogeographic consequences of different introduction histories of invasive Australian Acacia species and Paraserianthes lophantha (Fabaceae) in South Africa», *Diversity and Distributions*, vol. 17, 5, doi:10.1111/j.1472-4642.2011.00784.x, p. 861–871, ISSN 1472-4642. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1472-4642.2011.00784.x. 68

ROY, M., B. VIGINIER, SAINT-MICHEL, F. ARNAUD, M. RATINIER et M. FABLET. 2020, «Viral infection impacts transposable element transcript amounts in Drosophila», *Proceedings of the National Academy of Sciences*, vol. 117, 22, doi:10.1073/pnas.2006106117, p. 12 249–12 257, ISSN 0027-8424, 1091-6490. URL https://www.pnas.org/content/117/22/12249. 8, 66, 120

RYAN, C. P., J. C. BROWNLIE et S. WHYARD. 2016, «Hsp90 and Physiological Stress Are Linked to Autonomous Transposon Mobility and Heritable Genetic Change in Nematodes», *Genome Biology and Evolution*, vol. 8, 12, doi:10.1093/gbe/evw284, p. 3794–3805, ISSN 1759-6653. 66

SAIKA, H., A. MORI, M. ENDO et S. TOKI. 2019, «Targeted deletion of rice retrotransposon Tos17 via CRISPR/Cas9», *Plant Cell Reports*, vol. 38, 4, doi:10.1007/s00299-018-2357-7, p. 455–458, ISSN 1432-203X. URL https://doi.org/10.1007/s00299-018-2357-7. 118

SCHLENKE, T. A. et D. J. BEGUN. 2004, «Strong selective sweep associated with a transposon insertion in Drosophila simulans», *Proceedings of the National Academy of Sciences*, vol. 101, 6, doi:10.1073/pnas.0303793101, p. 1626–1631, ISSN 0027-8424, 1091-6490. URL https://www.pnas.org/content/101/6/1626. 10, 11

SCHLÖTTERER, C., R. TOBLER, R. KOFLER et V. NOLTE. 2014, «Sequencing pools of individuals - mining genome-wide polymorphism data without big funding», *Nature Reviews. Genetics*, vol. 15, 11, doi:10.1038/nrg3803, p. 749–763, ISSN 1471-0064. 15

SCHMIDT, P. S., C.-T. ZHU, J. DAS, M. BATAVIA, L. YANG et W. F. EANES. 2008, «An amino acid polymorphism in the couch potato gene forms the basis for climatic adaptation in Drosophila melanogaster», *Proceedings of the National Academy of Sciences*, vol. 105, 42, doi:10.1073/pnas.0805485105, p. 16 207–16 211, ISSN 0027-8424, 1091-6490. URL https://www.pnas.org/content/105/42/16207. 25

SCHNABLE, P. S., D. WARE, R. S. FULTON, J. C. STEIN, F. WEI, S. PASTERNAK, C. LIANG, J. ZHANG, L. FULTON, T. A. GRAVES, P. MINX, A. D. REILY, L. COURTNEY, S. S. KRUCHOWSKI, C. TOMLINSON, C. STRONG, K. DELEHAUNTY, C. FRONICK, B. COURTNEY, S. M. ROCK, E. BELTER, F. DU, K. KIM, R. M. ABBOTT, M. COTTON, A. LEVY, P. MARCHETTO, K. OCHOA, S. M. JACKSON, B. GILLAM, W. CHEN, L. YAN, J. HIGGINBOTHAM, M. CARDENAS, J. WALIGORSKI, E. APPLEBAUM, L. PHELPS, J. FALCONE, K. KANCHI, T. THANE, A. SCIMONE, N. THANE, J. HENKE, T. WANG, J. RUPPERT, N. SHAH, K. ROTTER, J. HODGES, E. INGENTHRON, M. CORDES, S. KOHLBERG, J. SGRO, B. DELGADO, K. MEAD, A. CHINWALLA, S. LEONARD, K. CROUSE, K. COLLURA, D. KUDRNA, J. CURRIE, R. HE, A. ANGELOVA, S. RAJASEKAR, T. MUELLER, R. LOMELI, G. SCARA, A. KO, K. DELANEY, M. WISSOTSKI, G. LOPEZ, D. CAMPOS, M. BRAIDOTTI, E. ASHLEY, W. GOLSER, H. KIM, S. LEE, J. LIN, Z. DUJMIC, W. KIM, J. TALAG, A. ZUCCOLO, C. FAN, A. SEBASTIAN, M. KRAMER, L. SPIEGEL, L. NASCIMENTO, T. ZUTAVERN, B. MILLER, C. AMBROISE, S. MULLER, W. SPOONER, A. NARECHANIA, L. REN, S. WEI, S. KUMARI, B. FAGA, M. J. LEVY, L. MCMAHAN, P. VAN BUREN, M. W. VAUGHN, K. YING, C.-T. YEH, S. J. EMRICH, Y. JIA, A. KALYANARAMAN, A.-P. HSIA, W. B. BARBAZUK, R. S. BAUCOM, T. P. BRUTNELL, N. C. CARPITA, C. CHAPARRO, J.-M. CHIA, J.-M. DERAGON, J. C. ESTILL, Y. FU, J. A. JEDDELOH, Y. HAN, H. LEE, P. LI, D. R. LISCH, S. LIU, Z. LIU, D. H. NAGEL, M. C. MCCANN, P. SANMIGUEL, A. M. MYERS, D. NETTLETON, J. NGUYEN, B. W. PENNING, L. PONNALA, K. L. SCHNEIDER, D. C. SCHWARTZ, A. SHARMA, C. SODERLUND, N. M. SPRINGER, Q. SUN, H. WANG, M. WATERMAN, R. WESTERMAN, T. K. WOLFGRUBER, L. YANG, Y. YU, L. ZHANG, S. ZHOU, Q. ZHU, J. L. BENNETZEN, R. K. DAWE, J. JIANG, N. JIANG, G. G. PRESTING, S. R. WESSLER, S. ALURU, R. A. MARTIENSSEN, S. W. CLIFTON, W. R. MCCOMBIE, R. A. WING et R. K. WILSON. 2009, «The B73 maize genome: complexity, diversity, and dynamics», *Science (New York, N.Y.)*, vol. 326,

5956, doi:10.1126/science.1178534, p. 1112–1115, ISSN 1095-9203. 7, 55

SEBERG, O. et G. PETERSEN. 2009, «A unified classification system for eukaryotic transposable elements should reflect their phylogeny», *Nature Reviews. Genetics*, vol. 10, 4, doi:10.1038/nrg2165-c3, p. 276, ISSN 1471-0064. 4

SESSEGOLO, C., N. BURLET et A. HAUDRY. 2016, «Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies», *Biology Letters*, vol. 12, 8, doi:10.1098/rsbl.2016.0407, p. 20160 407. URL https://royalsocietypublishing.org/doi/full/10.1098/rsbl.2016.0407. 7, 22, 26, 56, 64, 86, 87, 88, 89, 91, 100, 113

SHEARER, P. W., J. D. WEST, V. M. WALTON, P. H. BROWN, N. SVETEC et J. C. CHIU. 2016, «Seasonal cues induce phenotypic plasticity of Drosophila suzukii to enhance winter survival», *BMC Ecology*, vol. 16, 1, doi:10.1186/s12898-016-0070-3, p. 11, ISSN 1472-6785. URL https://doi.org/10.1186/s12898-016-0070-3. 21, 25

SLOTKIN, R. K. et R. MARTIENSSEN. 2007, «Transposable elements and the epigenetic regulation of the genome», *Nature Reviews Genetics*, vol. 8, 4, doi:10.1038/nrg2072, p. 272–285, ISSN 1471-0064. URL https://www.nature.com/articles/nrg2072. 6

SPRADLING, A. C., D. STERN, A. BEATON, E. J. RHEM, T. LAVERTY, N. MOZDEN, S. MISRA et G. M. RUBIN. 1999, «The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes», *Genetics*, vol. 153, 1, p. 135–177, ISSN 0016-6731. 6

STANKE, M., M. DIEKHANS, R. BAERTSCH et D. HAUSSLER. 2008, «Using native and syntenically mapped cDNA alignments to improve de novo gene finding», *Bioinformatics (Oxford, England)*, vol. 24, 5, doi:10.1093/bioinformatics/btn013, p. 637–644, ISSN 1367-4811. 70, 92, 101

STAPLEY, J., A. W. SANTURE et S. R. DENNIS. 2015, «Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species», *Molecular Ecology*, vol. 24, 9, doi:10.1111/mec.13089, p. 2241–2252, ISSN 1365-294X. URL http://onlinelibrary.wiley.com/doi/abs/10.1111/mec.13089. 56, 68

STEPHAN, W. et H. LI. 2007, «The recent demographic and adaptive history of Drosophila melanogaster», *Heredity*, vol. 98, 2, doi:10.1038/sj.hdy.6800901, p. 65–68, ISSN 1365-2540. URL https://www.nature.com/articles/6800901. 68, 116

STOCKTON, D., A. WALLINGFORD, D. RENDON, P. FANNING, C. K. GREEN, L. DIEPENBROCK, E. BALLMAN, V. M. WALTON, R. ISAACS, H. LEACH, A. A. SIAL, F. DRUMMOND, H. BURRACK et G. M. LOEB. 2019, «Interactions Between Biotic and Abiotic Factors Affect Survival in Overwintering Drosophila suzukii (Diptera: Drosophilidae)», *Environmental Entomology*, vol. 48, 2, doi:10.1093/ee/nvy192, p. 454–464, ISSN 0046-225X. URL https://academic.oup.com/ee/article/48/2/454/5292620. 22

STOCKTON, D. G., A. K. WALLINGFORD et G. M. LOEB. 2018, «Phenotypic Plasticity Promotes Overwintering Survival in A Globally Invasive Crop Pest, Drosophila suzukii», *Insects*, vol. 9, 3, doi:10.3390/insects9030105, ISSN 2075-4450. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6164111/. 25

STRITT, C., S. P. GORDON, T. WICKER, J. P. VOGEL et A. C. ROULIN. 2017, «Recent Activity in Expanding Populations and Purifying Selection Have Shaped Transposable Element Landscapes across Natural Accessions of

the Mediterranean Grass Brachypodium distachyon», *Genome Biology and Evolution*, vol. 10, 1, doi:10.1093/gbe/evx276, p. 304–318, ISSN 1759-6653. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5786231/. 114

SULTANA, T., A. ZAMBORLINI, G. CRISTOFARI et P. LESAGE. 2017, «Integration site selection by retroviruses and transposable elements in eukaryotes», *Nature Reviews Genetics*, vol. 18, 5, doi:10.1038/nrg.2017.7, p. 292–308, ISSN 1471-0064. URL https://www.nature.com/articles/nrg.2017.7. 10, 88

SUN, C., D. B. SHEPARD, R. A. CHONG, J. LÓPEZ ARRIAZA, K. HALL, T. A. CASTOE, C. FESCHOTTE, D. D. POLLOCK et R. L. MUELLER. 2012, «LTR Retrotransposons Contribute to Genomic Gigantism in Plethodontid Salamanders», *Genome Biology and Evolution*, vol. 4, 2, doi:10.1093/gbe/evr139, p. 168–183, ISSN 1759-6653. URL https://doi.org/10.1093/gbe/evr139. 88

SUN, H., J. DING, M. PIEDNOËL et K. SCHNEEBERGER. 2018, «findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies», *Bioinformatics (Oxford, England)*, vol. 34, 4, doi:10.1093/bioinformatics/btx637, p. 550–557, ISSN 1367-4811. 100

TALAVERA, G. et J. CASTRESANA. 2007, «Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments», *Systematic Biology*, vol. 56, 4, doi:10.1080/10635150701472164, p. 564–577, ISSN 1063-5157. URL https://academic.oup.com/sysbio/article/56/4/564/1682121. 99

TALLA, V., A. SUH, F. KALSOOM, V. DINCĂ, R. VILA, M. FRIBERG, C. WIKLUND et N. BACKSTRÖM. 2017, «Rapid Increase in Genome Size as a Consequence of Transposable Element Hyperactivity in Wood-White (Leptidea) Butterflies», *Genome Biology and Evolution*, vol. 9, 10, doi:10.1093/gbe/evx163, p. 2491–2505. URL https://academic.oup.com/gbe/article/9/10/2491/4091610. 7, 56, 66, 87, 119

TAMURA, K., S. SUBRAMANIAN et S. KUMAR. 2004, «Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks», *Molecular Biology and Evolution*, vol. 21, 1, doi:10.1093/molbev/msg236, p. 36–44, ISSN 0737-4038. 99, 101

THABET, S., A. NAMOUCHI et H. MARDASSI. 2015, «Evolutionary Trends of the Transposase-Encoding Open Reading Frames A and B (orfA and orfB) of the Mycobacterial IS6110 Insertion Sequence», *PLoS ONE*, vol. 10, 6, doi:10.1371/journal.pone.0130161, ISSN 1932-6203. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4473070/. 121

TOCHEN, S., D. T. DALTON, N. WIMAN, C. HAMM, P. W. SHEARER et V. M. WALTON. 2014, «Temperature-related development and population parameters for Drosophila suzukii (Diptera: Drosophilidae) on cherry and blueberry», *Environmental Entomology*, vol. 43, 2, doi:10.1603/EN13200, p. 501–510, ISSN 1938-2936. 21

TOXOPEUS, J., R. JAKOBS, L. V. FERGUSON, T. D. GARIEPY et B. J. SINCLAIR. 2016, «Reproductive arrest and stress resistance in winter-acclimated Drosophila suzukii», *Journal of Insect Physiology*, vol. 89, doi:10.1016/j.jinsphys.2016.03.006, p. 37–51, ISSN 1879-1611. 22

VACAS, S., J. PRIMO, J. J. MANCLÚS, MONTOYA et V. NAVARRO-LLOPIS. 2019, «Survey on Drosophila suzukii Natural Short-Term Dispersal Capacities Using the Mark-Release-Recapture Technique», *Insects*, vol. 10, 9, doi:10.3390/insects10090268, ISSN 2075-4450. 22

VAN'T HOF, A. E., P. CAMPAGNE, D. J. RIGDEN, C. J. YUNG, J. LINGLEY, M. A. QUAIL, N. HALL, A. C. DARBY et I. J. SACCHERI. 2016, «The industrial melanism mutation in British peppered moths is a transposable element», *Nature*, vol. 534, 7605, doi:10.1038/nature17951, p. 102–105, ISSN 1476-4687. 11, 56, 64

VARÓN-GONZÁLEZ, C., A. FRAIMOUT, A. DELAPRÉ, V. DEBAT et R. CORNETTE. 2020, «Limited thermal plasticity and geographical divergence in the ovipositor of Drosophila suzukii», *Royal Society Open Science*, vol. 7, 1, doi:10.1098/rsos.191577, p. 191 577, ISSN 2054-5703. 25

VARÓN-GONZÁLEZ, C., A. FRAIMOUT et V. DEBAT. 2020, «Drosophila suzukii wing spot size is robust to developmental temperature», *Ecology and Evolution*, vol. 10, 7, doi:https://doi.org/10.1002/ece3.5902, p. 3178–3188, ISSN 2045-7758. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.5902. 25

VENDRELL-MIR, P., F. BARTERI, M. MERENCIANO, J. GONZÁLEZ, J. M. CASACUBERTA et R. CASTANERA. 2019, «A benchmark of transposon insertion detection tools using real data», *Mobile DNA*, vol. 10, 1, doi:10.1186/s13100-019-0197-9, p. 53, ISSN 1759-8753. URL https://doi.org/10.1186/s13100-019-0197-9. 17, 18, 66

VICARIO, S., E. N. MORIYAMA et J. R. POWELL. 2007, «Codon usage in twelve species of Drosophila», *BMC Evolutionary Biology*, vol. 7, doi:10.1186/1471-2148-7-226, p. 226, ISSN 1471-2148. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2213667/. 97, 113

VIEIRA, C., D. LEPETIT, S. DUMONT et C. BIÉMONT. 1999, «Wake up of transposable elements following Drosophila simulans worldwide colonization», *Molecular Biology and Evolution*, vol. 16, 9, p. 1251–1255, ISSN 0737-4038. 56

VILLANUEVA-CAÑAS, J. L., G. E. RECH, M. A. R. DE CARA et J. GONZÁLEZ. 2017, «Beyond SNPs: how to detect selection on transposable element insertions», *Methods in Ecology and Evolution*, vol. 8, 6, doi:10.1111/2041-210X.12781, p. 728–737, ISSN 2041-210X. URL https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.12781. 10, 67

VITTE, C. et O. PANAUD. 2005, «LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model», *Cytogenetic and Genome Research*, vol. 110, 1-4, doi:10.1159/000084941, p. 91–107, ISSN 1424-8581, 1424-859X. URL https://www.karger.com/Article/FullText/84941, publisher: Karger Publishers. 87

WALSER, J.-C., B. CHEN et M. E. FEDER. 2006, «Heat-Shock Promoters: Targets for Evolution by P Transposable Elements in Drosophila», *PLOS Genetics*, vol. 2, 10, doi:10.1371/journal.pgen.0020165, p. e165, ISSN 1553-7404. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0020165, publisher: Public Library of Science. 119

WALSH, D. B., M. P. BOLDA, R. E. GOODHUE, A. J. DREVES, J. LEE, D. J. BRUCK, V. M. WALTON, S. D. O'NEAL et F. G. ZALOM. 2011, «Drosophila suzukii (Diptera: Drosophilidae): Invasive Pest of Ripening Soft Fruit Expanding its Geographic Range and Damage Potential», *Journal of Integrated Pest Management*, vol. 2, 1, doi:10.1603/IPM10010, p. G1–G7. URL https://academic.oup.com/jipm/article/2/1/G1/2193867. 22

WATERHOUSE, R. M., M. SEPPEY, F. A. SIMÃO, M. MANNI, P. IOANNIDIS, G. KLIOUTCHNIKOV, E. V. KRIVENTSEVA et E. M. ZDOBNOV.

2018, «BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics», *Molecular Biology and Evolution*, vol. 35, 3, doi:10.1093/molbev/msx319, p. 543–548, ISSN 1537-1719. 99

WEBER, C. C., B. NABHOLZ, J. ROMIGUIER et H. ELLEGREN. 2014, «Kr/Kc but not dN/dS correlates positively with body mass in birds, raising implications for inferring lineage-specific selection», *Genome Biology*, vol. 15, 12, doi:10.1186/s13059-014-0542-8, ISSN 1465-6906. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4264323/. 97, 113

WEILGUNY, L. et R. KOFLER. 2019, «DeviaTE: Assembly-free analysis and visualization of mobile genetic element composition», *Molecular Ecology Resources*, vol. 19, 5, doi:10.1111/1755-0998.13030, p. 1346–1354, ISSN 1755-098X. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6791034/. 3, 4, 19, 90, 91, 101, 123, 159

WHITNEY, K. D., T. GARLAND et JR. 2010, «Did Genetic Drift Drive Increases in Genome Complexity?», *PLoS Genetics*, vol. 6, 8, doi:10.1371/journal.pgen.1001080. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2928810/. 7, 88

WICKER, T., H. GUNDLACH, M. SPANNAGL, C. UAUY, P. BORRILL, R. H. RAMÍREZ-GONZÁLEZ, R. DE OLIVEIRA, K. F. X. MAYER, E. PAUX, F. CHOULET et INTERNATIONAL WHEAT GENOME SEQUENCING CONSORTIUM. 2018, «Impact of transposable elements on genome structure and evolution in bread wheat», *Genome Biology*, vol. 19, 1, doi:10.1186/s13059-018-1479-0, p. 103, ISSN 1474-760X. URL https://doi.org/10.1186/s13059-018-1479-0. 87

WICKER, T., F. SABOT, A. HUA-VAN, J. L. BENNETZEN, P. CAPY, B. CHALHOUB, A. FLAVELL, P. LEROY, M. MORGANTE, O. PANAUD et OTHERS. 2007, «A unified classification system for eukaryotic transposable elements», *Nature Reviews Genetics*, vol. 8, 12, p. 973–982. URL http://www.nature.com/nrg/journal/v8/n12/abs/nrg2165.html. 4, 87

WICKER, T., F. SABOT, A. HUA-VAN, J. L. BENNETZEN, P. CAPY, B. CHALHOUB, A. FLAVELL, P. LEROY, M. MORGANTE, O. PANAUD, E. PAUX, P. SANMIGUEL et A. H. SCHULMAN. 2009, «Reply: A unified classification system for eukaryotic transposable elements should reflect their phylogeny», *Nature Reviews Genetics*, vol. 10, 4, doi:10.1038/nrg2165-c4, p. 276–276, ISSN 1471-0056, 1471-0064. URL http://www.nature.com/articles/nrg2165-c4. 4

WINGETT, S. W. et S. ANDREWS. 2018, «FastQ Screen: A tool for multi-genome mapping and quality control», *F1000Research*, vol. 7, doi:10.12688/f1000research.15931.2, p. 1338, ISSN 2046-1402. URL https://f1000research.com/articles/7-1338/v2. 100

WRIGHT, S. I., N. AGRAWAL et T. E. BUREAU. 2003, «Effects of Recombination Rate and Gene Density on Transposable Element Distributions in Arabidopsis thaliana», *Genome Research*, vol. 13, 8, doi:10.1101/gr.1281503, p. 1897–1903, ISSN 1088-9051, 1549-5469. URL http://genome.cshlp.org/content/13/8/1897. 12, 13, 65, 115, 118

YANG, Z. 2007, «PAML 4: phylogenetic analysis by maximum likelihood», *Molecular Biology and Evolution*, vol. 24, 8, doi:10.1093/molbev/msm088, p. 1586–1591, ISSN 0737-4038. 94, 102, 109

ZANNI, V., A. EYMERY, M. COIFFET, M. ZYTNICKI, I. LUYTEN, H. QUESNEVILLE, C. VAURY et S. JENSEN. 2013, «Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters», *Proceedings of the National*

Academy of Sciences of the United States of America, vol. 110, 49, doi:10.1073/pnas.1313677110, p. 19 842–19 847, ISSN 1091-6490. 65, 98

ZHANG, Y.-Y., D.-Y. ZHANG et S. BARRETT. 2010, «Genetic uniformity characterizes the invasive spread of water hyacinth (Eichhornia crassipes), a clonal aquatic plant», *Molecular ecology*, vol. 19, doi:10.1111/j.1365-294X.2010.04609.x, p. 1774–86. 68

ZHUANG, J., J. WANG, W. THEURKAUF et Z. WENG. 2014, «TEMP: a computational method for analyzing transposable element polymorphism in populations», *Nucleic Acids Research*, vol. 42, 11, doi:10.1093/nar/gku323, p. 6826–6838, ISSN 0305-1048. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4066757/. 18

PhD THESIS SUMMARY

Transposable Elements (TEs) are selfish genomic elements that multiply by copying and pasting themselves at various locations in the host genome. This particular property allows them to maintain in populations, and even to proliferate, without necessarily bringing any advantage to their host. With the exception of a few examples of presumed adaptive copies, it has even been shown that these sequences have a negative, or neutral impact. So far, their dynamics remains unclear. The invasive species *Drosophila suzukii* offers a unique opportunity to study TE dynamics. Originally from Asia, it began an invasion of the American and European continents in 2008, allowing to test the effects of environmental and demographic changes on TE dynamics. During this PhD, a portrait of TE content in *D. suzukii* was first drawn in a reference genome, revealing the presence of 47% ET, mainly concentrated in gene-poor regions. TE abundance was then quantified in 22 populations of *D. suzukii* but also in related species. Comparison of TE contents between invasive and native populations indicates an accumulation of these sequences during the invasive process. This accumulation is associated with a decrease of the effective size proxy $\widehat{\theta_W}$, suggesting a role of a relaxation of purifying selection against TEs in their proliferation. At the species level, the comparison of TE contents between *D. suzukii* and related species shows an accumulation of TEs in *D. suzukii* lineage over the last four million years. This accumulation, concentrated in gene-poor regions, is likely responsible for the high percentage of TEs in *D. suzukii* genome. The study of dN/dS suggests that this increase in TE content is not associated with a relaxation of purifying selection. Beyond the potential role of variations in selection intensity, the impact of environmental changes and genetic variability on TE dynamics was studied in *D. suzukii* populations. Results indicate a complex role of genetic variability, with ET abundance associated with ~5,000 genomic regions, but no impact of bioclimatic variables on TE activity. However, the study of insertion polymorphism in the populations shows that environmental selection pressures could increase the frequency of some TE copies, including six insertions potentially involved in the adaptation of invasive populations.