Analyse bioinformatique des événements de transferts horizontaux entre espèces de drosophiles et lien avec la régulation des éléments transposables

Laurent Modolo

N° 258-2014 Année 2014

Thèse de l'université de lyon

Présentée

devant L'UNIVERSITÉ CLAUDE BERNARD LYON 1

pour l'obtention

du Diplôme de doctorat
(arrêté du 7 août 2006)

soutenue publiquement le
1er décembre 2014

par

Laurent Modolo

Analyse bioinformatique des événements de transferts horizontaux entre espèces de drosophiles et lien avec la régulation des éléments transposables

Directrice de thèse : Emmanuelle Lerat

Jury :	Sandrine Charles	Examinateur
	Clément Gilbert	Examinateur
	Emmanuelle Lerat	Directrice de thèse
	Catherine Matias	Rapporteur
	Oliver Panaud	Rapporteur
	Arnaud le Rouzic	Examinateur

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. François-Noël GILLY**
Vice-président du Conseil d'Administration	M. le Professeur Hamda BEN HADID
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur Philippe LALLE
Vice-président du Conseil Scientifique	M. le Professeur Germain GILLET
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme. la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. le Professeur F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. FLEURY
Département Chimie Biochimie	Directeur : Mme le Professeur C. FELIX
Département GEP	Directeur : M. H. HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur G. TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : Mme J.C. PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y. VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. P. FOURNIER
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. C. VITON
Institut Universitaire de Formation des Maîtres	Directeur : M. A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Administrateur provisoire : M. N. LEBOISNE

Résumé

Les éléments transposable (ET) sont des séquences d'ADN qui ont la capacité de se déplacer (transposer) au sein des génomes. Ces éléments ont été détectés dans la majorité des espèces dont le génome a été séquencé. Leur activité de transposition a différentes conséquences sur le génome de l'hôte et peut représenter une importante source de variabilité génétique. En particulier, chaque nouvelle insertion d'ET va correspondre à une nouvelle mutation de la séquence d'ADN du génome hôte et va de plus contribuer à une augmentation de la taille de celui-ci. Comme pour les autres types de mutations, la majorité de ces insertions sont sans effet (neutres) ou délétères. De plus les séquences d'ET forment des répétitions, ce qui augmente la probabilité de recombinaison entre différentes portions non-homologues du génome de l'hôte. Pour contrebalancer les effets négatifs liés à l'activité des ET, il existe chez leurs hôtes des mécanismes régulant l'activité de transposition. Une fois qu'un ET est régulé, l'accumulation progressive de mutations dans sa séquence conduit fatalement à la perte définitive de son activité de transposition.

Pour comprendre le succès et le maintien de ces séquences répétées aux propriétés particulières, de nombreuses études sur les ET ont été menées afin de comprendre leur dynamique. Les travaux que j'ai effectués au cours de cette thèse s'inscrivent dans cette démarche, avec d'une part l'étude des transferts horizontaux (TH) d'ET, un moyen d'échapper aux mécanismes de régulation de leur transposition, et d'autre part l'étude de leur régulation par les mécanismes de l'ARN interférence par des approches de bioinformatique et de statistique. Je me suis particulièrement intéressé au modèle biologique de la drosophile, pour lequel de nombreuses données génomiques sont disponibles.

Dans la première partie de ma thèse, je me suis intéressé à l'étude des TH entre deux espèces proches de drosophiles, *Drosophila melanogaster* et *D. simulans*. J'ai développé une nouvelle méthode bioinformatique permettant la détection de séquences transférées horizontalement par l'analyse comparative de deux génomes eucaryotes. Cette étude m'a permis de mettre en évidence l'importance des TH d'ET entre ces deux espèces de drosophiles. Mes résultats étayent un modèle de dynamique des ET dans lequel ces séquences échappent aux mécanismes de régulation de l'hôte par TH vers un nouvel hôte qui ne possède pas encore ce contrôle. Afin d'obtenir la sensibilité nécessaire à la détection de tous les événements de TH entre deux espèces proches, ce travail m'a aussi conduit à m'intéresser à la problématique des tests multiples unilatéraux. J'ai ainsi développé une nouvelle méthode de contrôle du taux de faux positifs moyen (*FDR*), permettant d'apporter la puissance des procédures utilisant un calcul du *FDR* local, pour la correction des tests multiples unilatéraux. Cette procédure peut être appliquée à d'autres problématiques que celle de la détection de TH.

Dans la deuxième partie de ma thèse, j'ai étudié la régulation des ET par la voie des petits ARN, un mécanisme de l'ARN interférence. Dans cette étude, j'ai analysé des données de séquençage de petits ARN, ainsi que d'ARN totaux issues de différentes populations de *D. simulans*. Ces populations ont été sélectionnées pour leur polymorphisme en terme de présence et nombre d'insertions d'ET, dans le but de détecter des différences au niveau des mécanismes de régulation par les petits ARN. Ce travail a conduit au développement d'un pipeline d'analyse permettant d'étudier des différences d'expression entre des séquences répétées. Certaines caractéristiques du jeu de données des petits ARN utilisé ont nécessité le développement d'une nouvelle procédure pour pouvoir les traiter. Cette méthode a été généralisée et implémentée dans le programme UrQt pour pouvoir être utilisée dans le contrôle qualité de données de séquences de nouvelle génération.

Mots-clés : évolution moléculaire, éléments transposables, drosophiles, transferts horizontaux, regulation épigénétique.

Abstract

Transposable elements (TEs) are repeated DNA sequences that are able to move (transpose) within their host genome. These elements have been found in the majority of species whose genomes have been sequenced. TE transposition has various consequences on the host genome and can be an important source of genetic variability. Each new TE insertion corresponds to a new mutation of the host DNA and contributes to the increase of its size. As any other mutations, a majority of these insertions do not have any effect (neutral) or are deleterious. Moreover, TE sequences form repeats, which increase the probability of recombination between non-homologous parts of the host genome. To counteract the negative effects of their TEs, regulation mechanisms of the TE transposition are present in the host genome. Once a TE is regulated, the progressive accumulation of mutations in its sequence will inevitably lead to the definitive loss of its transposition capacity.

Many efforts have been made to understand the success and the maintaining of these peculiar repeated sequences. My work during this thesis is part of this effort, with the study of horizontal transfers (HTs) of TEs enabling them to escape host regulation mechanisms, and the study of this regulation by RNA interference using bioinformatic and statistical approaches. My work focused on the drosophila biological model for which numerous genomic data are available.

The first part of my thesis concerns the study of HTs between two closely related species, *Drosophila melanogaster* and *D. simulans*. I have developed a new bioinformatic method for the detection of HTs between two eukaryotic genomes. This work pointed out the importance of the HT of TEs between these two drosophila species. My results support the model of TE dynamic where these sequences escape the host regulation mechanisms by a HT toward a new host devoid of this control. In order to obtain the necessary sensitivity for the detection of all HT events between two closely related species, the development of this method brought me to work on the unilateral multiple testing problematic. I have developed a new procedure to control the expected false discovery rate (*FDR*) that brought the power of local *FDR* based procedures to unilateral multiple testing. This work can be applied to other topics than HT detection.

The second part of my thesis focuses on the regulation of TEs by the small RNA pathway, an RNA interference mechanism. For this study, I have analyzed sequencing data of small RNAs and total RNAs from different *D. simulans* populations. These populations were selected for their polymorphism in presence and copy number of TEs to study differences in the small RNA regulation mechanisms. For this work, I have developed an analysis pipeline, to study differences of expression between repeated sequences. Some features of the small RNA dataset required the development of a new procedure to parse them. This procedure was extended and implemented in the software UrQt, to be used for the quality control of next generation sequencing data.

Keywords: molecular evolution, transposable elements, drosophila, horizontal transfers, epigenetic regulation.

Remerciements

Je tiens tout d'abord à remercier les personnes qui ont contribué directement à la réalisation de cette thèse :

Emmanuelle Lerat, pour m'avoir encadré tout au long de ma thèse. Merci, d'avoir été disponible, patiente et toujours de bon conseil. Merci surtout pour ces nombreuses heures à relire mes travaux.

Franck Picard, sans qui ma thèse n'aurait pas pris la tournure qu'elle a prise. Merci, pour ton dynamisme et ta bonne humeur ainsi que de m'avoir fait découvrir les joies des statistiques. Merci aussi pour toutes les opportunités que tu m'a données au cours de cette thèse.

Clément Goubert, qui a été un fantastique co-bureau pendant plus de deux années. Merci pour ta gentillesse et ta bonne humeur, et pour tous ces bons moments.

Valèria Romero, qui n'est pas restée en France assez longtemps. Merci, pour tes réflections et tous ces débats passionnés.

Marie Fablet, pour ses idées et ses conseils. Merci d'avoir été toujours disponible pour partager tes connaissances et discuter.

Alain Celisse, pour sa patience et sa rigueur. Merci d'avoir accepté de m'encadrer pendant près d'une semaine à Lille.

Stéphane Robin, pour ses idées et ses conseils. Merci d'avoir été plus qu'un membre de mon comité de pilotage.

Ensuite, je tiens à remercier ma famille, sans laquelle je ne serais pas l'homme que je suis devenu.

Merci à mes parents Anne et Ignace pour m'avoir toujours soutenu et cru en moi.

Merci à mes trois soeurs Inès, Juliette et Adèle pour être elles-mêmes, ne changez rien.

Merci aussi à mes grands parents Bernadette, Daniela et René pour avoir toujours porté un grand intérêt à ce que je fais.

Merci à Eugénie pour deux années de bonheur et pour m'avoir soutenu dans les moments de stress. Merci beaucoup pour tous les efforts que tu as fait pendant la fin de ma thèse.

Merci à mes amis, Andrea, Benjamin, Camille, Claire, Fabien, Laura, Louis-Marie, Matthieu, Maximilien, Myriam, Vincent pour avoir été là pendant toutes ces années.

Merci à tous les autres membres de l'équipe éléments transposables Abdou, Annabelle, Cristina, Elias, Hélène, Matthieu et Virginie pour cette bonne ambiance pendant ces trois années. Merci aussi à tous les thésards et moins thésards du LBBE pour toutes les activités extra-thèse organisées pendant ces années.

Enfin merci à toutes les personnes que j'oublie et qui seront vexées de ne pas se retrouver sur cette page.

Cette thèse est dédiée à mes trois soeurs, Inès, Juliette et Adèle.

Table des matières

Introduction

Avant la découverte des éléments transposables (ET), notre vision du génome, qui est l'ensemble de l'information génétique contenue dans chaque cellule d'un individu, était très différente de notre vision actuelle. En effet, les observations de la molécule d'acide désoxyribonucléique (ADN) qui encode cette information génétique sous la forme d'une succession de quatre bases azotées (Adénine – A, Thymine – T, Guanine – G et Cytosine – C) avaient mené à l'hypothèse d'une molécule très stable dans le temps. Les ET ont été mis en évidence par Barbara McClintock dans les années 1950 au cours de travaux sur des variétés de maïs phénotypiquement instables qui ont été plus tard récompensés par un prix Nobel de médecine. Ces analyses ont montré qu'il existait un contrôle génétique de l'activité de transposition des ET et que, dans certains croisements chez les maïs, ce contrôle pouvait se relâcher et permettre la mobilisation de ces séquences. En effet, la principale caractéristique des ET est d'être mobile et de pouvoir se répliquer (ou transposer) d'un point à l'autre du génome. La découverte de ces séquences a donc changé fondamentalement notre vision de la dynamique des génomes en passant du paradigme d'un génome figé qui évolue lentement dans le temps, à celui d'un génome changeant dans lequel des remaniements importants des séquences génétiques peuvent brusquement avoir lieu.

Les ET possèdent une capacité à se multiplier dans le génome de leur hôte qui peut être très similaire aux modes d'infection utilisés par les virus à ADN ou les rétrovirus à ARN. L'ARN ou acide ribonucléique est une molécule similaire à l'ADN qui est formée (transcrite) à partir de celui-ci. Cette molécule est moins stable que celle d'ADN et possède des bases Uracile (U) qui remplacent les bases T. De plus, si la séquence

d'ADN des ET contient généralement toute l'information nécessaire pour le mécanisme de transposition, comme pour les virus, leur activité reste dépendante des ressources de l'hôte. Pour finir, la majorité de l'activité de ces éléments consiste à se multiplier sans "considération" pour les effets provoqués chez l'hôte. C'est pourquoi ces séquences ont été très tôt considérées comme des parasites génomiques. Ainsi, même si ces éléments constituent une part importante de la plupart des génomes eucaryotes et que la distinction entre l'ADN de l'hôte et celui des ET est parfois difficile à faire, nous opposerons tout au long de cette thèse ce qui concerne l'hôte de ce qui concerne les ET.

Tout d'abord décrits comme de l'ADN "poubelle" [1] ou "égoïste" [2, 3], les ET peuvent être considérés comme de parfaits exemples de parasites génomiques. Bien que ces éléments aient été découverts dans les années 1950, il a fallu attendre les années 1970 pour qu'une étude sur l'acquisition d'une résistance à un antibiotique causée par une insertion d'ET chez des bactéries renouvelle l'intérêt porté à ces séquences répétées [4]. La plupart des familles d'ET possède des séquences promotrices et des séquences codantes nécessaires à leur transposition, par conséquent leur présence en de multiple copies peut représenter une réserve importante de matériel génétique qui peut être recruté par l'hôte pour de nouvelles fonctions [5]. Ce phénomène est appelé exaptation. Il aura fallu moins d'un siècle depuis leur découverte pour que le statut des ET passe aux yeux de la communauté scientifique de parasite génomique à celui de véritable moteur évolutif [6, 7]. Les ET ont été retrouvés dans la plupart des grands groupes eucaryotes [8] et au delà puisqu'on les trouve aussi chez les bactéries et les archées, l'ensemble de ces deux super-règnes étant dénommé les procaryotes. Les procaryotes, qui sont caractérisés par des génomes de plus petite taille que ceux des eucaryotes, possèdent aussi un plus faible nombre de copies d'ET. Au cours de cette thèse nous nous sommes principalement intéressés à la dynamique des ET chez les eucaryotes. L'histoire des ET chez ces organismes peut être retracée jusqu'à l'ancêtre commun des eucaryotes [9]. Les eucaryotes semblent donc avoir co-évolué avec ces séquences jusqu'à nos jours [10] et les relations complexes qui existent entre les ET et le génome de leurs hôtes font l'objet de nombreuses études [7].

De part leurs mécanismes de transposition, que nous présenterons par la suite, chaque nouvelle insertion d'ET peut provoquer des dommages importants au niveau du génome de l'hôte. Pour expliquer le maintien de ces éléments malgré leurs nombreux effets délétères, il est nécessaire de comprendre les mécanismes sous-jacents à leur activité. Nous présenterons donc dans la première partie de cette introduction les différents types d'ET ainsi que leurs modes de transposition. Si certaines insertions d'ET peuvent se révéler adaptatives, comme pour certains autres types de mutations dans le génome, la ma-

jorité d'entre elles sont neutres (sans effet), ou délétères. Il est donc essentiel pour les génomes hôtes de posséder des mécanismes de contrôle de la transposition des ET, pour assurer leur propre stabilité. Ces mécanismes peuvent se décomposer en deux grandes familles, que nous introduirons par la suite, selon qu'il s'agisse de contrôle pré- ou post-transcriptionnel. Enfin, devant l'efficacité de ce contrôle il est important de comprendre les mécanismes ayant permis aux ET d'échapper à leur inactivation et de "survivre" jusqu'à aujourd'hui. Nous présenterons donc dans une dernière partie les modèles expliquant le maintien des ET dans les génomes et leurs mécanismes associés.

Bien que les travaux effectués au cours de ma thèse se concentrent sur un aspect bioinformatique et biostatistique de l'étude dela dynamique des ET, il est important de comprendre les mécanismes biologiques impliqués dans le maintien de ces séquences d'ADN. Ce premier chapitre se veut donc une introduction générale aux ET avec une revue de la littérature sur ces éléments chez les eucaryotes et plus particulièrement sur le modèle drosophile que j'ai utilisé tout au long de mon projet de recherche. La description des outils bioinformatiques spécifiques à l'étude de ces séquences répétées sera effectuée au chapitre suivant.

1 Les mécanismes de transposition des éléments transposables

L'histoire évolutive des ET est complexe et les relations qui existent entre les séquences de ces éléments peuvent être décrites selon différents critères. Par conséquent, il existe différentes classifications des ET, comme celle établie par Finnegan [11], ou celle établie par Wicker et al. [12]. Cette dernière, plus récente, est celle que nous allons présenter ici. En effet, si la proportion d'ET dans un génome est très variable suivant les espèces (85% du génome du maïs contre 15% de celui de *Drosophila melanogaster* [13, 14]) il existe aussi des différences importantes dans la composition en famille d'ET d'une espèce à une autre [15]. Ces différentes familles d'ET peuvent être réparties en deux grandes classes, selon que leur mécanisme de transposition repose sur un intermédiaire ARN, ce qui constitue les éléments de classe I ou rétrotransposons, ou sur un intermédiaire ADN pour les éléments de classe II ou transposons [12]. Les éléments de classe I, comme les éléments de classe II, se décomposent en différentes grandes familles en fonction des domaines protéiques contenus dans leurs séquences, de particularités structurales, ainsi que de leur cycle de transposition.

Nous présenterons dans cette première partie les différents types d'ET et leurs mé-

canismes de transposition.

1.1 Les éléments de classe I

Les éléments de la classe I, ou rétrotransposons, peuvent être décomposés en trois groupes. Nous distinguerons les rétrotransposons à LTR qui possèdent de longues répétitions terminales (LTR pour l'anglais "long terminal repeat") encadrant leurs séquences, des rétrotransposons sans LTR qui en sont dépourvus. Le troisième groupe est constitué d'ET non autonomes qui dépendent de la machinerie moléculaire encodée par d'autres éléments de la classe I pour transposer et qui n'ont pas non plus de LTR. Les éléments de la classe I partagent tous un mécanisme de transposition réplicatif au cours duquel une copie de l'élément va être transcrite en ARN puis rétro-transcrite en une nouvelle copie ADN. Cette nouvelle copie est ensuite insérée à un autre endroit du génome appelé site cible de transposition.

1.1.1 Les rétrotransposons à LTR

Les rétrotransposons à LTR regroupent 4 grandes familles, *Copia*, *Gypsy*, *Bel-Pao*, ainsi que les rétrovirus endogènes (*ERV*). Ces éléments ont une taille variant de 5 à 9 kilo-base (kb) et sont encadrés par des LTR pouvant atteindre 1 kb. Entre ces deux LTR, leur séquence contient deux à trois cadres ouverts de lecture (ORF pour l'anglais "open reading frame") qui correspondent respectivement aux gènes *gag*, *pol*, et optionnellement *env* (pour la famille *ERV*). La séquence *gag* (group-specific antigen) code pour une protéine de structure similaire à celle des rétrovirus. La séquence *pol* code pour une polyprotéine constituée de quatre domaines : une protéase (PR), une intégrase (INT), une transcriptase inverse (RT) et une RNaseH (RH) (voir Figure 1). La présence de ces quatre domaines est nécessaire au cycle de transposition des rétrotransposons à LTR décrit dans la Figure 1. Le mode d'intégration de ces éléments dans le génome est très similaire à celui des rétrovirus à ARN.

Ce cycle de transposition commence par une phase au cours de laquelle le rétrotransposon est transcrit en ARN. Cet ARN simple brin va ensuite servir de guide à la synthèse d'un ADN dans le cytoplasme par la transcriptase inverse. L'ARN primaire est dégradé par le domaine RNaseH pour laisser place à la synthèse d'un brin d'ADN complémentaire à celui synthétisé à partir de l'ARN. Enfin, l'ADN double brin formé va retourner dans le noyau, avant d'être inséré quelque part dans le génome grâce à l'action de l'intégrase. Les différents domaines d'intégrase ciblent des sites d'insertion spécifiques (TIS pour l'anglais "target insertion site") d'une longueur de 4 à 6 nucléotides en fonction des ET.

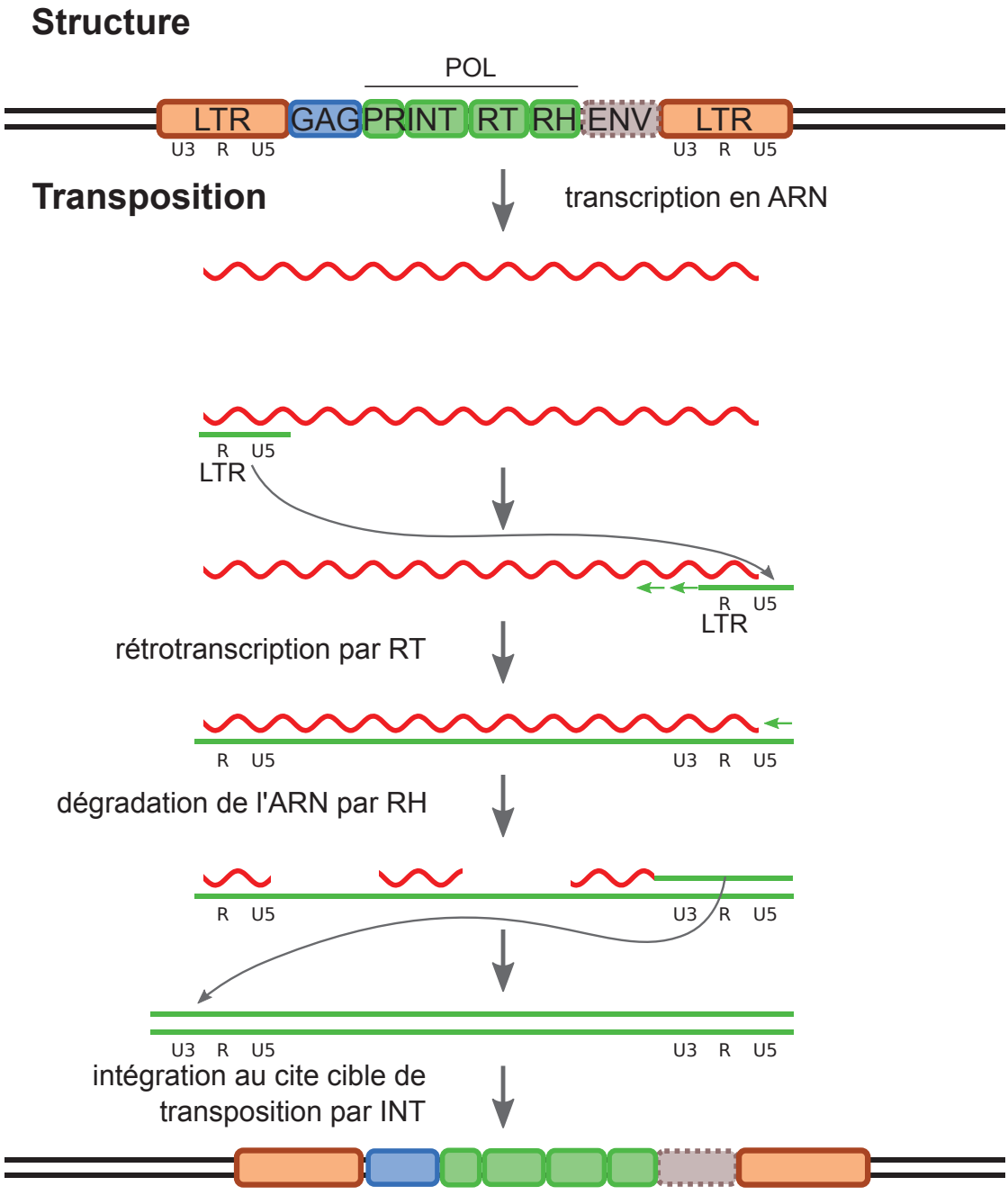


FIGURE 1 : Structure et cycle de transposition d'un rétrotransposon à LTR. L'ADN est représenté par une ligne droite, en noir pour l'ADN génomique et en vert pour celui de l'ET, alors que l'ARN est représenté par une ligne ondulée en rouge. Les différents domaines protéiques sont représentés par des rectangles de couleurs. *gag* , *pol* et *env* correspondent aux 3 ORF, avec PR pour protéase, INT pour intégrase, RT pour transcriptase inverse et RH pour RnaseH.

Ces TIS étant assez courts pour être retrouvés le long du génome de l'hôte, simplement par hasard, cela permet à ces éléments de transposer en de nombreuses positions.

Pour les éléments de la famille *ERV*, la présence du gène *env* qui code pour une protéine d'enveloppe très similaire à celle des rétrovirus, peut déterminer leur infectiosité et permettre leur transmission non sexuée [16, 17]. Néanmoins, malgré leurs fortes ressemblances avec les rétrovirus, les éléments *ERV* sont plus proches des autres rétrotransposons à LTR des rétrotransposons à LTR type *gypsy* ou *copia* [18].

Bien qu'ils ne possèdent pas de LTR, les éléments *DIRS* ("Dictyostelium intermediate repeat sequences") sont aussi classés avec les rétrotransposons à LTR de part la proximité phylogénétique de leurs séquences [19]. Dans leur cas, les LTR sont remplacées par des répétitions inversées [20] et l'intégrase est remplacée par une tyrosine recombinase.

1.1.2 Les rétrotransposons sans LTR

Contrairement aux autres types d'ET, la séquence des rétrotransposons sans LTR n'est pas encadrée de motifs non codant spécifiques. Ces éléments possèdent cependant une structure particulière qui permet de les identifier. Les rétrotransposons sans LTR peuvent être décomposés en plusieurs grandes familles. On distingue les éléments autonomes des éléments non autonomes. Les éléments autonomes regroupent le clade des "Long Interspersed Nuclear Elements" (*LINE*) et celui des éléments de type *Penelope*. Les éléments non autonomes sont représentés par les "Short Interspersed Nuclear Elements" (*SINE*) et les *SVA* (éléments composites constitués de séquences de *SINE*, d'un nombre variable de répétitions en tandem ainsi que de séquences d'éléments *Alu*) [21]. Les rétrotransposons sans LTR comme les rétrotransposons à LTR peuvent transposer de manière autonome, tandis que les éléments *SINE* et *SVA* sont dépendants de la machinerie moléculaire encodée par les *LINE*.

Les *LINE* ont une taille de 5 à 8 kb et leurs différentes sous-familles (*R2*, *RTE*, *Jockey*, *L1* et *I*) codent toutes une transcriptase inverse (RT) et une endonucléase (voir Figure 2). Ces éléments possèdent aussi une queue poly-A à leur extrémité 3'. Les *LINE* ne possèdent pas d'intégrase et ont un cycle de transposition différent des rétrotransposons à LTR (voir Figure 2). En effet, après la phase de transcription en ARN, l'endonucléase va effectuer une coupure simple brin dans un fragment d'ADN cible. L'ARN va ensuite servir de modèle à la synthèse d'un brin d'ADN depuis l'extrémité du site de coupure par la transcriptase inverse. Le second brin d'ADN cible est à son tour clivé et le second brin de l'ET est synthétisé par les mécanismes de réparation de l'ADN de la cellule. Les rétrotransposons sans LTR possèdent deux ORF. Pour l'ORF 2, EN correspond à

endonucléase, RT à transcriptase inverse et RH à RNAseH.

Au cours de leur transposition, il arrive que les éléments de type *LINE* soient tronqués en 5', ce qui forme de nouvelles copies inactives connues comme copies "dead-on-arrival" [22].

Les éléments *SINE*, qui sont dépendants des *LINE* pour leur transposition [23], sont des séquences de 80 à 500 bp qui possèdent un promoteur de polymérase III en 5' leur permettant d'être transcrits, ainsi qu'une queue poly-A. C'est cette queue polyA qui est reconnue par la machinerie des *LINE* [24]. Le cycle de transposition de ces éléments est identique à celui des *LINE* et utilise les mêmes protéines.

Enfin les éléments de la famille *Penelope* [25] possèdent un ORF unique de 2,5 kb codant pour une transcriptase inverse et une endonucléase. Par opposition aux *LINE*, ces éléments sont encadrés par des répétitions terminales de 480 pb.

1.2 Les éléments de la classe II

Comme les éléments de la classe I, ceux de la classe II, appelés transposons, peuvent être décomposés en trois sous-classes : une sous-classe contenant les éléments possédant des répétitions terminales inversées à chaque extrémité de leur séquence (TIR pour l'anglais "terminal inverted repeat") et les éléments de type *Crypton* (trouvés chez les champignons mais aussi les oomycètes, les diatomées et les insectes [26], une sous-classe contenant des éléments Hélitrons et une sous-classe contenant les éléments de type *Maverick*. Les élément de la classe II ont tous en commun une absence de rétrotranscription (et d'intermédiaire ARN) pendant leur transposition.

1.2.1 Les transposons à ADN

Contrairement aux autres ET qui possèdent un mode de transposition réplicatif, le cycle de transposition des transposons à ADN ne leur permet pas directement de se multiplier dans les génomes. La séquence des transposons à ADN peut mesurer jusqu'à 5 kb et est encadrée par des TIR. La plus grande partie de la séquence de ces éléments est constituée d'un cadre de lecture codant une transposase (voir Figure 3). Ces éléments peuvent être regroupés en de nombreuses super-familles comme *Tc1-Mariner*, *hAT*, *Transib*, *P* ou *PiggyBac* [12, 27], et chez certaines familles la séquence protéique de la transposase présente un motif de trois acides aminés, DDE, caractéristique.

Le cycle de transposition des transposons à ADN est décrit dans la figure 3, avec dans un premier temps l'excision du transposon par la transposase, qui reconnaît les extrémités TIR et crée ainsi une rupture double brin à extrémités adhésives de l'ADN.

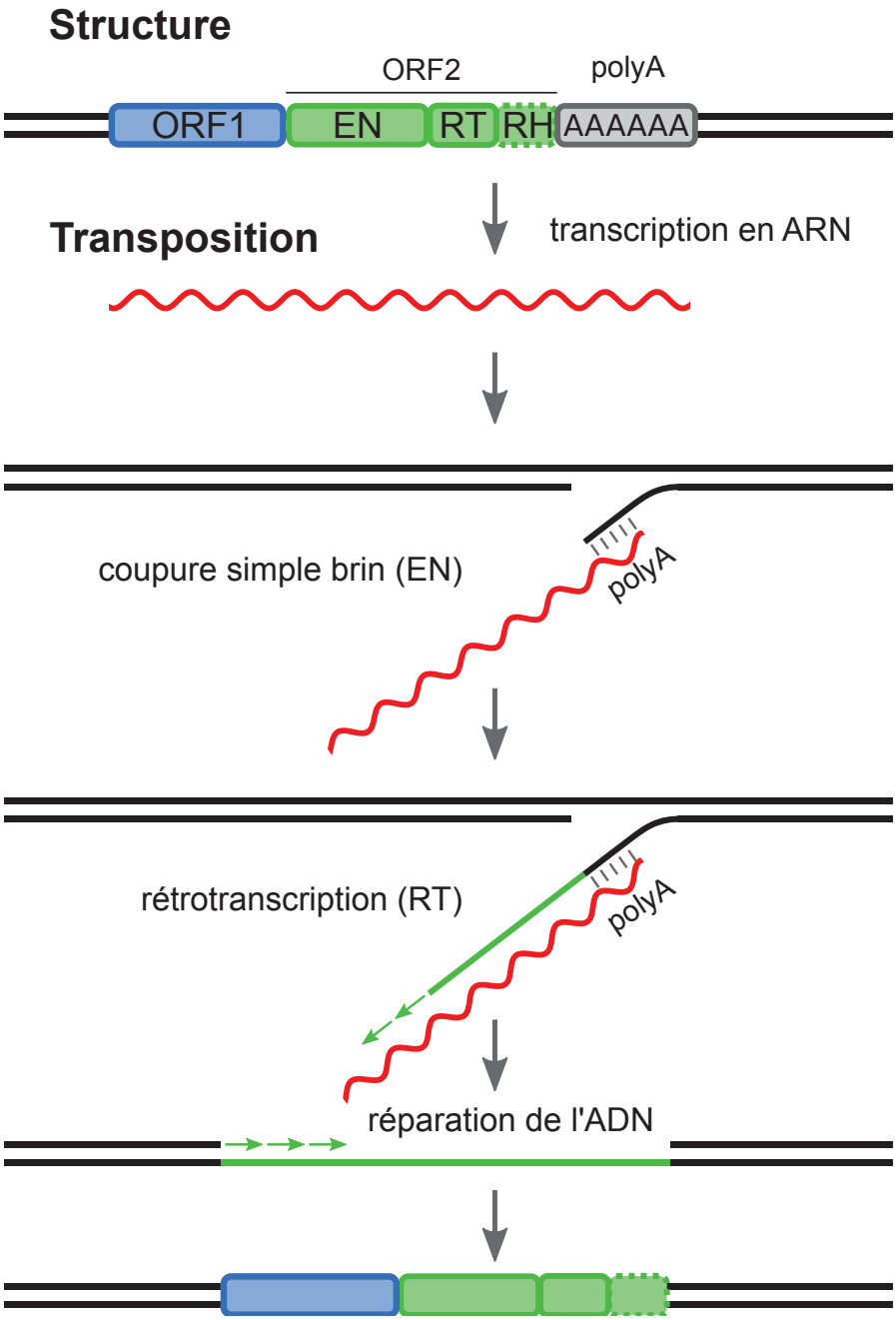


FIGURE 2 : Structure et cycle de transposition d'un rétrotransposon sans LTR. L'ADN est représenté par une ligne droite, en noir pour l'ADN génomique et en vert pour celui de l'ET, alors que l'ARN est représenté par une ligne ondulée en rouge. Les différents domaines protéiques sont représentés par des rectangles de couleurs. Les rétrotransposons sans LTR possèdent deux ORF. Pour l'ORF 2, EN correspond à l'endonucléase, RT à la transcriptase inverse et RH à la RNAseH.

La séquence double brin du transposon est ensuite insérée au niveau d'un site cible de transposition. Cette insertion va générer une signature de 2 à 11 nucléotides appelée "target site duplication" (TSD). Pour finir, la cassure double brin formée lors de l'excision du transposon est réparée par les mécanismes de réparation de l'ADN, soit en recollant les deux brin d'ADN, soit en reformant une copie du transposon à partir du chromosome homologue.

Il existe un autre groupe de transposons encadrés par des TIR, qui est celui des "miniature inverted repeat transposable elements" (*MITE*). Ces séquences d'une taille d'environ 500 pb ne possèdent pas d'ORF et, comme pour la relation des *SINE* avec les *LINE*, dépendent du matériel encodé par les transposons à ADN pour leur transposition [28]. La transposition de ces éléments se fait selon le même modèle que celui décrit dans la Figure 3.

1.2.2 Les Hélitrons

Les éléments de type Hélitron possèdent une structure ainsi qu'un mécanisme de transposition très différents des autres ET [29]. Découverts chez les plantes (*Arabidopsis thaliana*) [30], ces éléments codent tous pour une protéine de 1000 à 3000 acides aminés (AA) avec un domaine d'initiation de réplication en cercle roulant et une hélicase à ADN (HEL) [31].

Le mécanisme de cercle roulant des Hélitrons cible des sites AT et ce sont les seuls éléments de classe II à ne pas former de duplication de leur site cible au cours de leur transposition. C'est cette absence de site cible de transposition qui a retardé leur découverte *in silico* (*i.e.* pas des techniques bioinformatiques) chez les eucaryotes, alors que ces éléments peuvent représenter à eux seuls plus de 2% de la taille de certains génomes, comme celui du maïs [32]. Bien que le mécanisme exact de transposition des Hélitrons soit toujours élusif, il semblerait que ces éléments puissent transposer de manière réplicative et par excision du génome [33].

Le mécanisme supposé de transposition des Hélitrons, déduit de leur similarité avec celui des éléments à cercle roulant des bactéries, est présenté dans la Figure 4 [29]. Tout d'abord, une transposase forme une coupure simple brin en amont de l'élément ainsi qu'au niveau du site cible de transposition avant de lier les deux extrémités 5' formées entre elles. La coupure simple brin du site donneur va ensuite servir de point de départ pour la réplication du brin intact, déplaçant ainsi le brin de la séquence de l'élément qui est lié au cite cible par son extrémité 5'. Ensuite, la séquence liée au site cible de transposition va être coupée à son extrémité 3', soit à la fin de la séquence de l'élément,

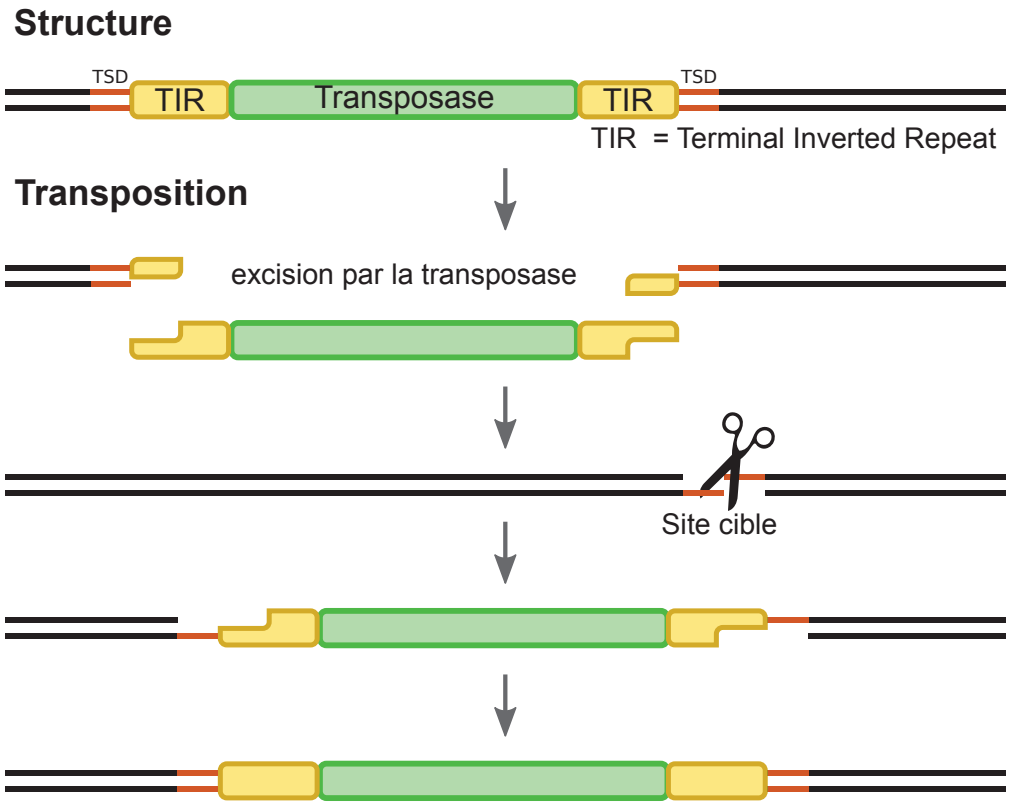


FIGURE 3 : Structure et cycle de transposition d'un transposon à ADN. L'ADN est représenté en noir et les différents domaines de la structure de l'élément sont représentés par des rectangles de couleurs. Au cour du cycle de transposition le site cible présenté en marron est clivé en une coupure à extrémités adhésives, ce qui va conduire à la duplication de ce site cible pour former le "Target Site Duplication" (TSD) de part et d'autre de la copie de l'élément au site d'insertion.

soit plus loin, selon si la fin de l'élément est correctement reconnue ou non. Dans le deuxième cas, cela peut conduire à la duplication de séquences voisines (représentées en bleu dans la figure). L'hétéroduplex formé au site cible de transposition est ensuite résolu pendant la réplication de l'ADN.

1.2.3 La classe Maverick

Découverte plus récemment en 2005 [34], la classe des éléments Maverick se distingue des autres transposons à ADN par plusieurs points. Ces éléments possèdent une taille importante allant de 9 à 22 kb et sont encadrés par de longs TIR [35]. Leurs longues séquences contiennent un grand nombre d'ORF (de 9 à 20), qui codent pour plusieurs protéines, dont une intégrase (INT), une ATPase, une protéase (PR) et une polymérase à ADN (POLB). Même si leur cycle de réplication reste inconnu, ces éléments sont supposés transposer comme des virus à ADN en formant des TSD de 6 pb au cours de leur insertion [34]. La POLB va permettre de synthétiser une copie de l'élément à partir d'une copie simple brin extrachromosomique (libre dans le noyau), qui va ensuite pouvoir être insérée dans le génome par l'intégrase à la manière des éléments de classe I. Les ET de la classe Maverick semblent dériver de virus à ADN ayant acquis la capacité de coder pour l'intégrase d'un rétroélément [36].

2 Effets des éléments transposables dans les génomes

Comme nous l'avons vu dans la section précédente, la fonction première des ET est de transposer et de se multiplier dans le génome de leur hôte. Une copie d'ET peut avoir différents effets pouvant se refléter sur le phénotype de l'hôte [37], en fonction de sa séquence et de son site d'insertion. Ces insertions peuvent provoquer des modifications qui vont changer l'expression des gènes environnant ou simplement être utilisés par l'hôtecomme matériel génétique nécessaire à l'acquisition de nouvelles fonctions [38, 39]. L'intérêt porté à l'étude des différentes familles d'ET a été fortement influencé par leurs effets sur leurs hôtes. Nous présenterons certains exemples de recrutement de ces séquences par leurs hôtes (exaptation) ainsi que les différents effets délétères qui ont été associés à l'activité des ET. Enfin, pour finir nous expliquerons les mécanismes permettant le succès des ET malgré leurs effets délétères.

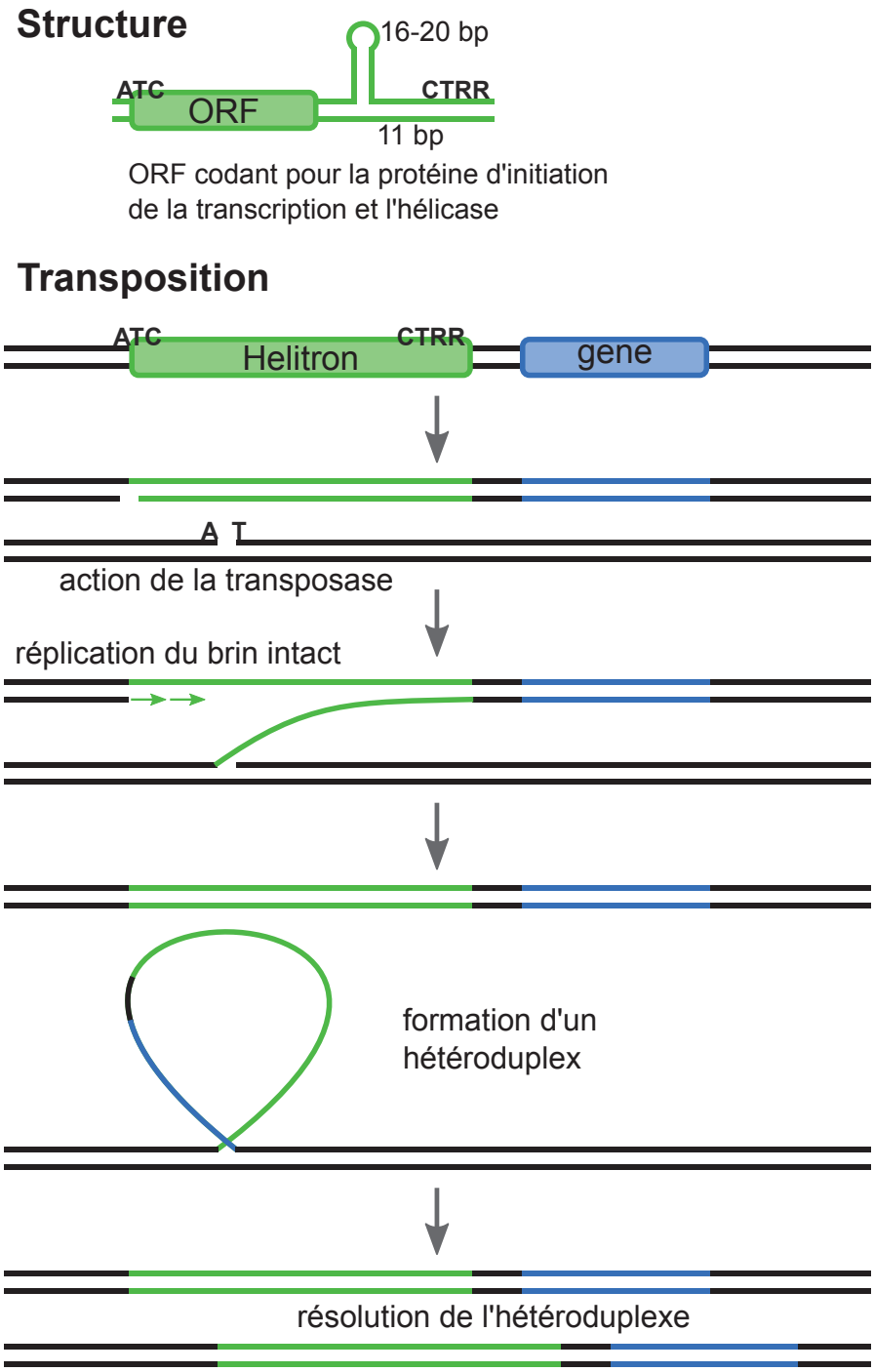


FIGURE 4 : Structure et cycle de transposition d'un hélitron. L'ADN génomique est représenté en noir et celui de l'élément en vert. Les différents domaines protéiques sont représentés par des rectangles de couleurs. Un fragment de gène est représenté en bleu pour illustrer le mécanisme de capture et de duplication résultant d'une transposition d'hélitron au cours de laquelle la réplication de l'élément ne s'est pas arrêtée au motif CTRR.

2.1 Exaptation de séquences d'éléments transposables

Comme nous l'avons dit au début de cette introduction, les séquences d'ET peuvent parfois être exaptées par leur hôte. Dans ce cas, ces séquences peuvent contribuer à la formation de nouvelles fonctions ou faciliter la colonisation de nouveaux environnements.

Il existe plusieurs exemples d'exaptation de séquences d'ET dans la littérature [40]. Chez les mammifères, les gènes *rag1* et *rag2* du système immunitaire sont un des exemples les plus connus d'exaptation de transposons à ADN [41]. Cette domestication a eu lieu il y a plus de 500 Ma chez les vertébrés. Les gènes *rag1* et *rag2* codent des enzymes similaires à des transposases et leur activité est au cœur du système de recombinaison *V(D)J* ("variable diverse and joining"). Ce système permet des réarrangements spécifiques d'ADN dans les cellules B et T du système immunitaire, qui sont utilisés pour obtenir une grande diversité de sites récepteurs pour leurs anticorps. Cette domestication de transposons à ADN joue donc un rôle primordiale dans les mécanismes de la réponse immunitaire acquise [40].

Un des exemples les plus remarquables de domestication de séquences provenant de rétrotransposons à LTR est celui de la formation du placenta chez les mammifères. L'histoire de l'évolution de cet organe est ponctuée de différentes exaptations de protéines de rétrotransposons à LTR. En effet, chez les marsupiaux et les euthériens (qui forment l'ensemble des mammifères possédant un placenta : les thériens), l'évolution du placenta est associée à un gène dérivé du gène *gag* d'un élément de la famille *gypsy* [42]. Chez les euthériens, chez qui le placenta a une structure plus complexe, il y a eu une seconde exaptation d'un gène similaire à *gag* provenant d'un autre rétrotransposon à LTR [43]. Enfin, l'acquisition d'un gène similaire au gène *env*, provenant d'un troisième type de rétrotransposon à LTR, a été détecté chez les primates et la souris [44].

Chez *Drosophila melanogaster*, la télomèrase, qui permet l'allongement de l'extrémité des chromosomes érodée à chaque division mitotique, a été perdue. Dans cette espèce, c'est l'activité de deux rétrotransposons sans LTR, *HeT-A* et *TART*, qui permet de compenser cette érosion [45]. En plus de présenter les caractéristiques d'une convergence fonctionnelle, ces deux éléments sont dépendants l'un de l'autre puisque seul *TART* code une transcriptase inverse et que le signal pour une association télomérique n'est présent que dans la séquence du gène gag de *HeT-A* [45]. Chez le riz et l'arabette, au moins 121 gènes seraient dérivés de séquence d'ET de la famille *Mutator* [46].

Il existe aussi de nombreux exemples de relation entre des adaptations environnementales et la présence d'ET [47]. Par exemple, chez la drosophile, une insertion d'ET a été associée avec des résistances à des virus et à un insecticide (l'azinphos méthyle

phosphate) [48]. Chez cette même espèce, des différences significatives en contenu d'ET ont été observées entre des populations proches mais exposées à des conditions environnementales différentes en Israël [49]. Ce type de variation a aussi pu être observé en corrélation avec des gradients climatiques entre le Sud et le Nord des États Unis d'Amérique ainsi que de l'Australie [50]. Pour conclure, certains auteurs vont même jusqu'à prédire que la présence et l'activité des ET chez les eucaryotes est un moteur essentiel de leur évolution et de la diversification de leurs espèces [51]. De ce point de vue, les effets délétères de certaines mutations provoquées par des ET au niveau des individus seraient moins importants que les gains apportés par le plus grand nombre d'allèles (différentes versions d'une séquence d'ADN pour une position donnée) qu'ils produisent et sur lesquels les différentes forces évolutives peuvent agir au niveau de la population. Ces mécanismes seront exposés plus en détail dans la troisième partie de cette introduction.

2.2 Effets délétères de l'activité des éléments transposables

Même s'il existe de nombreux cas d'exaptation liés à la présence d'ET [40], ces séquences répétées peuvent aussi être un important fardeau évolutif pour leurs hôtes. Un des effets les plus simples des insertions d'ET sur les gènes est l'induction d'une perte de fonction. Il a été démontré par exemple chez l'Homme que plus de 65 maladies étaient directement causées par des insertions d'ET [52]. L'activité de transposition des ET peut aussi modifier l'ordre des exons, allonger leurs séquences ou bien encore les dupliquer. Ainsi, les éléments *Alu* seraient responsables de plus de 30% des duplications de segments d'ADN chez l'Homme [53]. Dans le cas des Hélitrons, leur capacité à emporter des séquences voisines d'ADN lors de leur transposition leur permet aussi de capturer et de déplacer des séquences géniques. Cette caractéristique a notamment permis à ces ET de mélanger et déplacer plus de 20 000 fragments de gènes chez le maïs, chez qui ils ont été qualifiés de "exon shufle machine" [54]. Les insertions d'ET peuvent aussi avoir des effets plus subtils en modifiant l'expression des gènes lorsqu'ils transposent dans des promoteurs ou plus généralement au niveau des régions régulatrices des gènes [55]. Elles peuvent aussi directement former des régions régulatrices surnuméraires de gènes, comme c'est le cas pour presque 25% des séquences cis-régulatrices (sur le même brin d'ADN que le gène régulé) chez l'Homme [56]. Il existe aussi des effets indirects de la présence d'insertion d'ET qui sont liés aux mécanismes de régulation de ces éléments. En effet, les modifications de la structure de l'ADN pour la régulation pré-transcriptionnelle des ET, que nous allons développer dans la partie suivante, peuvent se propager et influencer l'expression

des gènes aux alentours [57]. Par exemple, il a été démontré chez certaines plantes que des insertions à plus de 3 kb d'un gène pouvaient en modifier l'expression [58].

L'effet des ET peut aussi se traduire à l'échelle du génome entier. Les mécanismes de recombinaison homologue sont responsables d'échanges de séquences d'ADN entre paires de chromosomes homologues chez les organismes diploïdes (qui possèdent chacun de leurs chromosomes en double). Ces mécanismes provoquent notamment la formation de nouvelles combinaisons de gènes à partir des deux génomes parentaux présents chez un individu, ce qui fait que la moitié de son matériel génétique présent dans ses gamètes est différente de celle transmise par chacun de ses parents. La présence à plusieurs endroits du génome de multiples séquences identiques d'ET peut aussi conduire à des recombinaisons entre séquences non-homologues et former selon les cas, des délétions, des insertions, des inversions, des translocations ou même des duplications de larges portions d'ADN, comme celles décrites dans la revue de Gray [59]. Ces effets sont par exemple responsables du phénomène de dysgénésie des hybrides qui conduit à des réarrangements chromosomiques importants ainsi qu'à une perte de la fertilité (stérilité) dans les individus issus d'un croisement entre des espèces différentes [60]. Enfin, de par leur activité de transposition, ces éléments sont aussi un fardeau énergétique puisqu'ils transposent en utilisant les ressources de la cellule hôte [61].

2.3 Maintien des éléments transposables dans les génomes

La citation suivante résume bien le statut de ces séquences dans les génomes et rappelle que malgré les nombreux cas d'exaptation, les ET ne sont pas des "ingénieurs" de la complexité des génomes eucaryotes : "The transposable elements propagate in a selfish mode 'because they can', and are then on occasion recruited 'because they are there'. Stochasticity dominates this process, which is not what an engineer usually wants." [62]. Même s'il existe des ET qui semblent cibler des sites de transposition minimisant l'effet de nouvelles insertions dans certains organismes [63], ce n'est pas le cas de toutes les insertions d'ET. Il semble donc naturel de considérer que le fardeau engendré pour l'organisme par la présence de ces parasites génomiques soit contre-sélectionné à l'échelle de la population. Cependant, pour les organismes diploïdes à reproduction sexuée la théorie peut donner des résultats contre-intuitifs [64].

Dans une population d'individus diploïdes sexués, la propagation des allèles est fortement impactée par le succès reproducteur associé à leur présence qui est appelé valeur sélective. En effet, comme chaque parent contribue à la moitié du matériel génétique de sa descendance, le matériel génétique des individus dont le succès reproducteur est plus

élevé va avoir une plus grande probabilité de se propager et d'augmenter en fréquence dans la population. À l'inverse, ce processus devrait conduire à une diminution de la fréquence des gènes ayant un impact négatif sur la valeur sélective des individus (voir Figure 5). Ce phénomène de sélection naturelle est par ailleurs d'autant plus important que le nombre théorique de génomes parentaux différents (ou taille efficace de la population) est grand. Quand la taille efficace de la population est faible, ce sont les effets des croisements aléatoires entre parents qui vont déterminer la composition génétique des individus de la génération suivante, et ainsi conduire à la fixation (c'est-à-dire la disparition d'autres allèles alternatifs) aléatoire d'allèles sans considération de valeur sélective. C'est ce que l'on appelle la dérive génétique (voir Figure 6). Pour les ET, les effets de la dérive génétique devraient favoriser l'augmentation en fréquence de leurs copies dans les populations. La dérive génétique aurait par exemple eu un effet permissif systématique sur les insertions d'éléments chez les vertébrés [65]. Par contre, les effets de la sélection sur les ET pour les organismes diploïdes à reproduction sexuée sont moins marqués, ce qui est expliqué par le fait que ce mode de reproduction favorise l'augmentation en fréquence des ET dans les populations [64].

Grâce à leur capacité à se multiplier, les ET ne sont pas soumis à la même ségrégation mendélienne que celle des autres allèles décrite précédemment. En effet, la capacité de transposer des ET leur permet d'assurer leur présence dans l'ensemble des combinaisons d'allèles possibles transmises à la descendance au lieu de la moitié de celles-ci. Ceci peut conduire au doublement de la fréquence de l'élément à chaque génération puisqu'il peut être présent dans toutes les combinaisons d'allèles parentaux. En conséquence, il faut des effets très importants liés à la présence de l'ET sur la valeur sélective de l'hôte (presque létaux) pour empêcher sa propagation dans la population. La reproduction sexuée peut donc être vue comme un formidable moyen de propager des ET dans les populations. Il est par ailleurs intéressant de constater qu'il existe de nombreux exemples d'interactions entre les ET et les mécanismes de différenciation sexuelle, comme par exemple dans le cas des chromosomes néo-X et néo-Y chez certaines espèces de drosophiles [66] ou celui du mécanisme de reproduction de la levure [67].

Les ET sont aussi présents chez les organismes où la reproduction est asexuée comme chez les procaryotes ou certains eucaryotes. Dans ce cas leur propagation dans la population peut être assurée par les nombreux échanges de plasmides bactériens qui ont lieu entre les individus d'une même colonie [68], ou encore d'autres mécanismes de transferts horizontaux que nous décrirons dans la dernière section de cette introduction. Néanmoins, l'absence de reproduction sexuée semble limiter l'invasion d'ET dans une espèce [69]. Les espèces unicellulaires possèdent généralement des tailles efficaces de population

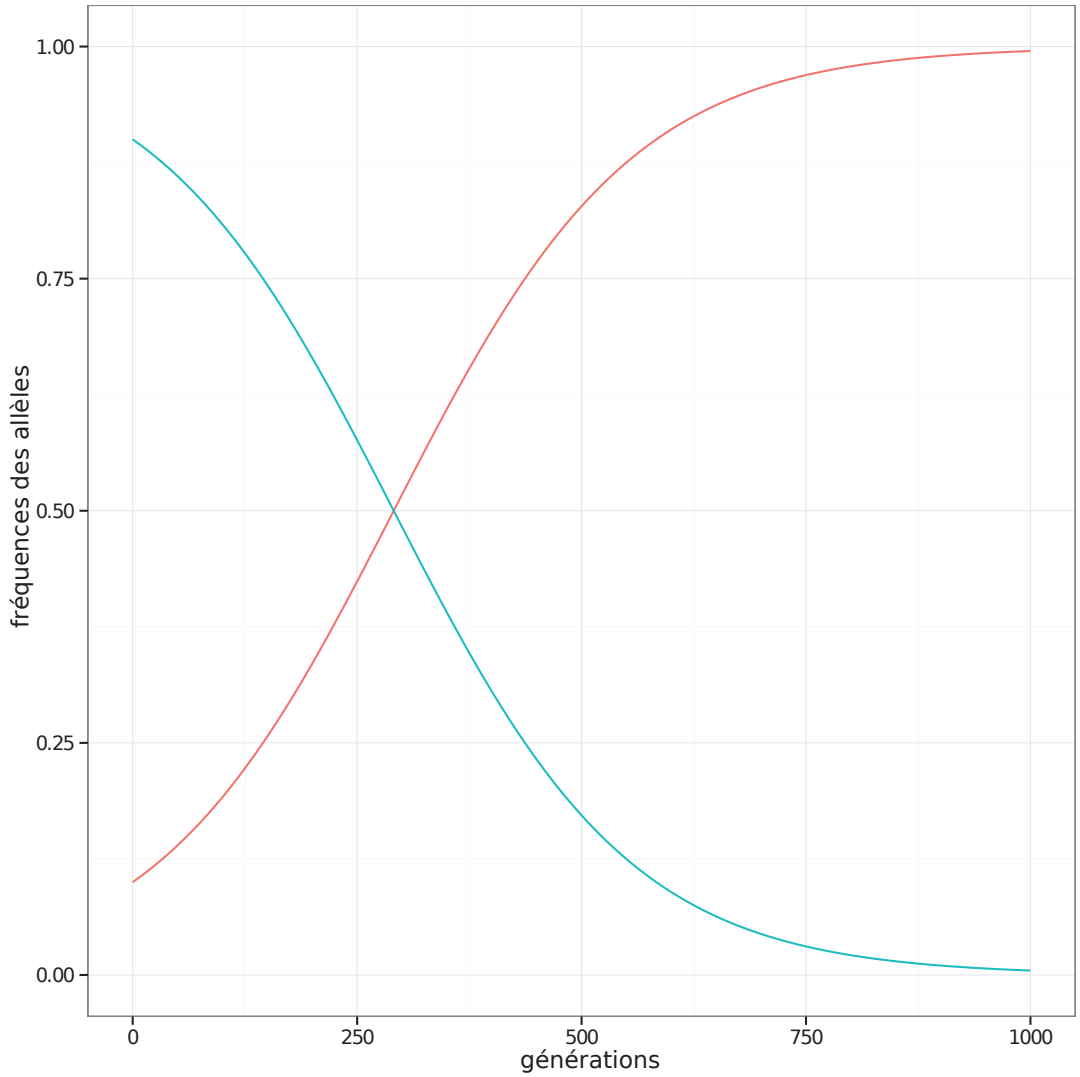


FIGURE 5 : Évolution de la fréquence de deux allèles d'un même locus dans une population de taille infinie au cours des générations. La fréquence des allèles est représentée en ordonnée et le temps en génération est représenté en absisse. L'allèle dont l'évolution de la fréquence est représentée en rouge n'est présent que dans 10% des individus au départ (temps zéro) et 90% de la population possède l'allèle alternatif présenté en bleu. L'augmentation de valeur sélective associée à la présence de l'allèle rouge conduit rapidement à sa fixation dans la population (sa fréquence atteint 1), alors que l'allèle bleu est éliminé de la population.

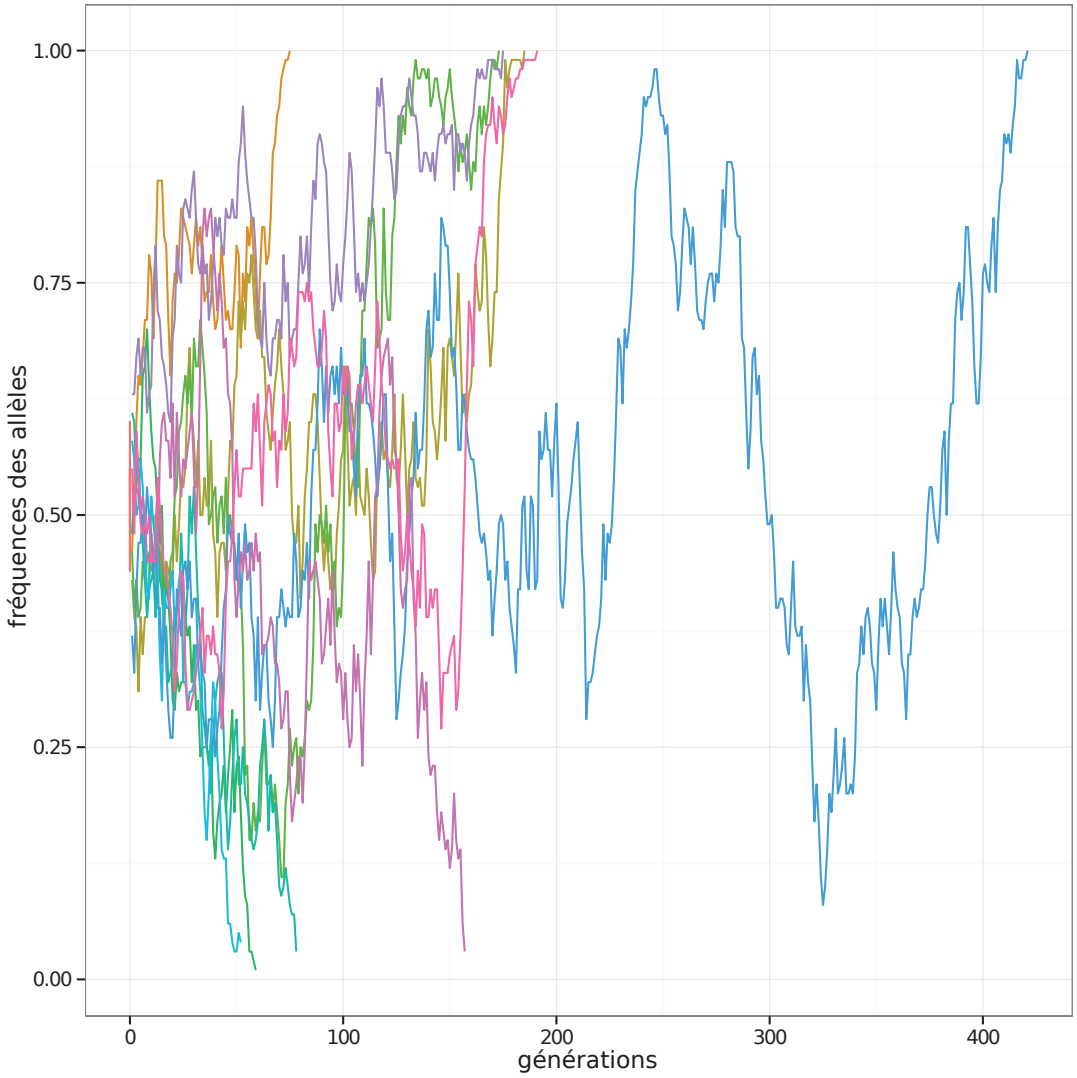


FIGURE 6 : Évolution de la fréquence d'un allèle possédant une fréquence initiale de 0.5 dans une population de taille finie. La fréquence des allèles est représentée en ordonnée et le temps en génération est représenté en absisse. La taille de la population est de 100 individus et est constante au cours du temps. Les différentes couleurs correspondent à 10 trajectoires possibles pour la fréquence de l'allèle étudié dans des conditions identiques. En effet, le devenir de cet allèle est soumis au hasard étant donné que cet allèle n'a pas d'effet sur la valeur sélective des individus.

qui sont de plusieurs ordres de grandeur supérieures à celles des organismes pluricellulaires. Ces différences de taille efficace devraient se traduire par des effets plus forts de la sélection sur le nombre de copies d'ET par rapport aux organismes multicellulaires. Une tendance associée à ce phénomène a été observée entre la proportion en ET du génome et la taille efficace estimée de nombreuses espèces [65].

3 Mécanismes de régulation des éléments transposables

De nombreux travaux de génétique des populations ont cherché à modéliser la dynamique des ET, et seuls les simulations produites à partir de modèles qui incorporent un terme de pénalisation proportionnel au nombre de copies d'ET dans un génome produisent des résultats réalistes [70, 71]. Ce facteur limitant l'activité des ET peut correspondre à plusieurs effets.

Le premier, moins réaliste, représenterait une saturation du génome en copies d'ET, ce qui signifierait soit que les ET diminuent leur taux de transposition lorsqu'ils dépassent une certaine densité dans le génome, soit qu'une fois la densité maximale atteinte, toute copie supplémentaire soit létale pour l'hôte. Ces hypothèses ne sont pas satisfaisantes pour plusieurs raisons. Comme nous l'avons vu précédemment, les effets des différentes insertions d'ET ne sont pas équivalents suivant leurs positions dans le génome. De plus, si le nombre de copies d'un ET augmente jusqu'à un maximum de copies supporté par l'hôte, le nombre de copies d'ET dans les génomes atteindrait des proportions bien supérieures à celles observées dans la nature avant de se stabiliser [72]. Un effet plus réaliste correspondrait à la mise en place de mécanismes de régulation de la transposition des ET par l'hôte. Il existe néanmoins un problème temporel pour la mise en place de ces mécanismes au niveau de la séquence d'ADN de l'hôte. En effet, le temps nécessaire à la sélection de mécanismes de régulation spécifique à un ET au niveau génétique et à la fixation de cette régulation dans la population, devrait être bien supérieure au temps nécessaire à un élément pour atteindre des proportions dommageables dans les génomes des individus de cette population [73].

La clef de la compréhension de la régulation des ET ne se situe pas au niveau génétique, mais au niveau épigénétique. Il existe d'autres manières pour la cellule de coder une information que par le biais de la séquence d'ADN, et l'on regroupe toutes ces informations supplémentaires sous le terme d'épigénétique. Cette information peut être transmise à la descendance d'une cellule ou d'un individu et être maintenue tout au long du programme transcriptionnel d'une cellule. L'information épigénétique est aussi beaucoup plus plastique que celle portée par l'ADN et peut être modifiée par différents

événements comme des stress, des changements environnementaux, ou lors des différentes phases du développement. Ces changements peuvent avoir lieu à des vitesses beaucoup plus importantes que celles des modifications de l'information génétique, puisque ces dernières doivent être disponibles à partir de mutations aléatoires des séquences avant d'augmenter en fréquence dans la population. Cette information épigénétique est par exemple à l'origine de l'existence de différents types cellulaires possédant la même séquence d'ADN, en modulant le programme transcriptionnel de ces cellules. Ces marques épigénétiques sont de différents types : la méthylation de l'ADN, les modifications d'histones ou l'ARN interférence par le biais d'ARN non codants.

Ce sont ces mécanismes épigénétiques qui semblent être au cœur de la régulation des ET [74]. Ils agissent à plusieurs niveaux : les modifications de la structure de l'ADN engendrées par la méthylation de l'ADN et les modifications d'histones forment une barrière pré-transcriptionnelle qui empêche la transcription des ET, tandis que les mécanismes d'interférence à ARN forment une barrière post-transcriptionnelle visant à dégrader les transcrits d'ET. Ces différentes voies de régulation sont étroitement interconnectées et dépendent souvent les unes des autres [75]. L'intérêt porté à la compréhension des mécanismes épigénétiques a donné lieu à de nombreux travaux pendant ces 10 dernières années, et chaque semaine de nouvelles études viennent renforcer nos connaissances sur ce sujet. La suite de cette section ne se veut donc pas une revue exhaustive des mécanismes de régulation épigénétique, mais un exposé des différents types de régulation épigénétique affectant l'activité des ET.

3.1 Régulation pré-transcriptionnelle

Il existe deux types de régulation pré-transcriptionnelle au niveau épigénétique qui sont la méthylation de l'ADN et les modifications d'histones. Ces deux modifications peuvent agir sur la transposition des ET en modifiant l'accessibilité de leurs séquences d'ADN à une éventuelle transcription. Dans cette section, nous décrirons ces deux mécanismes, sans détailler la façon dont ils peuvent cibler les séquences d'ET, qui sera décrite dans la section suivante portant sur les mécanismes d'interférence par ARN et les mécanismes de régulation post-transcriptionnelle.

3.1.1 Les modifications d'histones

La molécule d'ADN n'est pas libre et nue dans le noyau mais est associée à plusieurs protéines pour former une structure appelée chromatine. Cette chromatine est composée de nucléosomes qui sont des structures responsables de la condensation de la molécule

d'ADN. Chacun de ces nucléosomes est formé par 4 paires de molécules que l'on appelle des histones. On peut distinguer 5 familles d'histones appelées H1, H2A, H2B, H3 et H4. Chaque nucléosome correspond ainsi à 146 bp d'ADN enroulés deux fois autour d'un groupement de 8 histones composées des paires d'histones H2A, H2B, H3 et H4 [76] (voir Figure 7). Ces nucléosomes sont espacés par un segment d'ADN d'environ 50 pb et la fonction des histones H1, appelées histones de liaison, est de lier les extrémités du brin d'ADN au niveau de leur sortie du nucléosome [77]. La structure des histones a la particularité de posséder une queue d'acides aminés qui est accessible à la surface du nucléosome. Ce sont les modifications biochimiques de ces queues d'histones qui vont servir de support à une partie de l'information épigénétique.

Les queues des histones peuvent subir différentes modifications post-traductionnelles, comme des acétylations ou des méthylations. L'acétylation des histones correspond à l'ajout d'un groupement acétyle à des acides aminés via une protéine appelée histone acétyl-transférase. Ce type de modification entraîne un relâchement de la structure de la chromatine [78]. La méthylation des histones, qui correspond au transfert de 1 à 3 groupements méthyle sur des lysines, des arginines ou des histidines, via une histone méthyl-transférase, peut conduire à une ouverture de la structure de la chromatine, ou au contraire à sa condensation. Il existe aussi d'autres types de modifications chimiques, comme par exemple des phosphorylations, qui auraient une fonction de signal pour la réparation des coupures double brin de l'ADN [79] ou des ubiquitylations [76]. Les différents types de modifications à différentes positions des queues d'histones forment un code complexe qui n'est pas encore totalement élucidé et qui peut être différent selon les organismes [80] (voir Figure 7). Ainsi par exemple, pour la queue de l'histone H3, la triméthylation de la 4ème lysine, ou H3K4me3, est corrélée positivement avec l'expression des gènes, alors que la bi ou tri-méthylation de la 9ème lysine, ou H3K9me2 et H3K9me3, sont impliquées dans la répression des gènes [81]. On notera que la mono-méthylation de cette même lysine 9, ou H3K9me1, est au contraire retrouvée au niveau des promoteurs actifs chez l'Homme. Si les effets de ces différentes modifications d'histones peuvent être corrélés avec la répression ou l'expression de la transcription de certaines séquences, nous ne possédons des données que sur un faible nombre de modifications d'histones ainsi que sur les interactions qui peuvent exister entre elles. Néanmoins ces modifications sont reconnues pour avoir un effet sur la structure de la chromatine et sur la modulation de l'expression de différentes parties du génome.

Il existe donc différents états de condensation de la chromatine en fonction des modifications des histones qui lui sont associées [82]. Ainsi, nous pouvons distinguer l'euchromatine qui est constituée de régions où la structure de la chromatine est relâchée

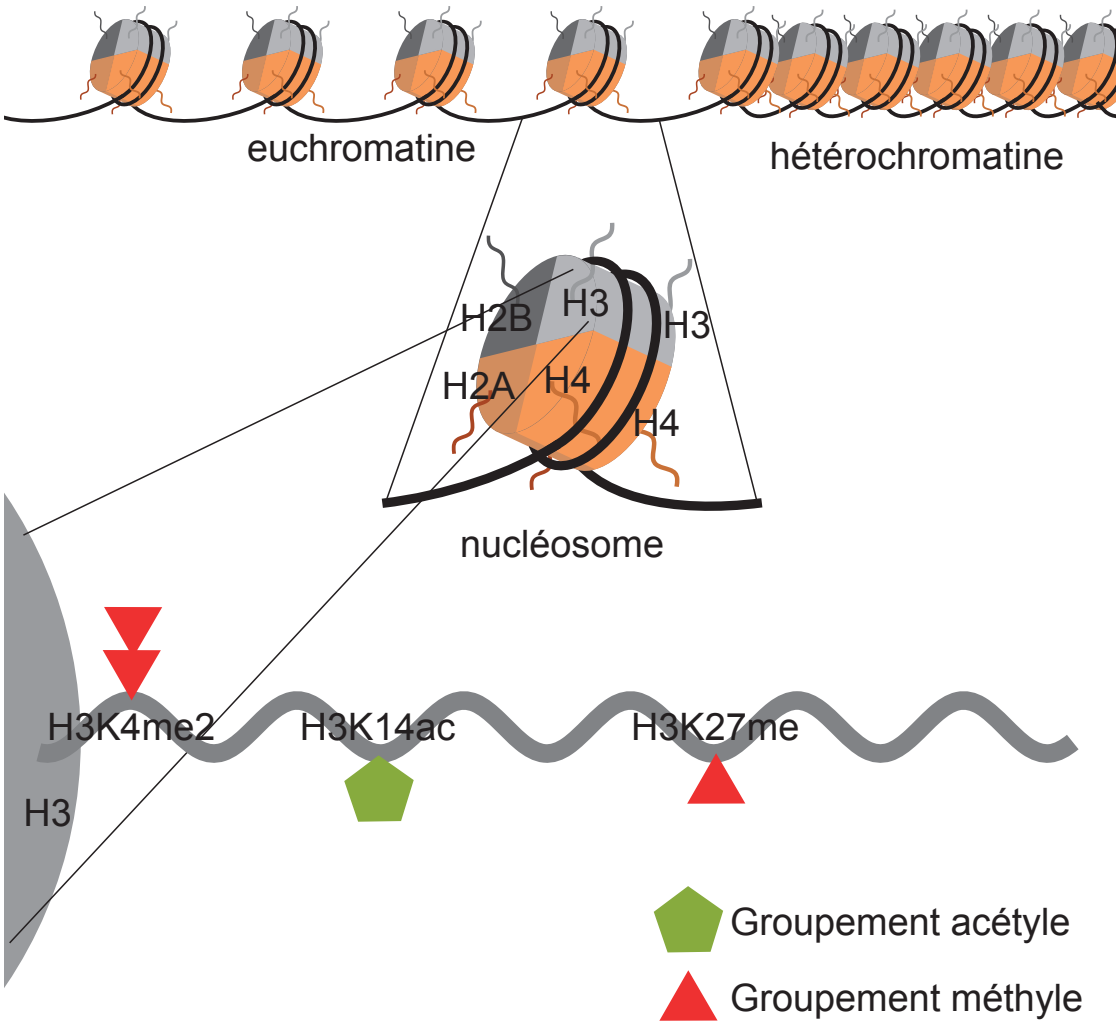


FIGURE 7 : Structure de la chromatine à différentes échelles. Les deux états de la chromatine, hétérochromatine et euchromatine sont présentés dans le haut de la figure avec l'ADN en noir enroulé deux fois autour de chaque nucléosome. La structure d'un nucléosome, formé de 8 histones est représentée au centre de la figure. Enfin un exemple de modification post-traductionnelle de queue d'histone est présenté au bas de la figure avec les marques H3K4me2, H3K14ac et H3K27me sur la queue d'une histone H3.

et où la molécule d'ADN est accessible, alors que la structure de l'hétérochromatine est beaucoup plus condensée [83] (voir Figure 7). Typiquement, c'est dans l'euchromatine que l'on retrouve la plupart des gènes, alors que l'hétérochromatine est pauvre en gènes et riche en séquences d'ET. Il existe deux types d'hétérochromatine, l'hétérochromatine constitutive et l'hétérochromatine facultative. L'hétérochromatine constitutive se situe principalement au niveau des télomères et des centromères des chromosomes et joue un rôle important dans les mécanismes de ségrégation des chromosomes ou ceux de la protection des télomères lors de la réplication de l'ADN [83]. Ce type d'hétérochromatine est un état irréversible de la chromatine qui est transmis au cours des divisions cellulaires. À l'inverse, l'hétérochromatine facultative est un état réversible de la structure de la chromatine qui permet par exemple de modifier le programme transcriptionnel des cellules au cours du développement. Ces modifications permettent aussi de limiter l'activité des ET en provoquant la condensation des séquences d'ET dans le génome. Chez la souris par exemple, les séquences de transposons à ADN sont associées avec la présence des marques répressives H3K9me3, H3K27me1 et H4K20me2, alors que les rétrotransposons à LTR sont associés à la marque H4K20me3 [84]. Chez la drosophile, les marques répressives H3K9me2 sont associées à la présence de séquences d'ET [85]. Chez cette espèce, les copies d'ET semblent s'accumuler dans les régions hétérochromatiques du génome où elles se retrouveraient piégés, avec peu de copies dans les régions euchromatiques [86].

3.1.2 La méthylation de l'ADN

Contrairement aux modifications d'histones qui influent sur la structure de la chromatine, la méthylation de l'ADN est une modification chimique directe de la séquence d'ADN, mais qui ne change pas l'information contenue dans celle-ci. Cette modification consiste en la méthylation du carbone en position 5 des cytosines, une des quatre bases constitutives de la molécule d'ADN. Chez certaines plantes, la méthylation peut aussi s'effectuer sur l'azote en position 6 des adénines [87]. La méthylation de l'ADN a particulièrement été étudiée chez les mammifères et les plantes chez qui elle est associée à une répression de la transposition [88]. Cet effet sur la transposition s'explique de deux façons : la méthylation de l'ADN a un effet négatif sur les interactions avec certains facteurs de transcription nécessaires au recrutement des ARN polymérases (responsables de la transcription en ARN) ; elle est aussi responsable du recrutement de protéines ("methyl-CpG-binding proteins") qui vont à leur tour recruter des protéines responsables de la condensation de la chromatine [89]. Le taux de méthylation de l'ADN est différent suivant les espèces. Par exemple, chez les mammifères, la majorité du génome est méthylé

à l'exception des régions riches en gènes [90], alors que le pourcentage de méthylation du génome des drosophiles serait inférieur à 1% [91]. Quand elle est présente, la méthylation, comme les modifications d'histones, permet de modifier le programme transcriptionnel d'une cellule. Ainsi, les profils de méthylation de l'ADN sont différents suivant les types cellulaires et sont modifiés au cours de la différentiation cellulaire [92, 93]. La méthylation de l'ADN est aussi impliquée dans la répression de l'activité de transposition des ET. Par exemple, pour les plantes, il semblerait que les séquences d'ET soient ciblées par les mécanismes de méthylation de l'ADN régulant ainsi leur répression [94]. Chez les mammifères, la déméthylation globale du génome, qui a lieu au cours du développement embryonnaire, coïncide avec une augmentation de l'expression des ET [95]. Un mécanisme similaire existe pendant les premières étapes du développement des cellules germinales au cours desquelles la méthylation de l'ADN est réinitialisée ce qui peut permettre aux ET de multiplier le nombre de leurs copies transmises de parent à enfant.

La méthylation de l'ADN et les modifications d'histones sont deux mécanismes de régulation de la transposition qui peuvent réprimer l'activité des ET. Nous avons vu que la méthylation de l'ADN pouvait provoquer des changements de la structure de la chromatine et que ces deux mécanismes étaient liés. Un exemple typique de perturbation des profils de modifications d'histones et de la méthylation de l'ADN est celui des cancers, au cours desquels ces instabilités épigénétiques importantes sont associées avec une augmentation importante de l'activité des ET [96]. Les mécanismes de régulation pré-transcriptionnelle sont donc un moyen efficace de réguler l'activité des ET (même si la méthylation de l'ADN ne semble pas être impliquée dans la régulation des ET chez la drosophile). Cependant, si ces modifications épigénétiques peuvent permettre de contrer la plupart des mécanismes de transposition des différentes familles d'ET, ils nécessitent de pouvoir cibler spécifiquement les séquences de ces éléments qui, comme nous l'avons vu, sont extrêmement diverses. Ce mécanisme de ciblage est assuré par la voie de l'interférence par ARN que nous allons maintenant présenter.

3.2 Régulation post-transcriptionnelle

Comme pour les mécanismes de régulation pré-transcriptionnelle, il existe différents mécanismes de régulation post-transcriptionnelle. Pour les ET, ce mode de régulation vise à détruire les transcrits d'ET, les empêchant ainsi de transposer dans le génome. Les différents mécanismes de régulation post-transcriptionnelle reposent sur la voie de l'interférence par ARN, dont la découverte a été récompensée par un prix Nobel de médecine en 2006. L'interférence par ARN permet la dégradation de transcrits ARN soit au cours

de leur transcription (régulation transcriptionnelle), soit après celle-ci (régulation post-transcriptionnelle). Ce processus repose sur des complexes formés d'un petit ARN et d'une molécule de la famille Argonaute. Ce type de molécule, qui possède une activité RNase-H, cible spécifiquement une molécule d'ARN complémentaire à la séquence du petit ARN, et la dégrade.

Il existe trois types d'interférence par ARN en fonction des petits ARN associés. Les miRNA, impliqués dans la régulation des gènes du génome hôte, les siRNA, découverts pour leur rôle dans la défense antivirale, et les piRNA, spécifiquement impliqués dans le contrôle des ET dans les cellules de la lignée germinale. Cependant les frontières entre ces différentes classes ne sont pas strictes : on trouve par exemple des ET contrôlés par des siRNA et des virus inhibés par des miRNA ou des piRNA. Les protéines Argonautes associées à ces petits ARN pour former les effecteurs de l'interférence par ARN peuvent être décomposées en deux grandes familles : la famille des protéines AGO impliquée dans la fonction des miRNA et des siRNA, et les protéines de la famille PIWI impliquées dans celle des piRNA [97]. Même si l'histoire évolutive des miRNA peut être liée à celle des ET, avec par exemple 55 des 462 miRNA de l'homme dérivant de séquences d'ET [98], cette voie de régulation concerne principalement l'expression des gènes de l'hôte et ne semble pas impliquée dans la régulation des ET, nous ne la présenterons donc pas ici.

3.2.1 Les "small interfering RNA"

La formation des "small interfering RNA" (siRNA) est semblable à celle des miRNA. Suivant les organismes, un fragment d'ARN double-brin (dsRNA pour l'anglais "double stranded RNA") qui peut être court ou long selon l'organisme, est découpé en siRNA par la RNase "Dicer" dans le noyau [99, 100]. Les siRNA double-brin formés sont ensuite exportés dans le cytoplasme. Puis le brin complémentaire de la séquence d'ARN est pris en charge par une protéine Argonaute pour former un complexe Ago-siRNA [101]. Chez la Drosophile, on peut distinguer la voie des siRNA issus de dsRNA endogènes qui sont transformés par un complexe Dicer-2 Loqs, et celle des siRNA issus de dsRNA exogènes qui sont transformés par un complexe Dicer-2 R2D2. Ces deux types de siRNA sont pris en charge par la protéine Ago-2 pour former des complexes siRNA-Ago2. Les dsRNA endogènes peuvent provenir de plusieurs sources dans le génome. Ils peuvent être le produit de transcription convergente, où les deux brins d'ADN sont transcrits dans des sens opposés, de la combinaison de transcrits complémentaires, ou de la transcription de clusters de piRNA. Chez *Arabidopsis thaliana* et *Schizosaccharomyces pombe*, l'activité des protéines RdRP (de l'anglais "RNA dependent RNA polymerase") permet aussi la

formation de ces dsRNA à partir d'ARN simple brin. Les complexes Ago-siRNA sont responsables de la dégradation d'ARN viraux, de la régulation des ET et de modifications de la structure de la chromatine [102]. Chez la Drosophile et la souris, certains siRNA proviennent de clusters de piRNA que nous allons présenter, mais ce phénomène pourrait être uniquement lié au fait que certains clusters de piRNA produisent des transcrits capables de s'assembler sous la forme de dsRNA [103, 104].

3.2.2 Clusters de piRNA

Contrairement à la voie des miRNA, la complémentarité élevée des siRNA et des piRNA avec leur cible implique une très grande diversité de séquences pouvant s'associer aux nombreuses familles d'ET. Ainsi, là où il n'existe que quelques centaines de miRNA différents, les différentes séquences de piRNA peuvent se compter par milliers. Pour comprendre la régulation des ET par ces deux mécanismes d'interférence par ARN, il est donc nécessaire de décrire l'existence de régions du génome riches en séquences d'ET appelées clusters de piRNA qui sont une des sources de cette diversité de séquences. Un cluster de piRNA est composé de nombreuses copies d'ET défectueuses dont la transcription produit de longs ARN non codants (lncRNA pour l'anglais "long non-coding RNA"). La taille de ces lncRNA peut varier de quelques kb, dans le cas d'une seule séquence d'ET, à plus de 100 kb pour les plus grands clusters [105]. Ces lncRNA permettent la formation de nombreux petits ARN pour les voies de répression des ET par les siRNA de même que les piRNA [106].

Les clusters de piRNA sont au cœur de la régulation des ET et certains modèles les considèrent comme des pièges à ET [107, 108]. Ces modèles décrivent un système similaire au système CRISPR chez les bactéries et chez les archées, qui ont la capacité d'intégrer activement des portions de séquences de virus ce qui leur permet de développer une réponse ciblée contre l'ARN de ceux-ci [109, 110]. À la différence de ce système, les clusters de piRNA semblent acquérir les nouvelles copies d'ET de manière passive. Une fois piégées, ces copies ne seront plus transcrites que sous forme de lncRNA. Pour pouvoir assurer la capture des copies d'ET ces clusters de piRNA doivent être nombreux dans le génome et l'insertion de copies d'ET dans ceux-ci ne doit pas provoquer d'effets délétères pour l'hôte [111]. Malgré sa simplicité, ce système, qui permet de dégrader spécifiquement un nouvel élément à partir d'une insertion de celui-ci, semble être très efficace puisque par exemple, chez *Drosophila melanogaster*, on n'observe qu'une seule séquence par famille d'ET dans ces clusters, ce qui pourrait correspondre à une régulation immédiate de l'élément inséré puisque chaque ET semble n'avoir pas eu le temps de s'insérer une

deuxième fois dans le cluster [108]. Une fois un cluster de piRNA formé, sa capacité à produire des petits ARN spécifiques de séquences d'ET peut être directement transmise à la descendance [112].

Chez la drosophile, la majorité de ces clusters est située à la frontière entre des régions d'hétérochromatine et d'euchromatine [111]. Les copies d'ET dans ces clusters sont soit orientées préférentiellement dans le sens inverse de la transcription du cluster, on parle alors de cluster unidirectionnel, ou bien indifféremment dans les deux sens, dans le cas des clusters doubles dont les deux brins d'ADN sont transcrits. Les clusters unidirectionnels sont majoritairement exprimés dans les cellules somatiques alors que les clusters doubles sont plutôt exprimés dans les cellules germinales [106]. Cette différenciation n'est cependant pas universelle puisque, chez la souris, les clusters de piRNA sont organisés de manière bidirectionnelle avec une partie du cluster possédant des copies sens et une autre possédant des copies anti-sens comme un enchaînement de deux clusters unidirectionnels [113, 114]. Les mécanismes contraignant le sens des insertions d'ET dans ces clusters restent mal compris.

3.2.3 Les "Piwi-interacting RNA"

À l'inverse des autres mécanismes de l'interférence par ARN, la formation des piRNA est indépendante des protéines Dicer et n'utilise pas de dsRNA [115]. Chez *D. melanogaster*, 90% des piRNA proviennent de 142 régions identifiées comme étant des clusters de piRNA [116]. Plusieurs modèles ont été proposés pour expliquer la genèse et la contribution des piRNA à la régulation des ET, mais ce sujet de recherche est encore récent et une grande partie des molécules impliquées dans ces mécanismes ainsi que leurs fonctions demeurent inconnues. Nous allons donc présenter le modèle de ce mécanisme de régulation chez *D. melanogaster*, qui est l'espèce dans laquelle il a été le plus étudié. Les mécanismes de régulation des piRNA semblent cependant être présents chez la plupart des animaux, avec la présence de deux ou trois protéines paralogues PIWI. Chez la souris et la drosophile, les trois protéines PIWI sont plus proches entre elles à l'intérieur de ces espèces qu'entre ces espèces, ce qui suggère des duplications indépendantes de ces protéines dans ces deux espèces. Néanmoins ces trois protéines ont des fonctions équivalentes entre ces deux espèces.

Chez la drosophile, ces trois protéines sont nommées Piwi, Aubergine (Aub) et Argonaute3 (Ago3). Les protéines Piwi ont une localisation nucléaire alors que les protéines Aub et Ago3 ont une localisation cytoplasmique [116]. Concernant les petits ARN qui s'associent à ces protéines, nous pouvons distinguer les piRNA primaires, majoritaire-

ment anti-sens aux transcrits d'ET et associés aux protéines Piwi, des piRNA secondaires associés aux protéines Ago3 et Aub suivant qu'ils sont respectivement dans une orientation sens ou anti-sens. Les piRNA primaires sont exprimés dans les cellules germinales ainsi que les cellules somatiques, alors que la présence des piRNA secondaires est limitée aux cellules germinales.

Même s'ils sont de tailles variables (entre 21 et 30 pb suivant les espèces) et qu'ils présentent une grande diversité de séquences, les piRNA ont des caractéristiques communes liées aux enzymes qui les produisent. Ainsi les piRNA anti-sens, qu'ils soient primaires ou secondaires, ont un fort enrichissement en uracile (U) en 5', alors que les piRNA secondaires sens présentent un fort biais en adénine (A) pour leur base numéro 10 [117].

Dans les cellules somatiques des ovaires de la drosophile, les piRNA primaires semblent être issus majoritairement de clusters dans lesquels les séquences d'ET sont orientées dans le sens inverse à celui de la transcription. La transcription de ces clusters unidirectionnels produit des lncRNA qui sont ensuite exportés dans le cytoplasme dans des structures granulaires appelées Yb-body. Avant leur passage dans les Yb-body, il semblerait que ces lncRNA séjournent dans une structure nucléaire appelée DotCOM [118]. C'est dans les Yb-body que ces lncRNA pourraient être découpés par l'endonucléase Zucchini en fragments d'ARN plus courts avec une uracile en 5' [105]. Ces fragments possédant une uracile en 5' sont ensuite liés à des protéines Piwi avec l'aide des protéines Shutdown (Shu) et Heat shock (HS) [119]. Une exonucléase inconnue, appelée Trimmer, va ensuite venir rogner l'extrémité 3' du petit ARN lié à Piwi et former un complexe Piwi-piRNA mature. Les Piwi-piRNA sont ensuite renvoyés dans le noyau de la cellule. Il semblerait que les Piwi-piRNA soient les seuls complexes de type Argonaute-piARN à ne pas découper leur cible ARN. Les complexes Piwi-piRNA ciblent les séquences d'ET en cours de transcription avec un brin d'ARN identifiable, où ils permettent la mise en place de la marque répressive H3K9me3 au niveau de la séquence de l'élément correspondant. Cette modification est effectuée par la protéine hétérochromatique 1 (HP1) qui est recrutée par ces complexes Piwi-piRNA [120, 121, 122, 123]. Chez la souris, chez qui la méthylation peut être utilisée pour la répression de séquences d'ET, les homologues de Piwi, MILI et MIWI2, semblent aussi provoquer la méthylation de leurs cibles [124]. Les complexes Piwi-piRNA sont donc responsables de modifications épigénétiques qui ciblent les séquences d'ET. Le processus de formation des piRNA primaires est décrit dans la Figure 8.

C'est dans les cellules germinales que le vrai combat contre la transposition des ET a lieu. En effet, ce sont ces cellules qui contiennent le matériel génétique transmis à la des-

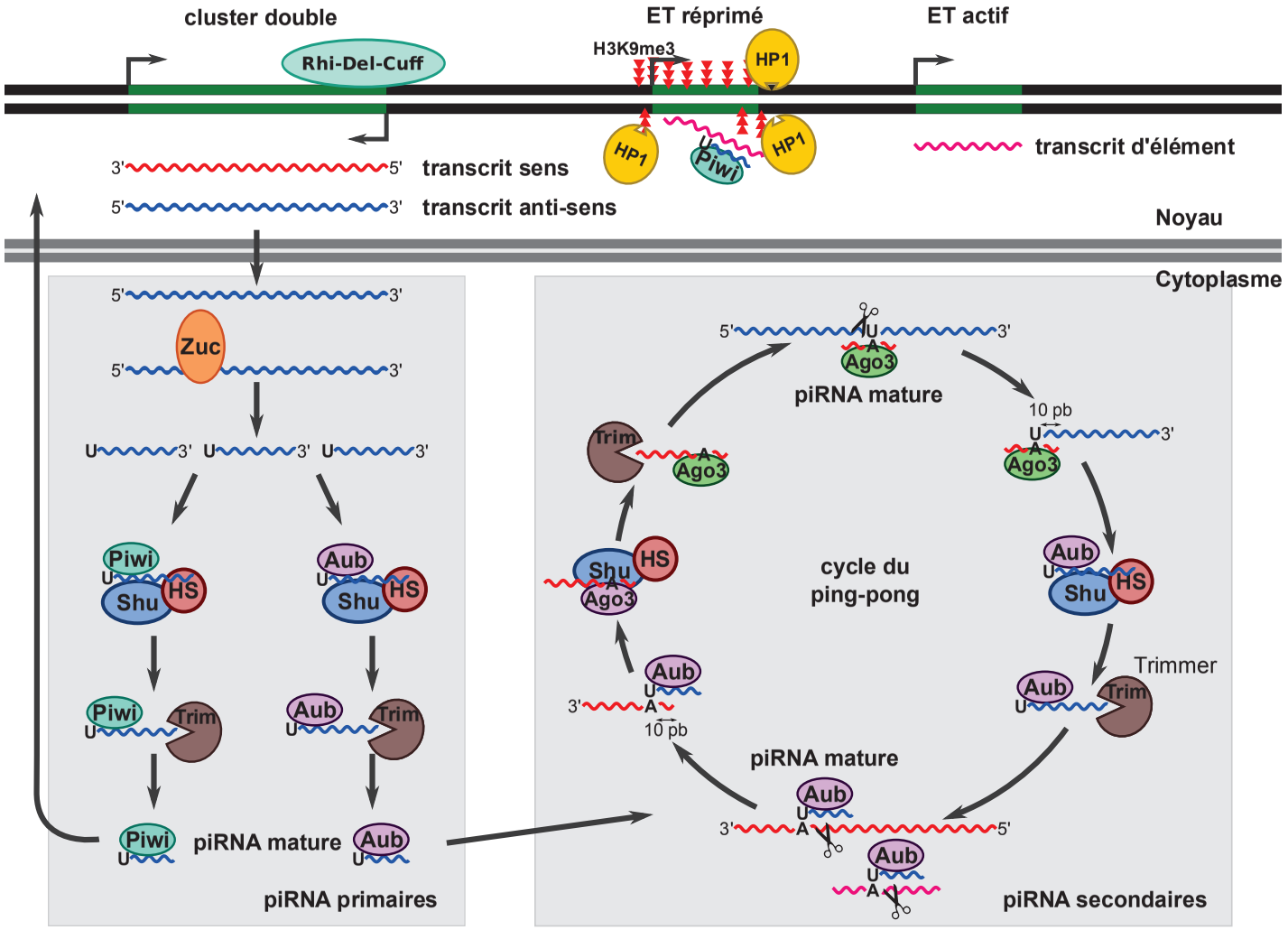


FIGURE 8 : Mécanismes de production des piRNA primaires et secondaires dans les cellules germinales. Pour la production des piRNA primaires dans les cellules somatiques, ce mécanisme est similaire mais commence par la transcription de clusters unidirectionnels qui ne produisent que des lncRNA anti-sens. Le nom des différentes protéines correspondent à Zucchini pour Zuc, Piwi pour Piwi, Aubergine pour Aub, Argonaute3 pour Ago3, Shutdown pour Shu, Heat shock pour HS, Trimmer pour Trim, protéine hétérochromatique 1 pour HP1 et au complexe Rhino-Cutoff-Deadlock pour Rhi-Del-Cuff. L'ADN génomique est représenté en noir alors que l'ADN correspondant à des séquences d'ET est représenté en vert (clusters de piRNA, ou copie d'ET). Les molécules d'ARN sont représentées par un trait ondulé, rouge dans le cas des transcrits sens aux séquences des copies d'ET et bleu dans le cas des transcrits anti-sens. L'activité RNaseH des complexes Aub-piRNA et Ago3-piRNA est représentée par une paire de ciseaux.

cendance, il est donc primordial d'y limiter l'activité des ET [125]. Il existe en addition à la voie des piRNA primaires, celle des piRNA secondaire dans les cellules germinales. Chez la drosophile, cette voie est initiée par la transcription de clusters dans lesquels le sens des séquences d'ET semble aléatoire. Ces clusters possèdent des promoteurs pour les deux sens de transcription. Les lncRNA produits sont ensuite exportés dans une structure cytoplasmique que l'on appelle le nuage. Chez la drosophile comme chez la souris, le transport des précurseurs de piRNA du noyau vers le nuage semble impliquer la protéine Maelstrom (Mael) [126]. C'est dans le nuage que les lncRNA anti-sens produits vont être découpés en petits ARN avec une uracile en 5'. Chez la drosophile, les protéines Piwi et Aub sont toutes deux capables de se lier à ces petits ARN. Comme pour les Piwi-piRNA, Trimmer rogne l'extrémité 3' des petits ARN liés aux protéines Aub et forme des complexes Aub-piRNA matures. Les complexes Aub-piRNA peuvent ensuite cibler des transcrits d'ET et les lncRNA sens et les dégrader. L'activité RNase-H de la protéine Aub forme une coupure du brin d'ARN complémentaire au petit ARN avec lequel elle est liée en position 10 du petit ARN formé (voir Figure 6). Les petits ARN sens formés possèdent donc une adénine en position 10 qui fait face à l'uracile en 5' des complexes Aub-piRNA. Ces petits ARN sont ensuite liés à la 3ème protéine PIWI, Ago3, après avoir été rognés en 3' formant des complexes Ago3-piRNA. Ces derniers peuvent à leur tour dégrader les transcrits anti-sens produits par les clusters de piRNA. Ce mécanisme forme une boucle d'amplification appelée ping-pong qui, à chaque cycle, dégrade des transcrits d'ET ainsi que des lncRNA en augmentant le nombre de complexes Aub-piRNA et Ago3-piRNA. Comme pour les complexes Piwi-piRNA, les protéines Shutdown et Heat Shock semblent avoir un rôle dans la liaison des piRNA avec les protéines Aub et Ago3. Le processus de formation des piRNA secondaires et le mécanisme du ping-pong sont décrits dans la Figure 8.

La boucle d'amplification du ping-pong semble aussi pouvoir être initiée par un dépôt maternel de complexes Aub-piRNA dans l'ovocyte [127]. Chez la drosophile, il existe aussi une interaction réduite entre Piwi et Ago3 qui permettrait de produire des complexes Piwi-piRNA à partir de ce mécanisme du ping-pong [128]. Ce ping-pong produit une signature caractéristique avec une population de piRNA anti-sens possédant une uracile en 3' et une population de piRNA sens avec une adénine en position 10 se chevauchant sur 10 pb. Cette signature a été retrouvée chez de nombreuses espèces, comme des éponges de mer, des vers plats, des papillons, des poissons, des grenouilles et des mammifères [125].

L'évolution des piRNA est rapide et leur adaptation à la présence d'un nouvel ET peut s'effectuer durant la vie d'un seul individu [129]. Des études récentes ont par ailleurs

démontré que des copies d'ET ciblées par les complexes Piwi-piRNA pouvaient rapidement être transformés en nouveaux clusters de piRNA [130, 131]. La clef de la formation de ces nouveaux clusters réside dans la différence fondamentale qui existe entre les clusters de piRNA unidirectionnels et les clusters doubles. En effet, pour les clusters doubles, la transcription des séquences sens d'ET devrait être reconnue par les complexes Piwi-piRNA et provoquer leur hétérochromatisation. Il semblerait que des complexes formés par au moins trois protéines, Rhino (Rhi), une protéine de la famille des HP1, Cutoff (Cuff), une protéine essentielle à la transcription des cluster doubles, et Deadlock (Del), soient la clef de ce paradoxe. Le complexe RDC (Rhi-Del-Cuff) pourrait reconnaître les modifications engendrées par l'activité des Piwi-piRNA et transformer ces régions hétérochromatiques en clusters de piRNA doubles. La nature des clusters formés par l'action de RDC serait donc épigénétique. Pour les clusters formés par RDC contenant des copies complètes d'ET, l'absence de Piwi provoque l'absence de RDC, et la réactivation de ces éléments [130]. Ainsi la production de Piwi-piRNA initiée par un dépôt maternel pourrait être nécessaire à la formation de clusters doubles façonnés par RDC [131].

Pour résumer, il existe différents mécanismes épigénétiques permettant la régulation des ET. Ces mécanismes sont efficaces ainsi que spécifiques à chaque séquence d'élément, et sont essentiels au maintien de l'intégrité des génomes. Ces mécanismes ne bloquent néanmoins pas totalement toute transposition dans la cellule puisque l'on observe tout de même des taux de transposition très faibles pour les familles contrôlées [132]. Par ailleurs, même si ces contrôles peuvent être maintenus au cours des générations, la nature volatile de l'information épigénétique fait que des événements de stress peuvent affaiblir ces contrôles et conduire à une augmentation importante du nombre de copies de certaines familles [133]. Ces stress pourraient donc être une alternative au phénomène de dérive génétique pour expliquer l'augmentation du nombre de copies d'ET [134].

4 Transferts horizontaux d'éléments transposables

Comme nous l'avons vu précédemment, il existe de nombreux mécanismes visant à contrôler l'activité des ET chez les eucaryotes. De plus, la plupart des séquences d'ET n'est pas sous sélection purificatrice, ce qui veut dire que les nouveaux variants formés par des mutations aléatoires de leurs séquences ne devraient pas être éliminés de la population par la sélection. L'accumulation de mutations aléatoires devrait donc conduire fatalement à l'inactivation définitive des ET et à leur disparition des génomes. Pour comprendre la dynamique et le maintien des ET dans les génomes, il convient de comprendre comment ces éléments peuvent échapper aux mécanismes permettant de contrôler leur

transposition.

Historiquement, les premiers modèles expliquant le maintien des ET dans les génomes reposaient sur l'existence d'un équilibre entre d'une part les forces antagonistes à la transposition des ET, comme la sélection naturelle et les différents mécanismes de régulation des ET, et d'autre part leur activité de transposition. Dans ces modèles, l'activité de transposition des ET doit être seulement suffisante à leur maintien dans les populations [70]. Cependant, le maintien d'un tel équilibre au cours du temps et des générations semble peu probable [135]. En effet, chaque mutation apparaissant au niveau des séquences des ET augmente la probabilité que cette copie perde son activité de transposition, tandis que chaque nouvelle insertion d'ET peut causer des effets délétères pour l'hôte.

Un modèle plus parcimonieux (*i.e.* implicant un minimum de causes) pour expliquer le maintien des ET est de considérer que ces séquences évoluent à des échelles plus grandes que celle d'une espèce, et qu'elles échappent aux mécanismes de contrôle en quittant le génome de leur hôte. Sous ce modèle, la survie des ET est liée à leur capacité à se transférer horizontalement vers une nouvelle espèce, ou population, dépourvue de mécanismes de régulation spécifiques contre ces éléments [136]. Un transfert horizontal (TH) correspond au mouvement d'information génétique à travers les différentes barrières reproductives qui existent entre des organismes plus ou moins proches. Cette définition large permet de regrouper les mécanismes permettant le passage d'information entre différentes espèces, comme ceux permettant ce passage entre individus de différentes populations. Les différents mécanismes de transferts horizontaux d'ET seront décrits dans une première partie. Ensuite nous présenterons les effets des TH sur les génomes et les espèces hôtes. Pour finir, nous introduirons le lien qui existe entre TH et dynamique des ET avec le modèle de naissance et mort [136].

4.1 Les mécanismes de transferts horizontaux

Les mécanismes de TH ont été particulièrement étudiés chez les procaryotes, chez qui ils sont fréquents et peuvent faciliter l'adaptation des individus à de nouveaux environnements [137, 138]. Ces organismes possèdent un panel de plusieurs mécanismes permettant le TH de matériel génétique d'un individu vers un autre [68]. Les transferts horizontaux chez les procaryotes peuvent ainsi être véhiculés par des agents de transferts de gènes (GTA pour l'anglais "gene transfert agents") encodés par la bactérie. Ces GTA sont similaires à des phages mais contiennent de l'ADN bactérien et ne possèdent pas les séquences d'ADN nécessaires à la production de leurs protéines [139]. Comme les GTA,

les plasmides peuvent être un vecteur permettant le transfert d'ADN entre les procaryotes. Ces séquences d'ADN circulaire peuvent cependant transporter des séquences d'ADN plus grandes et cibler une plus grande diversité d'hôtes [140]. Dans le cas des procaryotes, il est donc facile de concevoir qu'une séquence d'ET puisse être impliquée dans un TH d'un individu à un autre. Chez les eucaryotes, il n'existe pas de mécanismes dédiés aux TH d'ADN entre individus. De plus, la physiologie des eucaryotes elle-même présente différentes barrières à la possibilité d'un TH. Ces barrières peuvent correspondre par exemple à la présence d'un noyau qui isole le génome du reste de la cellule. Ou dans le cas des eucaryotes à reproduction sexuée au fait que seul un petit nombre de cellules germinales contient l'information génétique qui va être transmise à la descendance.

4.1.1 Les transferts horizontaux chez les eucaryotes

Les ET possèdent différentes caractéristiques qui peuvent en faire de meilleurs candidats à un TH que les gènes chez les eucaryotes. La première de ces caractéristiques est leur capacité à s'intégrer dans une séquence d'ADN. Deuxièmement, les caractéristiques des différentes familles d'ET pourraient aussi faciliter leur TH. Par exemple, pour les éléments de la classe I, certains rétrotransposons à LTR possèdent un gène d'enveloppe fonctionnelle et peuvent être infectieux comme c'est le cas pour l'élément gypsy chez *D. melanogaster* [16]. De manière plus générale, le fait que certains ET forment un intermédiaire d'ADN double brin stable durant leur cycle de transposition, devrait augmenter leur probabilité d'être transférés horizontalement [141]. De plus, durant leur transposition, certains ET peuvent adopter une conformation circulaire appelée épisome. Outre une stabilité plus importante que leur formes linéaires, ces épisomes peuvent conserver leur capacité d'intégration dans les génomes [142] ce qui pourrait faciliter leurs TH et leur intégration dans un nouvel hôte. Ces différences pourraient par ailleurs être en relation avec la proportion des différents types d'ET impliqués dans des TH. Par exemple chez les drosophiles, sur une centaine de cas de TH détectés, 52% concernent des transposons à ADN, 42% des rétrotransposons à LTR et 5% de rétrotransposons sans LTR [141].

Depuis le cas historique du TH de l'élément *P* chez *D. melanogaster* depuis une espèce d'un autre groupe de drosophile, *D. willistoni* [143], de nombreux cas de TH d'ET ont été découverts entre eucaryotes. Par exemple celui de l'élément *Mariner* entre *Zaprionus tuberculatus* et *D. mauritiana* [144] ou celui de l'élément *hobo* entre différentes espèces de drosophiles [145]. Ces exemples se sont multipliés au cours des vingt dernières années, avec un fort biais de détection chez les animaux, avec 94,37% de TH identifiés dans ce règne, contre 4,30% chez les plantes et 1,32% chez les champignons [146]. De plus, chez

les animaux, une écrasante majorité de ces TH a été détectée chez les drosophiles (54%). Plus récemment et grâce aux techniques de biologie moléculaire actuelles, la diversité des espèces analysées pour la détection des TH a augmenté avec l'augmentation du nombre de génomes séquencés disponibles. Par exemple, de multiples TH de quatre éléments de la famille de transposons ADN *hAT* ont été détectés entre différents tétrapodes comme la souris, la grenouille ou le lézard [147, 148]. L'élément *OC1* (un transposon à ADN de type *hAT*) a été observé comme transféré horizontalement entre différentes espèces de marsupiaux [149]. Les TH d'ET semblent donc être un phénomène commun à un grand nombre d'espèces au sein des eucaryotes. Chez les plantes par exemple, une étude récente a permis la détection de 32 cas de TH de rétrotransposons à LTR entre 40 génomes de plantes [150]. Cette étude estime à plus de 2 millions le nombre de transferts de ces éléments dans l'ensemble des espèces de plantes si l'on considère que leur taux reste comparable à ceux trouvés pour ces 40 espèces. Néanmoins pour la plupart des études de transferts horizontaux d'ET chez les eucaryotes, les mécanismes liés à ces transferts restent inconnus.

4.1.2 Les vecteurs comme support des transferts horizontaux

Les premiers mécanismes suspectés pour le transport de ces séquences d'ET entre individus sont le transfert par l'intermédiaire de virus à ARN pour les éléments de la classe I ou de virus à ADN pour les ET de la classe II, ou bien même de bactéries. L'utilisation de ces vecteurs pourrait faciliter le transport de séquences d'ET dans les cellules germinales de l'hôte. Cependant, la grande taille efficace des populations de ces vecteurs non sexués devrait aussi être corrélée avec une forte sélection pour éliminer les séquences d'ET de leurs génomes [151, 152]. Cette élimination constante de séquences d'ET dans le vecteur étant un obstacle à la détection de lien entre vecteur et TH d'ET, ce n'est que récemment que des preuves de ce mécanisme ont été détectées. Ainsi, le transfert d'un élément de type *Maverick* entre une guêpe parasitaire et un bracovirus a été récemment décrit [153]. Dans ce modèle le bracovirus est un symbiote obligatoire de la guêpe qui va réduire les défenses immunitaires des différents lépidoptères (papillons) que cette guêpe parasite. Les interactions entre cette guêpe et ce bracovirus ont fini par permettre le transfert et la fixation d'une séquence d'ET dans le génome du bracovirus. Ce bracovirus pourrait ensuite transférer cette séquence d'ET à d'autres espèces qu'il infecte. Un autre exemple plus récent concerne la présence de séquences d'ET transférées chez un baculovirus qui aurait permis sa dissémination dans différentes espèces de papillons. Dans ce cas c'est le nombre de génomes de baculovirus étudiés qui a permis la détection de séquences d'ET

[154].

Un autre type de vecteur pour le TH d'ET est celui des parasites. Dans ce modèle, les étroits échanges hôte-parasite pourraient permettre à des ET de passer d'un organisme à l'autre. En effet, la présence de séquences d'ET dans les fluides organiques est connue depuis longtemps [155] et pourrait permettre leur passage par le sang. Par exemple le TH de quatre ET (*SPIN*, *OC1*, *hAT1* et *ExtraTerrestrial*) à été récement mis en évidence entre différentes espèces d'animaux d'Afrique du Sud et d'Amérique du Sud, comme une espèce d'opposum, de singe, de grenouille ou d'escargot. En amérique du sud, l'espèce d'opposum et de singe sont toutes deux parasitées par le même invertébré suceur de sang, *Rhodnius prolixus*, chez qui ces ET sont aussi présents et qui pourrait donc être le vecteur de ce TH [156].

Enfin un dernier mécanisme de TH d'ET entre deux espèces proches est celui de l'introgression. Le mécanisme d'introgression correspond au transfert d'une séquence d'ADN entre deux populations par l'intermédiaire d'un hybride (voir Figure 9). Ce mécanisme facilite le passage d'une séquence d'ET dans la lignée germinale puisque cette séquence est déjà présente dans la partie du génome de l'hybride provenant de la population donneuse. Ensuite un rétro-croisement entre l'hybride et un individu de la population receveuse peut permettre le transfert de l'ET selon le hasard des crossing-over lors des divisions méiotiques chez l'hybride. Pour faire un lien avec la régulation post-transcriptionnelle des ET, nous pouvons souligner que l'introgression par le bais d'un individu femelle peut permettre la transmission d'un contrôle épigénétique de l'ET et augmenter la valeur sélective de l'individu recevant le fragment d'ARN introgressé dans la population receveuse.

4.1.3 Les transferts horizontaux dans les populations

Nous avons décrit plusieurs mécanismes permettant le TH de séquences d'ET entre individus de différentes espèces ou populations eucaryotes. Dans le cas d'une séquence transférée possédant une valeur sélective neutre, la fixation de cette séquence dans la population receveuse ne peut s'effectuer que sous l'action de la dérive génétique [157]. Cependant, l'activité de transposition des ET leur permet d'augmenter leur probabilité d'être présent dans le matériel génétique transmis à la descendance et leur probabilité de propagation dans la population (comme expliqué précédemment). Cette différence entre ET et gène explique sans doute les différences qui existent entre le nombre de TH réussis impliquant des ET par rapport à celui impliquant des gènes détectés chez les eucaryotes [158]. Nous avons vu qu'il existe de nombreuses barrières à la réalisation et au succès

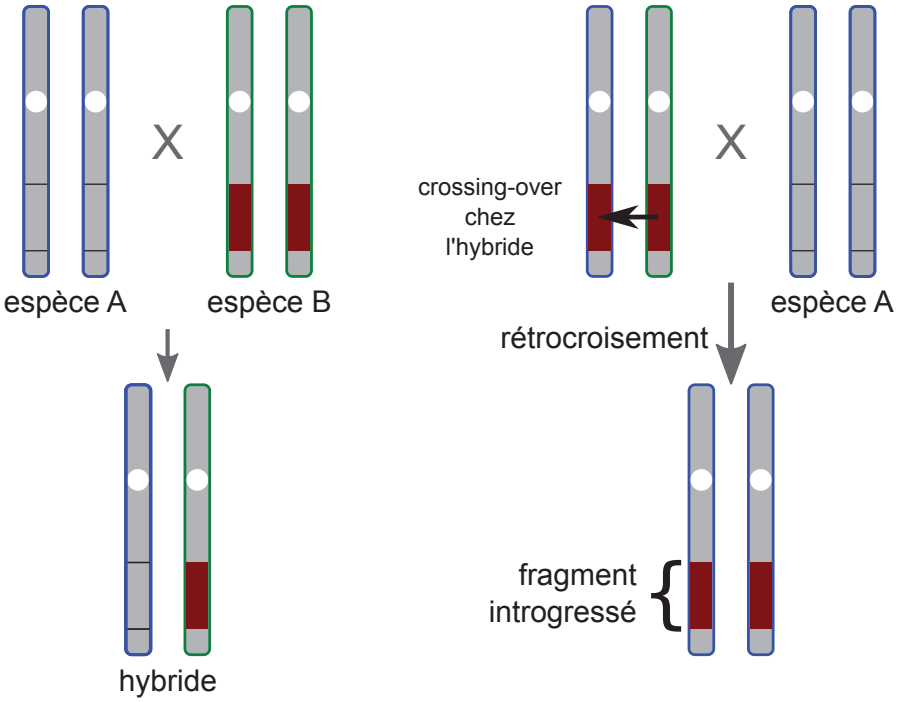


FIGURE 9 : Formation d'un hybride par un croisement entre une espèce A en bleue et une espèce B en vert. Un fragment d'ADN, en marron, va être transféré entre la partie du génome de l'hybride provenant de l'espèce B vers celle provenant de l'espèce A par crossing-over. Ensuite, un rétrocroisement avec un individu de l'espèce A va permettre l'introduction de la partie du génome de l'hybride contenant le fragment d'ADN marron dans l'espèce A.

d'un TH chez les eucaryotes. Une explication possible pour l'observation de nombreux événements de TH repose sur une grande probabilité d'événements de TH. Dans ce cas, même si la probabilité de fixation de l'ADN transféré dans la population receveuse est faible, de multiples TH ont lieu au cours du temps et aboutissent à la fixation de certaines des séquences transférées dans la population [157]. Ce scénario est particulièrement réaliste pour les transferts par vecteurs pour lesquels la probabilité d'intégrer la séquence transférée dans les cellules germinales est faible mais où les interactions entre les espèces et leurs parasites sont souvent nombreuses et sur une longue période.

4.2 Explosion de transposition

Pour décrire la dynamique des ET ainsi que le rôle des TH dans leur maintien dans les génomes, il est nécessaire de décrire le mécanisme d'explosion de transposition. Une explosion de transpositions a lieu suite à une perte partielle ou complète des mécanismes de régulation d'un ET. L'activité de l'élément concerné par cette absence de régulation va alors fortement augmenter, ce qui conduit à une multiplication rapide de son nombre de copies. Cette explosion de transposition peut être provoquée par des stress importants venant modifier les marques épigénétiques impliquées dans cette régulation, ou bien peut correspondre au TH d'un ET dans un génome qui ne possède pas encore de mécanismes de régulation spécifiques à cet élément.

4.2.1 Nombre de copies d'éléments transposables et explosion de transposition

Il existe différents exemples d'explosion de transposition dans la littérature [159]. Par exemple une augmentation de 40 copies par génération de l'élément *mPing* a été récemment observée chez des populations de riz [160]. Chez des espèces de drosophiles, des explosions de transposition de différents ET ont été documentées, comme le cas de l'élément *DINE-1* chez *D. yakuba*, celui de l'élément *412* chez *D. simulans* ou encore ceux des éléments *mdg1*, *mdg2*, *mdg3*, *copia*, *FB*, *P*, et *Doc* chez *D. melanogaster* (voir [161] pour une revue de toutes les explosions de transposition observées en laboratoire). Au niveau de la population, cette augmentation de l'activité d'un ET devrait favoriser une augmentation de la fréquence d'individus possédant au moins une copie de l'élément [135]. Cependant, si la transposition d'un élément dans les cellules germinales favorise sa propagation dans la population, ces transpositions peuvent se faire vers un très grand nombre de positions possibles dans le génome. Par conséquent la probabilité que la même insertion soit transmise par plusieurs combinaisons parentales est faible.

En l'absence d'avantage sélectif associé à la présence d'une insertion particulière dans le génome, on s'attend donc à ce que la fréquence des insertions individuelles d'ET dans la population soit faible. Par exemple, même chez *D. melanogaster* chez qui la plupart des familles d'ET sont jeunes et actives [162], une étude populationnelle récente a montré que seulement 1/3 des insertions d'ET étaient fixées dans une population du Portugal [163]. Pour le reste des insertions, environ 50% d'entre elles sont à des fréquences inférieures à 0.2, tandis que la dernière partie d'entre elles (les derniers 17%) se situe à des fréquences intermédiaires entre 0.2 et 0.95. La signature d'une explosion de transposition et d'une activité récente d'ET au niveau populationnel est donc la présence d'un grand nombre d'insertions d'ET en faible fréquence et d'une forte proportion d'individus possédant au moins une copie de l'élément [163].

4.2.2 Explosion de transposition et spéciation

L'abondance et la composition des familles d'ET est souvent spécifique à chaque espèce, ce qui peut suggérer un lien entre les ET et les mécanismes de spéciation [5]. Le modèle derrière ce processus est celui des équilibres ponctués [134, 164], qui décrit l'évolution des espèces comme un cycle de longues périodes stables ponctuées de courtes périodes d'intense spéciation. Des explosions de transposition pourraient alors apporter les changements génétiques nécessaires à la formation de nouvelles espèces. Par exemple dans le phénomène de dysgénésie des hybrides, un défaut de mécanismes de régulation spécifiques à certains ET peut rapidement créer une barrière reproductive entre une population possédant les ET en question et une autre qui en est dépourvue [165, 166]. Les changements génétiques importants qui sont les conséquences d'explosion de transposition peuvent aussi conduire à de nouveaux phénotypes et à l'isolation zygotique nécessaire pour une évolution divergente de deux populations [134].

Le fait que le contrôle des ET soit épigénétique a deux conséquences principales [21] : les effets de l'environnement qui modifient les marques épigénétiques ont une influence directe sur la dynamique des ET et les copies potentiellement actives des ET peuvent être maintenues dans un état de stase. Par conséquent, chaque génome transporte les outils nécessaires pour des changements importants de son organisation prêts à être activés par des stimulus environnementaux. Plusieurs études ont démontré l'effet de changements environnementaux, comme la température, le régime alimentaire, ou des stress, comme facteurs modifiant le taux de transposition des ET [167, 168, 169]. La capacité de transposition et les recombinaisons provoquées par les ET peuvent aussi avoir des effets directs sur l'évolution des hôtes en augmentant le nombre d'allèles différents soumis aux effets

de la sélection [6]. Les explosions de transpositions pourraient alors augmenter le potentiel adaptatif de l'hôte et favoriser la colonisation de nouveaux environnements [159]. Pour les cas les plus extrêmes, ces explosions de transposition pourraient produire une variabilité génétique suffisante pour que la sélection puisse promouvoir des combinaisons impossibles à effectuer par les autres mécanismes évolutifs [134, 164]. Cependant, si les modèles de l'évolution par équilibres ponctués incorporent les ET comme une source de variabilité génétique qui peut être mobilisée en cas de stress, les interactions avec les TH restent mal comprises.

4.3 Le modèle de naissance et mort

Le modèle de naissance et de mort des ET a été proposé en 2010 par Schaack et collaborateurs [136]. Ce modèle propose une solution élégante pour réconcilier l'efficacité des mécanismes de régulation de la transposition des ET et le fait que ces séquences soient toujours actives à ce jour. Les transferts horizontaux d'ET sont au cœur de ce modèle [158] dans lequel la naissance d'un ET correspond à son introduction dans un nouvel hôte, alors que sa mort correspond à l'inactivation de l'intégralité de ses copies (voir Figure 10).

4.3.1 Étapes du cycle de naissance et mort

La première étape du cycle de naissance et mort correspond à la naissance d'un ET dans un nouvel hôte par son arrivée par TH. Comme nous l'avons vu précédemment, la probabilité de TH dans le matériel génétique des cellules germinales d'un individu est faible [170]. À ces barrières physiologiques peuvent venir s'ajouter les effets populationnels de la dérive qui peuvent éliminer un faible nombre d'allèles porteurs de l'ET de la population [171]. Par conséquent, pour être observable, cette naissance doit être en fait multiple, avec un grand nombre de TH chez plusieurs individus de la population receveuse au cours du temps.

La deuxième étape du cycle de vie va correspondre à la colonisation du génome de l'hôte par l'ET. Cette colonisation doit se faire sur deux plans : par une augmentation du nombre de copies par génome, et une augmentation de la fréquence de la présence de ces copies dans les individus de la population. Comme nous l'avons vu dans les deuxième et troisième parties de cette introduction, la nature des ET et des mécanismes liés à leur régulation devrait favoriser cette invasion. En effet, on ne s'attend pas à avoir des mécanismes de régulation de la transposition de l'ET à l'intérieur d'un génome vierge de copies de cet élément, ce qui devrait conduire à une explosion de transposition

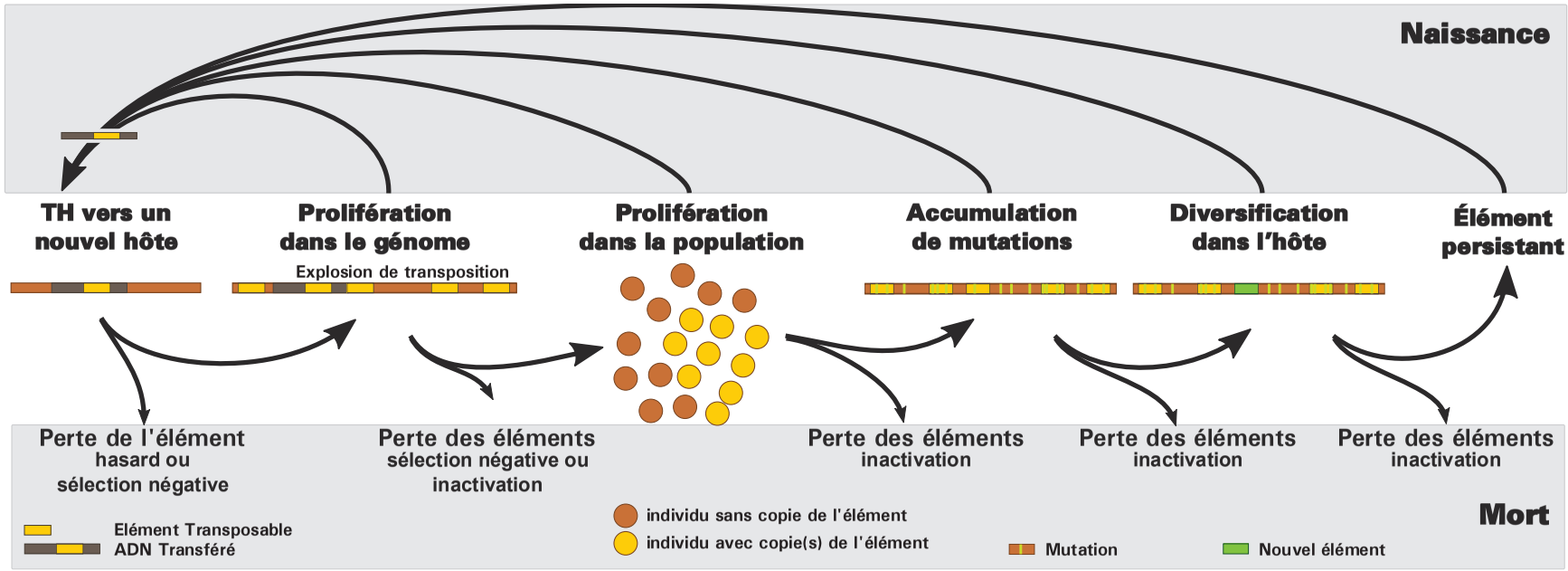


FIGURE 10 : Schéma du modèle de naissance et mort d'un élément transposable adapté de [136]. Les deux cadres gris regroupent les différents mécanismes conduisant à la mort d'un ET pour le cadre du bas, et les transferts horizontaux conduisant à sa naissance dans le cadre du haut.

et à l'augmentation en fréquence des génomes possédant des copies de l'ET dans une population à reproduction sexuée [171].

Pour finir, la troisième étape de la vie d'un ET correspond à son évolution à l'intérieur du génome de l'hôte, après la mise en place de mécanismes contrôlant sa transposition [129, 172]. Avec le temps, les différentes copies vont commencer à accumuler des mutations dégradant leur capacité de transposition ce qui va conduire à la mort de l'élément ou à la formation de nouvelles familles d'ET qui vont pouvoir recommencer ce cycle de naissance et de mort. Si la plupart des copies vont êtres perdues au cours du temps par attrition, certaines d'entre elles peuvent êtres exaptées et survivre en tant que séquences conservées [173]. Ces copies vont alors pouvoir persister dans la population grâce aux effets de la sélection. De manière surprenante, la régulation d'une copie d'ET dans un génome pourrait aussi contribuer indirectement à l'augmentation en fréquence de certaines de ses insertions dans la population [174]. Ce phénomène est expliqué par le gain en valeur sélective apporté par les copies transformées en cluster de piRNA. De plus en présence de ces copies transformée, les autres copies régulées peuvent être considérées comme neutres avec la présence de piRNA spécifiques à leurs séquences.

Il faut bien garder à l'esprit que les étapes de ce cycle de vie présentées ici de manière séquentielle, sont en réalité un continuum dans la nature. Par exemple, l'exaptation ou l'inactivation de copies par des mutations peuvent concerner les premières copies transférées horizontalement, ou les événements de transferts horizontaux peuvent survenir tout au long de la période de fixation de l'ET dans la population, créant ainsi de nouveaux foyers d'individus transmettant ces copies dans la population. Un aspect important de ce modèle pour expliquer le maintien des ET dans les génomes est que chaque copie de l'élément à chaque étape de son cycle de vie va pouvoir être la source de nouveaux TH vers d'autres organismes vierges de cet élément.

Ce modèle explique bien les profils d'ET observés chez les plantes, les poissons et les insectes. Dans ces espèces, on observe en effet généralement un grand nombre de petites familles d'ET qui sont composées de copies récentes de ces éléments [61]. En revanche, les profils d'ET observés chez les grands mammifères correspondraient plutôt à des phénomènes de co-évolution sur des échelles temporelles beaucoup plus importantes qui correspondraient au modèle d'un équilibre entre transposition et régulation. Chez ces espèces, on observe plutôt un petit nombre de familles d'ET composées d'un très grand nombre de copies dont la majorité est inactive. Le cas de la famille *LINE*-1 chez les mammifères est un exemple exceptionnel de longévité, puisque cette famille aurait été maintenue dans leurs lignées, sans TH, depuis plus de 100 Ma [175].

4.3.2 Naissance et mort des éléments transposables dans les populations

Le modèle de naissance et mort décrit précédemment peut aussi s'appliquer à l'échelle inter-populationnelle au sein d'une espèce. Jurka et al. [176] présentent un modèle dans lequel les individus d'une espèce sont regroupés en différentes populations d'individus appelés dèmes. Chacun de ces dèmes possèdent une taille efficace plus faible que celle de l'espèce ce qui provoque des effets plus importants de la dérive génétique. La fixation d'un ET est donc favorisée à l'échelle de chaque dème. Ces multiples environnements permissifs pourraient donc favoriser l'invasion du génome des individus de chaque dème par différentes familles d'ET. Enfin puisqu'il s'agit d'individus d'une même espèce, ces différentes familles d'ET peuvent ensuite être transférées de dème en dème, ce qui leur permet d'échapper aux mécanismes de régulation contrôlant leur activité de transposition.

Après un temps suffisamment long il serait même possible, dans ce modèle, que les mécanismes de régulation de certains ET ne soient plus maintenus dans certains dèmes. Par exemple, ces mécanismes pourraient ne plus être nécessaires avec la dégradation des copies correspondantes dans le génome de l'hôte. Ces mécanismes pourraient aussi n'être plus assez spécifiques par rapport à la version de l'élément toujours active dans la population. Dans ce cas, l'ET initialement fixé dans un dème pourrait effectuer une seconde colonisation de celui-ci. C'est ce qui semble être le cas pour l'élément *Penelope* qui serait en train d'envahir une seconde fois les populations de *Drosophila virilis* [172].

En plus de pouvoir expliquer le maintien et la diversité des ET au sein d'une espèce, ce modèle en dème décrit les mécanismes nécessaires à l'isolement reproducteur de certains dèmes et à la formation de nouvelles espèces. En effet, les différences de composition en ET entre populations peuvent former une véritable barrière reproductrice post-zygotique. Dans ce cas, bien que les individus soient assez proches morphologiquement et génétiquement pour pouvoir s'accoupler (ce qui constitue une absence d'isolement pré-zygotique), la survie d'embryons formés est compromise par une trop forte activité des ET. C'est le phénomène de dysgénésie des hybrides qui pourrait expliquer un des mécanismes de la formation de nouvelles espèces.

5 Conclusion

Nous avons pu voir au cours de cette introduction l'intérêt de l'étude des ET, ainsi que les différents mécanismes qui semblent gouverner leur dynamique. Les transferts horizontaux auraient un rôle prépondérant dans le maintien de ces séquences dans les

génomes actuels, et les explosions de transposition qui leur sont associées peuvent avoir des conséquences considérables sur l'histoire évolutive des espèces hôtes. D'autre part il existe des mécanismes complexes pour réguler l'activité de ces éléments dans les génomes, et les interactions qui existent entre les phases d'invasion, de régulation, et de perte d'un ET sont encore mal connues.

Les travaux effectués dans cette thèse s'inscrivent dans l'effort de compréhension de la dynamique des ET dans les génomes, avec le développement et l'utilisation d'outils bioinformatiques pour leur étude. Ces analyses ont été effectuées sur le modèle drosophile, pour lequel de nombreuses données concernant le génome et les ET sont disponibles. Dans le chapitre suivant de cette thèse nous présenterons les méthodes pour permettre l'analyse bioinformatique des ET dans les génomes dont la séquence ADN est disponible [14]. Nous présenterons dans le chapitre 2 une nouvelle méthode pour la détection de transferts horizontaux entre génomes complets d'eucaryotes et son application à deux espèces de drosophiles, *D. simulans* et *D. melanogaster*. La procédure de correction pour les tests multiples utilisée par cette méthode est présentée dans le chapitre 3. Le chapitre 4 présentera l'étude de la régulation des ET par la voie des piRNA chez différentes populations de *D. simulans*. Et le chapitre 5 présentera un nouveau programme, UrQt, dont l'utilisation a été nécessaire pour le traitement de des données sur les piRNA.

1

Analyse bioinformatique des séquences d'éléments transposables

Avec la progression des technologies de séquençage, nous avons accès à de plus en plus de séquences génomiques. Ce progrès a débuté avec l'avènement des technologies de séquençage en 1977 [177], qui ont permis le séquençage de génomes complets de différentes espèces eucaryotes modèles dans les années 1990 et 2000 [178, 179, 180]. Ce type de technologies, appelées plus tard séquençage de première génération, permet d'obtenir des lectures d'ADN de plusieurs centaines de paires de bases (pb). Cependant ce type d'approches reste chères pour des projets de séquençage de génomes complets et il a fallu attendre 2005, avec la commercialisation des méthodes de séquençage de deuxième génération, pour que les projets de séquençage de génomes eucaryotes se démocratisent [181]. Ces technologies de seconde génération ou NGS (de l'anglais "next-generation sequencing"), permettent d'obtenir un très grand nombre de petites lectures d'ADN à un bien plus faible coût que les méthodes de première génération. De plus, l'amélioration de ce type de séquençage depuis 10 ans a permis une augmentation du nombre de lectures produites (plusieurs dizaines ou centaine de millions) ainsi que de leur taille (plusieurs centaines de pb) [182]. Le développement de technologies dites de troisième génération pourrait permettre de repousser encore plus ces limites avec des lectures de plusieurs milliers de pb [183].

L'avènement des technologies de séquençage a grandement contribué au développe-

ment de la bioinformatique et des méthodes d'analyse de ces séquences. Cependant, comme les séquences répétées peuvent représenter une part importante du génome de leur hôte, une des premières étapes d'une grande partie de ces analyses consiste à les exclure des données à analyser. En effet, historiquement l'analyse de ces séquences d'ADN a été plutôt tournée vers l'étude des gènes, et la présence de séquences d'ET codant pour des protéines venait perturber leurs prédictions. Un des programmes les plus connus pour effectuer cette étape est RepeatMasker dont la principale fonction est de masquer les séquences d'ET dans les séquences d'ADN [184]. Même si les séquences d'ET sont souvent mises de coté et peu étudiées dans la plupart des projets de séquençage de génomes, il existe de nombreux programmes et de bases de données dédiées à leur étude.

Plusieurs articles présentant une compilation des différentes méthodes d'identification et de classification des séquences d'ET ont été publiés [185, 186, 187, 188]. Ces articles portent principalement sur les approches "classiques" de l'étude des séquences d'ET qui sont applicables à des génomes séquencés et assemblés. L'assemblage d'un génome consiste à la résolution, par des méthodes bioinformatiques, du puzzle formé par les millions de courtes lectures produites lors de son séquençage [189]. Les méthodes d'identification des séquences d'ET peuvent être décomposées en trois grands groupes. Le premier groupe utilise des méthodes de similarité avec la détection de séquences déjà connues et présentes dans des bases de données comme par exemple RepBase [8]. Le deuxième groupe profite des différentes caractéristiques structurales des séquences d'ET pour essayer de les identifier, comme la présence de LTR ("long terminal repeat") pour les rétrotransposons à LTR. Enfin le dernier groupe utilise le fait que les séquences d'ET sont répétées dans les génomes pour les identifier, en cherchant par exemple les k-mers (ou mots de k lettres) sur-représentés dans le génome. De plus, l'augmentation de la quantité de données NGS ouvre de nouvelles possibilités pour l'étude et l'identification des ET chez les eucaryotes, comme de caractériser les séquences d'ET de génomes non assemblés [190], ou alors de pouvoir étudier leur polymorphisme d'insertion à partir de données de séquençage de plusieurs individus d'une même population [191]. Ce sont aussi ces technologies NGS qui permettent d'étudier plus précisément les mécanismes de régulation des ET ainsi que leur expression, et qui ont permis la caractérisation de leur régulation par les piRNA chez la drosophile [116]. L'analyse bioinformatique des séquences d'ET est un sujet de recherche dynamique pour lequel de nombreux programmes sont publiés chaque année. C'est pourquoi il est nécessaire de publier en parallèle des comparaisons de ces programmes ainsi que des conseils sur leur utilisation.

C'est dans ce contexte que ma directrice de thèse, Emmanuelle Lerat, a été invitée à rédiger le chapitre "Identification and analysis of transposable elements in genomic

sequences" du livre intitulé "Genome analysis, current procedure and application". J'ai eu l'opportunité de participer activement à la rédaction de ce chapitre en me chargeant des parties dédiées à l'analyse des séquences d'ET dans l'ère des technologies NGS.

Ce chapitre de livre, inclus ci-après, a été rédigé à la fin de l'année 2012 et est donc représentatif de l'état de l'art à cette période. Plusieurs autres programmes pour l'analyse bioinformatique des ET ont été publiés depuis. Pour n'en citer que quelques uns, la recherche d'ET par homologie de séquence peut maintenant être effectuée de manière beaucoup plus sensible avec le programme NHMMER, qui permet d'identifier des séquences d'ET fortement dégradées dans un génome [192]. Ce programme utilise un modèle probabiliste basé sur des chaînes de Markov pour identifier des séquences homologues entre elles et est notamment utilisable dans les versions les plus récentes de RepeatMasker. En ce qui concerne l'alignement de lecture NGS contre une référence, la version stable de Bowtie2 a été publiée [193] . Ce programme permet d'avoir accès à la plupart des avantages d'une recherche par BLASTN, comme la possibilité d'avoir des portions de séquences séparées les unes des autres ou d'avoir plusieurs nucléotides qui ne correspondent pas à la séquence cible dans un alignement, combinée à la rapidité des algorithmes basés sur la transformation Burrows-Wheeler [194]. L'étude de la phylogénie des différentes copies d'un ET peut apporter des informations importantes sur la dynamique de cet élément dans un génome. Le programme AnTE permet de reconstruire des arbres phylogénétiques de copies d'ET en prenant en compte les différences qui existent entre les copies dont la transposition a formé de nouvelles copies et celles qui n'ont pas transposé [195]. Plus spécifiques à l'assemblage de séquences d'ET à partir de données NGS, les programmes RepARK et TEdna ont été publiés cette année [196, 197]. Ces programmes utilisent le fait que les ET sont présents en plusieurs endroits du génome pour les assembler à partir de la liste des k-mers sur-représentés dans les lectures. Cette idée avait déjà été utilisée par le programme RepeatExplorer, pour pouvoir caractériser le contenu en ET d'un génome à partir d'un échantillonnage des lectures NGS [190].

Les méthodes d'assemblage de séquences NGS spécifiques aux ET permettent de contourner le problème de l'assemblage des séquences répétées dans un génome. La difficulté de l'assemblage des régions répétées d'un génome (dont les séquences d'ET) est principalement liée à la petite taille des lectures NGS. En effet, dans le cas où la taille d'une lecture est inférieure à celle d'une répétition (ce qui est souvent le cas pour les ET), il est difficile d'assigner cette lecture à une position spécifique du génome. La contrepartie de ce problème est que pour le séquençage uniforme d'un génome, par exemple avec une couverture de 10x (signifiant que chaque portion du génome apparaît en moyenne dans 10 lectures différentes), la séquence d'un ET présent en 30 copies aura

une couverture moyenne de 300x. Réciproquement, pour une couverture inférieure à 0,5x les seules parties du génome que l'on peut assembler sont les régions répétées. Ainsi pour un échantillonnage de 0,5x, un ET présent en 30 copies aura une couverture de 15x, et un élément présent en 100 copies une couverture de 50x. La couverture d'un ET dépend donc de son nombre de copies dans le génome séquencé.

Il existe une analogie entre ces différences de couverture pour un ET donné et les différences de nombre de transcrits ARN produits à partir de différentes portions d'un génome. Le séquençage de transcriptomes, ou RNA-Seq, est similaire à celui de l'ADN avec l'ajout d'une phase de transcription inverse pour transformer les ARN d'une cellule en ADN avant de les séquencer. Étant donné que l'activité de différents gènes peut être très variable suivant le type de gène, de tissus ou les conditions, il existe des différences importantes entre le nombre de lectures associées au transcrit d'un gène, par rapport à un autre. Des différences de couverture entre les transcrits, en plus de la présence de transcrits alternatifs possédant des portions de séquences communes, nécessitent l'utilisation de programmes d'assemblage spécifiques au RNA-Seq [198]. Ces caractéristiques du RNA-Seq sont comparables avec les données de séquençage de copies d'ET. En effet, suivant le nombre de copies d'un ET présent dans le génome, on retrouve aussi ces variations de couverture d'un ET à l'autre, et suivant le polymorphisme de ces copies ou la proximité phylogénétique de ces ET, certaines portions de ces séquences peuvent être communes à plusieurs copies [199].

L'idée de ce parallèle entre RNA-Seq et séquençage génomique d'ET m'a permis d'initer une collaboration pour le développement du pipeline d'analyse DNApipeTE. DNApipeTE permet, à partir de données NGS brutes, d'obtenir une estimation de la proportion des différentes familles d'ET dans un génome ainsi que leurs séquences consensus. Une version préliminaire de l'article présentant cette nouvelle méthode d'annotation d'ET et de quantification d'ET à partir de données NGS est présentée en annexe.

Identification and Analysis of Transposable Elements in Genomic Sequences

Laurent Modolo et Emmanuelle Lerat

Genome Analysis, current procedures and applications
Chapitre 9, *165-181.*

Publié en 2014

Éditeurs : J. Fagerberg, D.C. Mowery et R.R. Nelson

Maison d'édition : Caister Academic Press

Identification and Analysis of Transposable Elements in Genomic Sequences

9

Laurent Modolo and Emmanuelle Lerat

Abstract

Genome sequences are composed of different compartments, among which transposable elements (TEs) represent one of the most important. Not only do these elements correspond to a particularly large proportion of genomes, they are also involved in different mechanisms implicated in the evolution of genomes, such as chromosome rearrangement and gene innovation. Thus, the precise determination of TEs in genomes is of significant importance. This step is becoming more and more complex with the emergence of new types of sequence data coming from next-generation sequencing (NGS) technologies. In this chapter, we present the current status of bioinformatic developments made in the detection and analysis of TEs in genomic sequences. We first present the classic tools dedicated to the identification of TEs in classic genomic data, which originate from whole-genome sequences. Because these sequences are significantly different from the new types of sequences generated by NGS and because the problem of repeats in these data is not trivial, we then present how it is possible to handle TEs in NGS data. We also provide some examples of tools designed to answer particular questions about TEs using NGS data and how these types of data are particularly valuable for deepening our knowledge of the dynamics of TEs. Although this is a still a fast-growing field for which new developments are made every day, we hope to provide a broader view of what currently exists in this field and what allows for TE analyses in genomic sequences.

Introduction

Eukaryotic genomes are composed of different elements, which are classified according to their function in the organism. The protein-coding genes, the non-coding genes (specifically rRNA, tRNA, and small RNA genes), and the regulatory elements associated with these genes are typically considered the most important groups. The remaining elements, which include pseudogenes and repeated sequences, have long been considered trivial in terms of genome functioning. However, this limited vision has begun to change over the past several years. Measured genome size has been demonstrated to be highly variable. However, this metric has been found to have no relationship with the 'complexity' of the organism. When considering the expected number of genes, this paradox has had the implication that a given genome can contain more DNA than required. Early genome sequencing projects helped to answer this question with the discovery that the functional genome, which predominantly consists of protein-coding genes, represents only a small percentage of the genome, whereas the more variable regions of the genome, the 'non-coding regions', were demonstrated to represent a higher proportion of the complete genome. With regard to the human genome, this finding has been particularly striking. The first estimations of the gene number in our own genome were radically revised after the human genome sequencing project, which revealed a small number of genes (fewer than 25,000) and more than 98% 'non-functional' DNA (Lander *et al.*, 2001). The main source of size variation in genomes was then identified as the non-coding regions of genomes.

Transposable elements (TEs) are a part of these still underestimated genomic components, which can play important roles in the functioning and in the evolution of organisms.

TEs are dispersed repeat DNA sequences that have the ability to move from one position to another along chromosomes. These elements typically encode for all the proteins necessary for their movement and possess internal regulatory regions, allowing for their independent expression. Different categories of TEs have been identified, and several attempts were made to classify them (Wicker *et al.*, 2007; Kapitonov and Jurka, 2008). Globally, two main classes have been described according to their transposition intermediates (retrotransposons use an RNA intermediate and form class I, while transposons use a DNA intermediate and form class II). Within each class, subclasses have been created to group sequences with the same structural features (Fig. 9.1). According to the classification system of Wicker *et al.*, five orders containing 17 superfamilies were described as class I elements, and two subclasses, containing 12 superfamilies, were described as class II elements. As the number of the newly sequenced genomes increases, new elements and potential new types of TEs are being discovered, which increases and enriches the complexity of TE classification.

Since B. McClintock discovered and first described these elements in maize in the 1950s (McClintock, 1956), TEs have been searched for and discovered in almost all eukaryotic organisms. Depending on the organism, the proportion of TEs can be highly variable and at times large, for example, 3% in yeast (Kim *et al.*, 1998), 15%

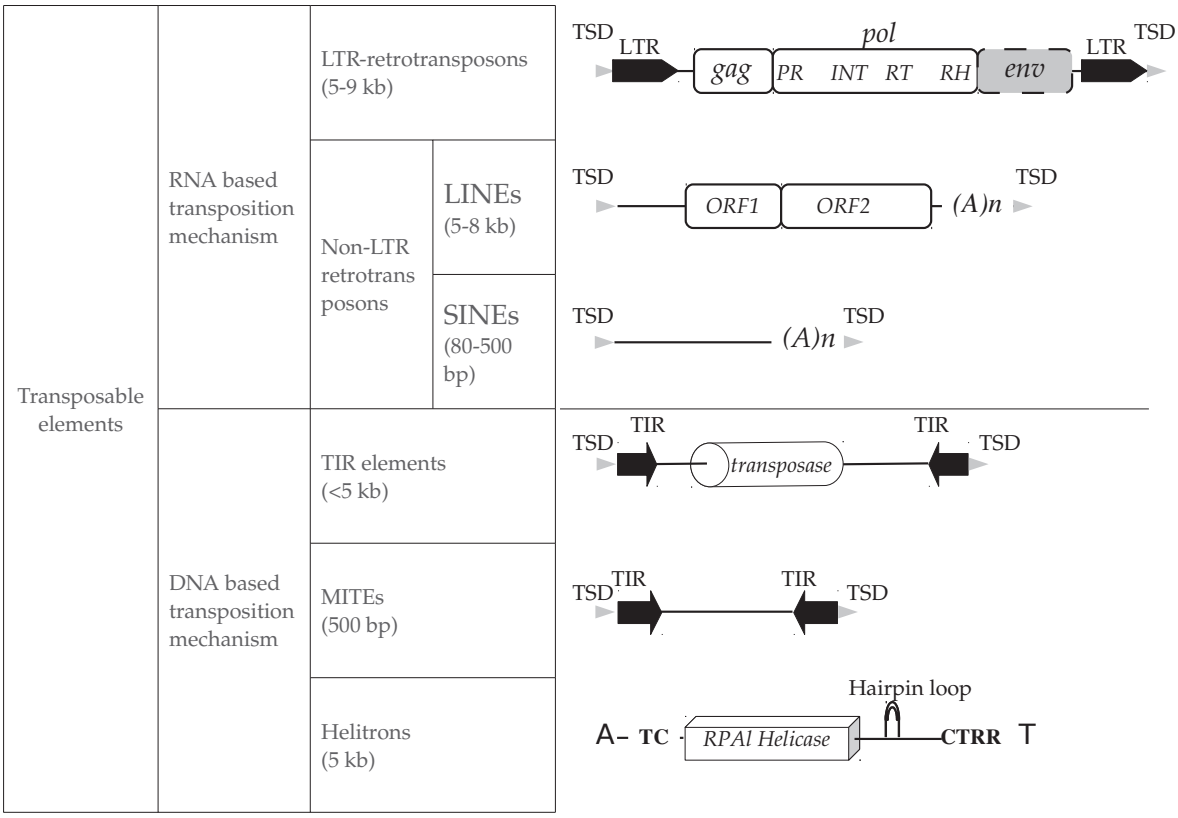


Figure 9.1 The different types of transposable elements (TEs) according to their transposition intermediates and their structural features. Almost all TEs possess two target site duplication (TSD) sequences at each extremity of the copy, corresponding to duplications of the insertion site. LTR-retrotransposons possess long terminal repeat (LTR) sequences at each extremity. Some LTR-retrotransposons encode for a third ORF specifying the ENV protein. The pol gene is formed by different domains, which encode for a protease (PR), an integrase (INT), a reverse transcriptase (RT) and an RNaseH (RH), respectively. The non-LTR retrotransposons contain a poly-A tail at their 3′ extremity. The LINEs possess two ORFs, whereas the SINEs have no coding capacity. Contrary to the MITEs, TIR elements contain an ORF encoding for a transposase. Both contain terminal inverted repeat (TIR) sequences at each extremity. The autonomous helitrons possess a coding capacity for a helicase and an RPA-like (RAPl) single-stranded DNA-binding protein.

in *Drosophila* (Dowsett and Young, 1982), 45% in human (Lander *et al.*, 2001) and more than 80% in maize (Schnable *et al.*, 2009). By proportion, TEs are directly linked to the host's genome size. Their proportion is typically very high in organisms with a very large genome, indicating their role in the expansion of the host's genome size. The variability of TE proportions in closely related species is linked to different parameters, such as the effective population size and reproduction mode of the host organism, the genetic drift, and the host's regulation of transposition activity. For example, in the sibling species *Drosophila melanogaster* and *D. simulans*, the proportion of TEs varies threefold (15% in *D. melanogaster* and 5% in *D. simulans*), although the two species have only been separated for 2.3 to 5.4 million years (Li *et al.*, 1999; Tamura *et al.*, 2004; Cutter, 2008). This discrepancy in TE content has been hypothesized to be linked to either a stronger selection against the deleterious effects of TE insertion in *D. simulans*, which could be due to a larger effective population size compared to *D. melanogaster*, or to a stronger resistance of *D. simulans* to an increase in TE copy number (Kimura and Kidwell, 1994; Vieira *et al.*, 1999). The self-incompatible plant *Arabidopsis lyrata* presents a larger genome (207 Mb) than does the self-compatible plant *A. thaliana* (125 Mb); however, these two species are only separated by 10 million years (Hu *et al.*, 2011). The size variation observed between these two plant species is predominantly due to TEs and appears, in part, to be related to differences in their mating systems (Lockton and Gaut, 2010).

Because of their presence in genomes, TEs have a significant impact on genome evolution and not only on the evolution of genome size (Lisch and Kidwell, 2000; Biémont and Vieira, 2006). TEs can promote mutations, which can be deleterious. For example, in humans, approximately 96 transposition events are directly linked to single-gene diseases (Hancks and Kazazian, 2012), and half of the spontaneous mutations observed in *Drosophila* are due to TEs (Eickbush and Furano, 2002). However, the effects of TEs can also increase the genetic diversity of an organism. The repetitive nature of TEs makes them responsible for chromosomal rearrangements via homologous recombination between copies, which, in some cases, can lead to the emergence of new species, such as has been hypothesized for *D. virilis*, for example (Evgen'ev *et al.*, 2000). When inserting in or near a gene, TEs can provide new regulatory elements, altering the expression of the gene, or they can contribute to gene innovation by providing a new coding region to a gene (Lisch and Kidwell, 2000). The implications of TEs in all epigenetic mechanisms have now been clearly established (Slotkin and Martienssen, 2007). All of these facts make TEs particularly important in the adaptation of organisms to environmental changes. During the years since their discovery, the status of TEs has moved from simple junk DNA to major players in genome evolution (for a historical review, see Biémont, 2010).

Thus, TEs are important components that cannot be neglected when analysing genome sequences. These elements can be quite numerous and, therefore, very important, and they should not be simply removed to ease gene annotations. The study of TEs is crucial for understanding their dynamics, which then allows us to better appreciate how genomes function and evolve. The question of identifying TEs in genomic sequences is a crucial point that has become more and more complex with the emergence of new types of sequencing data. In this chapter, we will first summarize the classic methods that exist to search and annotate TEs in assembled genome sequences. We will then identify the difficulties that can be encountered when using next-generation sequencing (NGS) data for this task and new methods that have been developed. Finally, we will provide examples concerning the specific analyses that can be performed on TEs using NGS data and how these new types of data constitute a significant advance in the field of TE dynamics.

Classic detection methods for TEs in genome sequences

Since the beginning of genome sequencing, significant efforts have been made to annotate the functional regions of genomes. The presence of transposable elements (TEs) and other repeats has made this task particularly difficult. Thus, to facilitate the annotation of genomes, methods to identify TEs and other repeats in genome

sequences were developed (Tang, 2007). Indeed, the ability to recognize these types of sequences has been a suitable starting point to allow for the assembly of a genome and also to ease the prediction of genes. This task is particularly important for genomes containing very high proportions of TEs. Moreover, given the importance of TEs in genome evolution, the identification of these sequences has been considered crucial to allow the access to entire populations of TE copies present in a given organism. Having access to all copies of a given TE family is particularly interesting for studying the evolution and dynamics of a TE family. For example, an analysis of the TE copies from the majority of families integrated into the *D. melanogaster* genome surprisingly demonstrated that the majority of these TEs had recently moved because these TE copies were almost identical within families and very few ancient copies were present (Bowen and McDonald, 2001; Lerat *et al.*, 2003). These observations are in favour of either the hypothesis of recurrent and numerous horizontal transfers of these elements or the hypothesis of a very high turnover in this genome to remove ancient and inactive copies. The identification of all the TE copies present in the human genome has allowed us to obtain information concerning the waves of amplification of the non-LTR retrotransposons LINE and SINE and the formation of retro-processed pseudogenes in this genome (Lander *et al.*, 2001; Ohshima *et al.*, 2003). Numerous methods dedicated to the identification and classification of TEs have been developed over the past 15 years. Several reviews have described these methods exhaustively and provide lists of available programs in each category (Bergman and Quesneville, 2007; Saha *et al.*, 2008a; Lerat, 2010; Janicki *et al.*, 2011). In this section, we will mainly summarize the different categories of existing programs and those, which are currently more used and more successful in performing their tasks.

Similarity- or library-based methods

The principle of these methods is to compare genome sequences to a library of TE reference sequences to search for the occurrence of the TEs in a genome. The library used can be defined by the user or can be a public database. The most widely used public database employed in this type of work is Repbase (Jurka *et al.*, 2005). This database contains the consensus sequences of different repeat sequences from a large set of eukaryotic organisms and is typically employed jointly with the program RepeatMasker (Smit *et al.*, 1996–2010), which performs a similarity search using the library as a reference. The main advantage of this type of method is that it is fast and accurate. Although, it obviously cannot discover new TE families, this method is still a good starting point to explore a new genome, particularly if TE sequences from closely related species are described.

Signature-based methods

This type of method uses particular structural features (such as nucleotide or protein motifs) of known TE classes to determine their occurrence in a genome sequence. Thus, this approach can locate new elements from a given class but will fail to discover new classes of elements. Another drawback of this method is that it will only discover nearly complete and potentially active copies and miss degraded ones. Thus, such an approach can be complemented using a library-based method once complete reference elements have been discovered based on their structure. Moreover, such an approach will depend on the level of knowledge available for a given class and if it is possible to determine fixed and shared features among several families of the same class. Signature-based programs typically concentrate on a particular type of TE.

For example, it is possible to specifically search for LTR-retrotransposons given several shared characteristics between families, such as the presence of an LTR (long terminal repeat) at each end of the sequence, the fact that the two LTRs are almost identical for complete and potentially active copies, a particular distance between them, or the presence of particular protein motifs in the ORFs contained inside the element. Different programs have been designed to detect LTR-retrotransposons, of which the most successful to date is LTRharvest (Ellinghaus *et al.*, 2008; Lerat, 2010). However, the user needs to determine the perfect parameters for the analysed genome to avoid the occurrence of numerous

false positives. Because of structural features such as the presence of a poly-A tail at the 3′ end of the sequence or target site duplications at each extremity of the copy, other programs have been designed to detect non-LTR retrotransposons (Szak *et al.*, 2002; Tu *et al.*, 2004; Lucier *et al.*, 2007). Particular DNA transposons known as MITEs have also been the subject of several programs because of their specific features, *i.e.* a short size (approximately 500 bp) and the presence of terminal inverted repeats (TIRs) at each end of a copy; and because the use of similarity-based methods to find these transposons is difficult due to their short size and a lack of coding capacity. Recently, the program MITE-Hunter (Han and Wessler, 2010) was developed to decrease the number of false positives typically obtained using other programs to locate MITEs.

De novo methods

With these types of approaches, it is possible to search for new types of elements because there is no *a priori* knowledge of the sequence itself. Indeed, these programs take advantage of the repetitive nature of TEs. These methods are particularly interesting when sequencing the genomes of species for which no close relatives are currently annotated and for which nothing is known about their repeat content. However, these methods are particularly sensitive to genomic coverage and to the quality of the sequence assembly. Another drawback is that these approaches will find any type of repeated sequences, even tandem repeats, satellites or segmental duplications, in addition to identifying TE sequences, which implies a classification step for the results to identify TEs. Moreover, TE families containing very few copies will not be detected.

There are two main approaches that are considered *de novo* methods. In the first approach, the genome sequence is first compared against itself to locate all the repeated sequences. Several programs use BLAST (Altschul *et al.*, 1990) to perform this step. The repeated sequences that are located are then grouped into clusters of similar sequence. A consensus sequence is then built for each cluster, and all the consensus sequences are used in a library-based approach to retrieve all occurrences within the genome. Among the most utilized programs that are currently used in the annotation of genomes, we can cite RECON (Bao and Eddy, 2002), PILER (Edgar and Myers, 2005), and BLASTER (Quesneville, unpublished).

In the second approach, the occurrence of multiple small words known as k-mers is searched for within the genome sequence. The k-mer can then be extended to obtain longer sequences. Among the existing programs using this approach, some have been used to discover TEs in genome sequences, for example, REPuter (Kurtz and Schleiermacher, 1999), RepeatScout (Price *et al.*, 2005), and ReAS (Li *et al.*, 2005). This last program has the peculiarity of running not on an assembled genome but on the unassembled reads of a whole genome shotgun sequence to avoid the problems related to a bad assembly.

TEs in the next-generation sequencing data era

The availability of next-generation DNA sequencing (NGS) technologies has revolutionized our approach to genomics (Margulies *et al.*, 2005). These technologies allow us to obtain huge amounts of data at a relatively low cost and with less bias than older technologies (Wicker *et al.*, 2006), thus opening new avenues to the study of TEs. These new types of data also imply that the classic methods described previously will no longer be adapted. To describe how to deal with TEs in NGS data, it is first important to understand why these data are different from older sequencing data and what methodologies are currently available to handle them before performing the TE studies.

With NGS technologies, not only has the volume of data generated dramatically increased, but the range of applications has broadened from methylation pattern detection (MeDIP-Seq) and the study of DNA–protein interactions (ChIP-Seq) to quantifying and detecting gene expression (RNA-Seq). Whatever the application, three steps can always be highlighted in sequencing methodologies. First, the DNA of interest is randomly fragmented and amplified. Second, the ends of each of these fragments are sequenced into reads. Finally, the original sequence of the

DNA is reconstructed from the reads. Currently, the first two steps are highly automated, and the only concern of the researcher is to determine the appropriate sequencing cost to balance the read length and the depth of coverage necessary for the study. Even if the read size has increased with the development of NGS technologies, the reconstruction of the original DNA sequences (assembly) is currently still the most challenging and time-consuming step.

NGS data are subject to particular artefacts that need to be taken into account before performing analyses. These artefacts are predominantly adaptor sequences originating from failed or short DNA insertions during library preparation or near identical reads originating from PCR error. This step of trimming and filtering can be performed using several tools, such as SeqTrim or Quake (Falgueras *et al.*, 2010; Kelley *et al.*, 2010). Another problem is the presence of sequencing errors. Variations between reads can be caused by real sequencing errors or by single-nucleotide polymorphisms (SNPs), accounting for polyploidy and pooled samples. Thus, it is important to perform error corrections that will be linked to the NGS technologies used to generate the data because the different sequencing methods produce different types of sequencing error.

Sequencing and analysing DNA using NGS

Once the reads are sequenced, an important step is to perform their assembly in order to reconstruct the complete genome. This step is complex but can be eased if a reference genome is available. In such a case, it is possible to map the reads directly onto the reference genome. As the number of reads can be very large, classic mapping programs such as the one from the BLAST suite (Altschul *et al.*, 1990) have become too computationally demanding. Thus, a number of alternative approaches have been developed over the past three years to handle NGS data. Two strategies exist for mapping reads onto a genome. The first uses a hash table of the reads, and the other uses a Burrows-Wheeler (BW) transform of the genome (Schbath *et al.*, 2012). Generally, a hash table can better address mismatches, whereas the more complex BW transform approach can easily handle repeats. There are different alternatives to deal with reads occurring at multiple positions, which are known as multi-reads. The program can either ignore them, keep the best matches, keep a specific number of them, or ignore the ones mapping to more than a specific number of locations (Treangen and Salzberg, 2012). Taking into account all of these considerations, it appears that the programs BWA (Li *et al.*, 2010) and Bowtie (Langmead *et al.*, 2009) can outperform the other programs on many criteria, such as computational time, the correct positions of the reads, the number of unmapped reads, and the multi-reads with no more than three mismatches (Schbath *et al.*, 2012). Other parameters that need to be taken into account in addition to mismatches are indels. For this question, BWA is currently the only mapping program using the BW transform that is able to handle indels. Mapping programs using the BW transform appear to be the most appropriate for the study of TEs because they can better handle genomic repeats. Moreover, even if this class of algorithms relies on heuristics to address mismatches, we expect to have a better error correction and less mismatch on TE sequences because they are sequenced with a better coverage than the rest of the genome. For example, with a coverage of 10×, we expect an average coverage of 50× for a given TE that is present with five copies in the genome.

Most of the time there is no reference genome, and the original DNA sequences have to be reconstructed *de novo*. The mapping approach can still be used with the LAST program, which can take into account the divergence between species to relax the mapping parameters and perform *xeno-mapping* (Frith *et al.*, 2010). Otherwise, two different approaches exist for assembling without a reference genome, the first using a seed approach and the other using a de Bruijn graph. In the first approach, the algorithm tries to elongate short sequences of k nucleotides (k-mers) using overlapping reads by computing an overlapping graph where all the paths in the graph consist of overlapping reads. Developed for Sanger technologies, the construction of such a graph is often computationally intractable in the case of NGS data (Pevzner *et al.*, 2001). Since the publication of the EULER program, most assemblers use

the de Bruijn graph approach to assemble reads (Fig. 9.2). The first step of this approach is to build an index of all the possible sequences of size *k* (*k*-mers, often between 24 and 27 bp). The graph itself is built by adding the information from each read, and the sequence of a read is represented by a path between nodes. The nodes correspond to the *k*-mers and their reverse complements to handle more efficiently the two strands of DNA. The addition of each read will correspond to the addition of more edges between the nodes. The construction of a contig is a byproduct of the graph itself. Retrieving the original DNA sequence is a matter of linearizing the information contained in the graph by following the most supported edge, i.e. the one with the highest number of reads (Pevzner *et al.*, 2001; Zerbino and Birney, 2008).

In practice, even with a high coverage, there are always some parts of a DNA sequence that are more difficult to assemble, such as repetitive elements. If the repeat length is larger than the read length, which is often the case for TEs, then the coverage cannot help to reconstruct the original sequence. In this case, a read cannot be associated with one particular copy of the repeated element. Unassembled or uncovered regions are going to form gaps in the assembly, thus decreasing the connectivity between the resulting sequences. Paired-end sequencing technology can be very useful for solving some of the problems caused by repetitive content and short or very short reads (Treangen and Salzberg, 2012). This type of technology is a specific way of sequencing a DNA fragment at both ends. As the size of the fragments is known, the two resulting reads can be reliably positioned relative to each other. To obtain pairs of reads that are more than 500 bp apart (called the insert size), a specific library must be built (a mate-pair or long paired-end library) (Bentley *et al.*, 2008). With paired-end sequencing, a coverage of 10× is sufficient to have at least one mate-pair spanning every instance of each repeat in the genome and to be able to anchor this repeat if the paired read is uniquely mappable (Wetzel *et al.*, 2011). This mate pair information is very useful to position and order the contigs between each other, to fill the gaps, and to build scaffolds (Fig. 9.3). Contrary to the read length, longer inserts will not correspond to a better assembly (Wetzel *et al.*, 2011). It appears that the best strategy is to use different insert sizes to be able to resolve small repeats with short inserts and long repeats with large inserts. The size of the insert can be specifically tuned to optimally assemble the repeat content of the genome under study (Wetzel *et al.*, 2011). For example, it was possible to obtain the same assembly quality found using the Sanger-based approach by using insert sizes of 180 bp, 3 kb, 6 kb and 40 kb for the mouse and human genomes (Gnerre *et al.*, 2011). As with the classic Sanger sequencing methods, the mis-assembly of the repetitive parts of a genome can lead to numerous errors in the reconstructed sequences (Phillippy *et al.*, 2008). For example, a collapse is formed when the assembler incorrectly joins reads originating from distinct repeat copies. On the contrary, expansions are formed

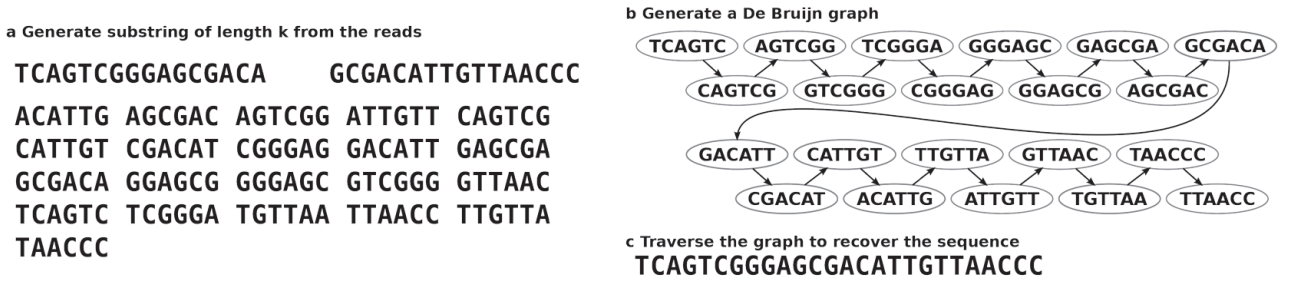


Figure 9.2 Overview of a sequence reconstruction using the de Bruijn graph-based strategy. (a) All substrings of length k (k-mers) were generated from the reads (here k=6). (b) Each unique k-mer represents a node in the de Bruijn graph. The read information is stored in the graph by connecting two nodes with an edge each time that two k-mers have a k – 1 overlap in the reads. Note that generally each node in the graph represents a k-mer and its complement. (c) Nodes from a chain of adjacent nodes link with each other by only one edge and are collapsed into a single node. The graph is then traversed to form contigs.

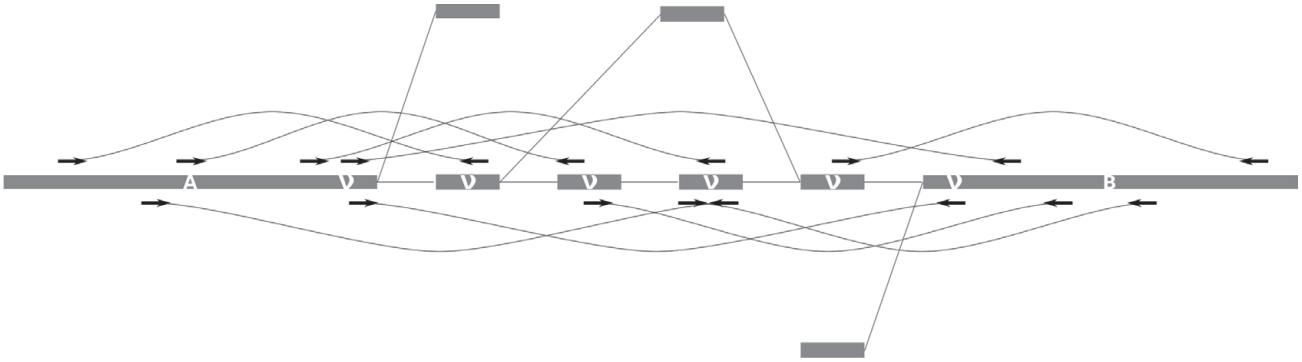


Figure 9.3 Utilization of paired-end information to resolve repeats and to fill gaps. A and B represent long contigs, and the small grey squares are other nodes in the graph. The dashed lines represent all the possible paths built from the read information. The reads are represented by black arrows, and the paired-end information is represented by grey curves linking the two arrowheads to each other. Finding the exact path in the graph from A to B is not a simple task. However, by using the paired-end information linking each pair of reads, we can explore a much simpler graph composed of the 'v' marked blocks.

when reads originating from different copies of a repeat are included in the assembly of one copy of this repeat. These cases are often associated with a greater or lesser density of reads than expected over the repeat(s), and the mate-pairs are stretched out or compressed. Sequence rearrangements are another type of assembly artefact and appear when blocks of DNA are separated by repeats. In this case, the order of these blocks can be wrongly recovered because they are anchored by similar repeats. A special case of rearrangements is inversion, where the two repeat copies are in opposite directions. Sometimes all the above cases of mis-assembly can be possible without violating any constrains on the paired-end information. The amosvalidate package is a collection of tools aimed at detecting mis-assemblies in an automated pipeline (Phillippy *et al.*, 2008).

All the DNA reconstruction problems that we have just described will appear in the de Bruijn graph by forming three types of structures (Zerbino and Birney, 2008). The first corresponds to 'tips', where a chain of nodes is disconnected from the rest of the graph at one end. The second corresponds to 'bubbles', where two paths start from the same node. The third corresponds to erroneous connections, which have no identifiable structure and correspond to chimeric reads. The 'bubble' structures of the graph can correspond to perfect circles, in the case of SNPs or can form densely connected 'tangles' in the case of repetitive sequences (Pevzner *et al.*, 2001). The shape of these 'tangles' can be used for the characterization of the TE, as described in Macas *et al.* (2007). There are many programs that can handle whole genome assemblies using the de Bruijn graph and that are capable of using paired-end information, such as the programs Velvet (Zerbino and Birney, 2008) and ABySS (Simpson *et al.*, 2009). More recently, SOAPdenovo (Li *et al.*, 2010) was used for the *de novo* assembly of the giant panda genome with an average of 20× coverage using 52 bp reads and 37 paired-end libraries with insert sizes ranging from 150 bp to 10 kb. This genome contains 36.1% TEs. The most recent program, ALLPATHS-LG (Gnerre *et al.*, 2011), was created to account for paired-end information with mixed insert size libraries. This program produces a better assembly quality than SOAPdenovo but it is slower (for mammalian-sized genomes, it requires three weeks when SOAPdenovo requires only three days (Gnerre *et al.*, 2011)). Even if the development of assemblers allows for an ever improving assembly of genomes, including their repeat content, there will always remain some unassembled fractions in the data. The study of these leftover contigs or reads can be a source of information about the TE content. These analyses can be conducted using mapping approaches of the data to known TE databases of DNA or protein sequences (Sun *et al.*, 2012).

Sequencing and analysing RNAs using NGS

New sequencing technologies have opened up new vistas of knowledge at the various levels of an organism's biology. Having access to an entire genome is particularly valuable, but having access to the landscape of gene expression allows us to go deeper into the analysis of genomes. Thus, RNA-Seq technology has been developed, which predominantly consists of adding a reverse transcription step to transform RNA into DNA before performing the regular NGS steps (Marioni *et al.*, 2008). The goal of this approach is to obtain the sequence and the abundance of all the transcripts of a given organism in a given condition. Using these types of sequencing methods, it is important to handle particular artefacts. In addition to the errors generated by the reverse transcription step, the reconstruction of the RNA sequences accounts for the different levels of expression for the different transcripts and for the mechanism of alternative splicing (Martin and Wang, 2011). Because of the presence of shared exons in different transcripts, expression counting or quantification is not trivial for RNA-Seq data (Trapnell *et al.*, 2010). Moreover, there is a competition between the most abundant transcripts, which can be over-sequenced, and the less abundant ones, which can be missed. To sequence and quantify less abundant transcripts, hybridization-based depletion methods to remove the more abundant transcripts can be utilized (He *et al.*, 2010). However, these depletion methods induce biases for the quantification and for the assembly of the most expressed transcripts.

Once the main artefacts are corrected, there are two strategies to analyse the RNA-Seq data. The first approach, known as 'map first', consists of mapping the reads onto a reference genome. This approach can be confronted with many problems, ranging from the correctness of the alignment, the possibility of losing the splicing information and the completeness of the reference genome. There are many programs for mapping RNA-Seq data to a reference genome, but it appears that the most commonly utilized is a combination of the program TopHat, which is able to discover splice junctions, and the program Cufflinks, which can be used for the quantification of transcripts (Trapnell *et al.*, 2010). The second approach, 'assembly first' (*de novo* method), consists of directly assembling reads to reconstruct the transcripts. Using this approach, the resultant transcripts can subsequently be mapped to a reference genome. A number of transcriptome assembly programs have been developed, and most use the de Bruijn graph approach. This data structure naturally handles the high redundancy of the data because each repeat or transcript is only present once in the graph. The most commonly used transcriptome assemblers are Velvet (Zerbino and Birney, 2008) and Trans-ABySS (Robertson *et al.*, 2010). More recently, the Trinity program was proposed (Grabherr *et al.*, 2011). Other approaches, such as KISSPLICE (Sacomoto *et al.*, 2012), have been developed to identify and quantify *de novo* polymorphisms such as alternative splicing, SNP and tandem repeats in RNA-Seq data.

Even with a reference genome, the best results are obtained using mixed strategies of the two approaches (Surget-Groba and Montoya-Burgos, 2010). Based on the confidence we have in the reference genome, there are two possibilities. When the reference genome is of a very good quality, it is possible to first align the reads onto it and then to assemble the reads using the mapped reads as long contigs. The assembling step allows for resolving of the reads coming from the expressed regions not present in the reference genome. In a case where the reference genome is not of a very good quality, another strategy is used, consisting of first making the assembly of reads before mapping them onto the reference genome. In this case, the errors present in the reference genome have little impact because they are not present in the assembled contigs. The mapping step is then used to resolve scaffolds from the more fragmented contigs obtained using the *de novo* approaches. This last approach was used to successfully assemble the transcriptome of the mosquito *Anopheles funestus* (Crawford *et al.*, 2010)

What do NGS data bring to TE analyses?

With the new technologies of DNA and RNA sequencing, new opportunities to study TEs have

appeared, once the difficulty for handling these repeated sequences taken into account when processing the data. In this section, we will exemplify different analyses that have allowed us to improve on our understanding of TEs.

TE identification in a genome survey

Even if the cost of NGS sequencing has continuously diminished it is still of interest to perform genomic surveys to characterize large genomes. The historic approach of using the end sequences of bacterial artificial chromosome (BAC) vectors is biased towards sequences that can be successfully loaded into BACs. NGS technologies allow us to perform more representative surveys of large genomes by sequencing at a low depth using a whole genome shotgun (WGS) approach. This approach was first used to further characterize the soybean genome, with particular attention paid to its repeat content (Swaminathan *et al.*, 2007). The first step of the data analysis is to characterize a maximum number of reads using databases of annotated sequences. For this step, classical programs such as BlastX and BlastN may be used. For example, in the survey analysis of the barley genome using 454 sequencing at 0.1× coverage, it was possible to determine 7.4% of the barley genes using BlastX on a database of predicted rice proteins and to characterize the presence of many TEs using BlastN on a TE plant database (Wicker *et al.*, 2009).

Another characteristic resulting from genome sampling with WGS and NGS technologies is to expect an increased coverage of the repetitive content. For example, with a genomic coverage of 0.01×, we can expect a coverage of 10× for each repeat occurring in the genome with 1000 copies (Macas *et al.*, 2007). With this characteristic, the identification of repeat content using a NGS survey is not limited to a homology search. The first program developed to use the expected increase in coverage for TE sequences is ReAS, which has already been referenced in section 1 of this chapter and allows us to assemble consensus TE sequences (Li *et al.*, 2005). The TEs must exist at a sufficient copy number to be recognized by their read number and must not be too degraded to have sufficient sequence similarities between their copies to be able to build a consensus sequence. The ReAS program starts by building a k-mer index of the reads. The high copy number k-mers are then picked out to retrieve the reads containing them. The reads are then assembled and expanded to recover the consensus sequences of the different elements. When possible, the contigs formed are linked using the paired-end information. The main drawback of this method is the fact that it is not designed for short or very short reads and cannot process reads less than 104 bp (Macas *et al.*, 2007). The AAARF (Assisted Automated Assembler of Repeat Families) program was designed to overcome this problem and can process short or very short reads (DeBarry *et al.*, 2008). This program uses one read as a query sequence to obtain its nucleotide coverage against the rest of the dataset using BlastN. This nucleotide coverage is then used to select the overlapping reads, which are then aligned using ClustalW to build a new query sequence. This program iteratively elongates each query sequence and assembles a set of TE contigs.

To recover repetitive sequences, it is also possible to cluster overlapping unannotated DNA sequences. This method was used to assemble 41% of the reads of the soybean genome into contigs using the Phrap program (Swaminathan *et al.*, 2007). A similar approach was used to assemble 31.6% of the unannotated reads into contigs for the barley genome survey (Wicker *et al.*, 2009). However, using this approach and because of the low depth of the data, some links are absent in the overlapping graphs, leading to contig fragmentation (Novák *et al.*, 2010). A slightly different approach was used to reconstruct the repeat sequences from a survey of the pea genome (Macas *et al.*, 2007). In this study, the program tclust from the TGICL package was used to cluster the reads based on a mutual similarity and to assemble each cluster into contigs. With this method, each cluster contains related repeat sequences, which can be used to better characterize the variability in the TE content of a genome. This method was successfully used to characterize the repeat content of the banana genome (Hribová *et al.*, 2010). One drawback of this more sensible approach is the formation of chimeric clusters, which are caused by the presence of reads spanning two TE sequences and form bridges

between two clusters (Macas *et al.*, 2007). More recently, a novel cluster-based approach was proposed with the program SeqGraphR (Novák *et al.*, 2010). This program uses a hierarchical agglomeration algorithm to cluster the reads and to characterize the TE sequences. This graph–based clustering allows for a better segregation of groups of unrelated sequences than tclust. Moreover, it allows for a better characterization of the cluster structure by computing various graph metrics to discriminate between different types of repeats. The assembly of TE sequences from a cluster results in the formation of consensus sequences. With this type of sequence representation, a significant amount of information about TE sequence variability is lost. SeqGraphR can provide an alternative representation of TEs by direct graph visualization. This approach can be very useful for deciphering contig assembly or for distinguishing between two closely related TEs. SeqGraphR was successfully used to characterize the TE content of three species of *Nicotiana tabacum* in a genome survey using 454 sequencing at 0.1× (Renny-Byfield *et al.*, 2011).

All previous methods work for genome survey data with relatively small dataset sizes, ranging from 33 Mb to 90 Mb, but may not be suitable for larger datasets. Moreover, approaches using read overlap information will only work for sparse genome survey data, where the only sequences that can be assembled come from repeat sequences. If the coverage reaches or exceeds 1×, most of the genome can be assembled, and the described approaches will lose their specificity for TE sequences. However, a genome survey dataset can always be generated from a deep genome sequencing output by randomly selecting a subset of reads to easily study the most abundant repeats. LTR retrotransposons can also be *de novo* identified using mapping approaches. For this class of TEs, the reads will pileup on the two LTRs, and a 'batman ears'-like structure will appear when using programs such as Tallymer or JellyFish if we allow for multi-read mapping (Kurtz *et al.*, 2008; Marçais and Kingsford, 2011).

Structural variant detection

One of the most interesting characteristics of TEs is their ability to replicate and to colonize a genome. This transposition activity can be studied between species or populations. NGS technologies allow us to study copy number variation (CNV) by sequencing pooled DNA from different individuals (or pool-Seq). The main advantages of these approaches are that the approaches are fast and not copy specific and offer a higher sensitivity compared to other technologies (Alkan *et al.*, 2011). To perform these types of studies, one needs an assembled reference genome, a database of the TE sequences (which can be built from the sequencing data) and, optionally, paired-end technology to better resolve the TE information.

For this purpose, the T-lex program was developed to compute the population frequencies of individual TE insertions (Fiston-Lavier *et al.*, 2011). This program is a pipeline using four modules. The first module uses RepeatMasker to identify TEs and their flanking regions in the reference genome. The second module uses MAQ (Li *et al.*, 2008) to determine the presence of TEs by mapping reads across the sequences formed by an identified TE and its flanking regions in the reference genome. The third module uses SHRiMP (Rumble *et al.*, 2009), which can align sequences with long gaps, to identify the absence of a TE insertion in the analysed populations by mapping reads spanning only the two flanking regions of the TE sequence in the reference genome. Finally, the last module combines the information of the previous modules to obtain the frequencies of each TE family in the populations. The second version of the T-lex program is able to automatically use paired-end information to detect novel TE insertions. By using a similar approach to T-lex, it was possible to successfully analyse the activity of TEs using pooled DNA samples from 114 isofemale lines of *D. melanogaster* (Kofler *et al.*, 2012). For this study, 80 million paired-end fragments were produced with the Illumina Genome Analyser IIx. These reads were mapped onto the reference genome, where all the repeats were first masked using RepeatMasker. The mapping step was then performed using the BWA-SW program (Li *et al.*, 2010). The authors were then able to identify novel insertions with at least three paired couples of reads, with, for each couple, one read mapping to a genomic locus and the others mapping to a TE sequence. In this study, novel TE insertions

were detected if they were present in at least 7% of the populations.

Without a reference genome, comparative studies of the TE content of different species can also be achieved using genome surveys. For example, an analysis of genomic gigantism in plethodontid salamanders was performed using a genome survey of six species sequenced with a coverage of 0.1× using 454 technology to obtain reads of a maximum length of 400 bp (Sun *et al.*, 2012). In this study, the RepeatModeler program (Smit and Hubley 2008–2010) was used to identify those repeats covered by a minimum of four reads. The REPCLASS program (Feschotte *et al.*, 2009) was then used to further classify the unknown repeats. The results demonstrated that the analysed salamander species accumulate large amounts of LTR-retrotransposons compared to other vertebrates.

Analysis of TE regulation by the host

With the prevalence of TEs and their capacity to invade a genome, it is crucial for the host to be able to regulate their activity to avoid too many deleterious effects. During the past few years, the links between TEs and the epigenetic systems of regulation, such as DNA methylation, histone modifications, and RNA interference, have been shown to be linked to the repeat content of a genome (Siomi and Siomi, 2008; Rebollo *et al.*, 2010). In particular, diverse RNA-mediated defences have been discovered in different eukaryotic organisms (Slotkin and Martienssen, 2007; Blumenstiel, 2011). These discoveries have allowed for the development of new models for TE dynamics in natural populations in which four phases have been described: an initial phase of TE invasion; a second phase of TE proliferation, leading to the appearance of TE insertion alleles initiating the production of small RNAs; and finally a quiescent state, leading to the stabilization of TE copy number (see (Blumenstiel, 2011) for a review).

With the development of NGS technologies, it has become easier to analyse the epigenetic control of TEs. Particular modifications can regulate TE activity, such as DNA methylation. This type of modification can be determined using a MeDIP-seq approach and has already been used in several organisms. For example, in black cottonwood, the TEs possessed variable methylation according to their family, with LTR retrotransposons being globally more methylated than other classes (Vining *et al.*, 2012).

RNA-Seq is also a reliable way to study active TEs because the complete RNA sequences of the TEs can be found in the data output. The control of TEs can also be studied when considering small RNAs such as piRNAs or siRNAs, which are expected to be copy specific to the TEs that they control. This type of analysis can also be used to validate *de novo* annotations of TEs. The results of these approaches are naturally linked with the condition or the stage where the different TEs are expressed. Due to piRNA regulation, the quantification aspect of RNA-Seq is of a lesser interest for the study of TEs because the number of transcripts is not directly correlated with the activity of a TE (Brennecke *et al.*, 2007). However, these studies allow us to obtain information on the potentially complete and active copies that are inserted into the genome. The analysis of small RNAs is typically performed by mapping the reads on to a reference genome to help identify clusters of small RNAs and to determine which copy in the genome is associated with a particular small RNA. This approach has been used in the analysis of the control of particular TEs in *D. melanogaster* (Brennecke *et al.*, 2007; Brennecke *et al.*, 2008; Grentzinger *et al.*, 2012) and in plant species (Hollister *et al.*, 2011). Of course, these genome-mapping approaches have limitations due to the differences that exist between individuals. For example, in the case of an analysis of *P*-elements, it was not possible to map them onto the *D. melanogaster* reference genome because this particular element is absent in the sequenced strain (Brennecke *et al.*, 2008). Thus, it was necessary to find another strategy, in this case, using those reads that did not map onto the reference genome.

The regulation of TEs can also be linked to the histone modifications of the DNA. Using ChIP-seq data, it is possible to determine what types of modifications are associated with TEs or with genes given their TE neighbourhood. For example, in mouse embryonic stem cells, an analysis of ChIP-seq data revealed that the

majority of gene promoters surrounded by numerous TEs were depleted of the bivalent marks H3K27me3 + H3K4me3 compared to genes surrounded by few or no TEs. This bivalent mark has been demonstrated to be specific to a 'poised state' of developmental genes that are temporarily repressed in embryonic stem cells but that will be activated later during development (Zhang and Mager, 2012). This previous analysis indirectly observed modifications associated with TEs. Generally, during the mapping step of reads sequenced using ChIP-seq analysis, the reads cannot be associated with TE copies because only uniquely mappable reads are conserved (those having a unique location on the genome). Thus, an alternative mapping approach was proposed by Huda *et al.* to allow for the direct examination of ChIP-seq reads associated with TEs (Huda *et al.* 2010). In that analysis, the authors took advantage of several ChIP-seq experiments, which allowed them access to a genome-wide map of 38 histone modifications in human $CD4^+$ T cells (Barski *et al.*, 2007; Wang *et al.*, 2008). They used the MAQ program to align the reads, allowing for redundant genomic locations. It was then possible to characterize what type of TE was present in the mapped reads. Their results demonstrated a high variation in TE histone modifications according to the TE family, with the older TE families and the TEs close to genes carrying more modifications than the younger TE families and those TEs distal to genes.

These different examples demonstrate how valuable NGS data are in the analysis of TEs, and also identify the need for specific tools to handle these data for studying TEs.

Conclusions

Transposable elements are important components of genomes that cannot simply be put aside when analysing genomes. It is important to understand how TEs function and evolve to better understand all of the impacts of TEs on genome evolution. Given the large amount of genomic data that has been continuously generated, this task becomes more and more difficult. However, these data allow us an access to new information that we did not have several years ago.

Since the first sequencing projects were undertaken, efforts have been made to develop programs that allow for the detection and analysis of TE sequences. Various programs have been developed that use different or complementary approaches to detect TE sequences. These tools have very different performances, but even the best ones cannot discover all the TE sequences in a genome because each has its own drawback(s) that prevents it from finding each and every TE (Saha *et al.*, 2008b; Lerat, 2010). Thus, the best approach to exhaustively describe the landscape of TEs in a genome is to use several of these different programs and to cross-reference the results. Similarly, the best approach to locate TE sequences in complete genomes appears to reside in the use of pipelines of programs. For example, the RepeatModeler pipeline includes different programs to build, refine and classify consensus sequences of putative interspersed repeats (Smit and Hubley, 2008–2010). The REPET pipeline has been built to integrate the findings of similarity- and *de novo*-based programs (Quesneville *et al.*, 2005). This pipeline was recently updated to retain those programs that provide the best results after the authors tested different *de novo* programs (Flutre *et al.*, 2011). Other pipelines have generally been developed to answer specific questions (see Lerat, 2010).

In all cases, another important step after the identification of putative TE sequences is the classification of the repeats into families. This is a difficult step because it must take into account the biological aspect of TEs, such as the fact that some copies can be fragmented and thus not only full-length elements exist in a genome and that TEs often insert inside each other, producing what are known as nested TEs. Some programs have been developed to integrate the classification step, such as the TEclass program, which tries to determine the main classes of unknown elements using machine-learning algorithms (Abrusán *et al.*, 2009), or the REPCLASS program, which uses different approaches to annotate TEs (Feschotte *et al.*, 2009).

With the new type of genomics data generated by next-generation sequencing technologies, it is necessary to develop new approaches to detect and analyse TEs. Indeed, most of the programs

that have been designed for classic genomic data cannot handle these new types of data. This is predominantly because NGSs produce small sequence fragments, which increase the difficulties involved in assembling the repeat content of a genome, but also because the amounts of these data are too large to be handled with the existing tools. However, new programs have been designed to take these problems into account. Even if all these programs are not specific to TEs, specific programs are now available to answer particular questions on TEs with regards to population genomics, and we can hope for new developments in more specific areas. Questions about the epigenetic regulation of TEs are particularly important at different levels, such as the impact of TEs in cancer development in humans. NGS data now offer the possibility of having access to this information, which will necessitate the development of particular tools specific for TEs.

Future trends

With the development of NGS data, access to individuals' sequence data should provide us with valuable information on the specific content of TEs, allowing us to be more precise in terms of the insertion profiles of TEs. Currently, the only available possibility is to compare data to reference genomes. However, this shortcut is not precise enough when delving deeper into our understanding of the mechanisms of TE dynamics.

Web resources

A list of existing TE detection tools is available at:

- http://bergmanlab.smith.man.ac.uk/?page_id=295
- Quesneville. *BLASTER suite* <http://urgi.versailles.inra.fr/Tools/Blaster>.
- Smit, AFA, Hubley, R and Green, P. *RepeatMasker Open-3.0*. 1996–2010 <http://www.repeatmasker.org>.
- Smit, AFA, Hubley, R. *RepeatModeler Open-1.0*. 2008–2010 <http://www.repeatmasker.org>.

References

Abrusán, G., Grundmann, N., DeMester, L., and Makalowski, W. (2009). Teclass – a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics *25*, 1329–1330.

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. *12*, 363–376.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Bao, Z., and Eddy, S.R. (2002). Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res. *12*, 1269–1276.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59.

Bergman, C.M., and Quesneville, H. (2007). Discovering and detecting transposable elements in genome sequences. Brief. Bioinform. *8*, 382–392.

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

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

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

Bowen, N.J., and McDonald, J.F. (2001). Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res. *11*, 1527–1540.

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

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

Crawford, J.E., Guelbeogo, W.M., Sanou, A., Traoré, A., Vernick, K.D., Sagnon, N., and Lazzaro, B.P. (2010). *De novo* transcriptome sequencing in *Anopheles funestus* using Illumina RNA-seq technology. PLoS One *5*, e14202.

Cutter, A.D. (2008). Divergence times in Caenorhabditis and Drosophila inferred from direct estimates of the neutral mutation rate. Mol. Biol. Evol. *25*, 778–786.

DeBarry, J.D., Liu, R., and Bennetzen, J.L. (2008). Discovery and assembly of repeat family pseudomolecules from sparse genomic sequence data using the assisted automated assembler of repeat families (AAARF) algorithm. BMC Bioinformatics *9*, 235.

Dowsett, A.P., and Young, M.W. (1982). Differing levels of dispersed repetitive DNA among closely related species of Drosophila. Proc. Natl. Acad. Sci. U.S.A. *79*, 4570–4574.

Edgar, R.C., and Myers, E.W. (2005). PILER: identification and classification of genomic repeats. Bioinformatics *21* (Suppl 1), i152–i158.

Eickbush, T.H., and Furano, A.V. (2002). Fruit flies and humans respond differently to retrotransposons. Curr. Opin. Genet. Dev. *12*, 669–674.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics *9*, 18.

Evgen'ev, M.B., Zelentsova, H., Poluectova, H., Lyozin, G.T., Veleikodvorskaja, V., Pyatkov, K.I., Zhivotovsky, L.A., and Kidwell, M.G. (2000). Mobile elements and chromosomal evolution in the virilis group of Drosophila. Proc. Natl. Acad. Sci. U.S.A. *97*, 11337–11342.

Falgueras, J., Lara, A.J., Fernández-Pozo, N., Cantón, F.R., Pérez-Trabado, G., and Claros, M.G. (2010). SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. BMC Bioinformatics *11*, 38.

Feschotte, C., Keswani, U., Ranganathan, N., Guibotsy, M.L., and Levine, D. (2009). Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome Biol. Evol. *1*, 205–220.

Fiston-Lavier, A.-S., Carrigan, M., Petrov, D.A., and González, J. (2011). T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. Nucleic Acids Res. *39*, e36.

Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification in *de novo* annotation approaches. PLoS One *6*, e16526.

Frith, M.C., Wan, R., and Horton, P. (2010). Incorporating sequence quality data into alignment improves DNA read mapping. Nucleic Acids Res. *38*, e100.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., *et al.* (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. U.S.A. *108*, 1513–1518.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., *et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Grentzinger, T., Armenise, C., Brun, C., Mugat, B., Serrano, V., Pelisson, A., and Chambeyron, S. (2012). piRNA-mediated transgenerational inheritance of an acquired trait. Genome Res. *22*, 1877–1888.

Han, Y., and Wessler, S.R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. *38*, e199.

Hancks, D.C., and Kazazian, H.H. (2012). Active human retrotransposons: variation and disease. Curr. Opin. Genet. Dev. *22*, 191–203.

He, S., Wurtzel, O., Singh, K., Froula, J.L., Yilmaz, S., Tringe, S.G., Wang, Z., Chen, F., Lindquist, E.A., Sorek, R., *et al.* (2010). Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. Nat. Methods *7*, 807–812.

Hollister, J.D., Smith, L.M., Guo, Y.-L., Ott, F., Weigel, D., and Gaut, B.S. (2011). Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Proc. Natl. Acad. Sci. U.S.A. *108*, 2322–2327.

Hribová, E., Neumann, P., Matsumoto, T., Roux, N., Macas, J., and Dolezel, J. (2010). Repetitive part of the banana (*Musa acuminata*) genome investigated by low-depth 454 sequencing. BMC Plant Biol. *10*, 204.

Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.-F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H., *et al.* (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat. Genet. *43*, 476–481.

Huda, A., Mariño-Ramírez, L., and Jordan, I.K. (2010). Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. Mob. DNA *1*, 2.

Janicki, M., Rooke, R., and Yang, G. (2011). Bioinformatics and genomic analysis of transposable elements in eukaryotic genomes. Chromosome Res. *19*, 787–808.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. *110*, 462–467.

Kapitonov, V.V., and Jurka, J. (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. Nat. Rev. Genet. *9*, 411–412.

Kelley, D.R., Schatz, M.C., and Salzberg, S.L. (2010). Quake: quality-aware detection and correction of sequencing errors. Genome Biol. *11*, R116.

Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A., and Voytas, D.F. (1998). Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Res. *8*, 464–478.

Kimura, K., and Kidwell, M.G. (1994). Differences in *P* element population dynamics between the sibling species *Drosophila melanogaster* and *Drosophila simulans*. Genet. Res. *63*, 27–38.

Kofler, R., Betancourt, A.J., and Schlötterer, C. (2012). Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. PLoS Genet. *8*, e1002487.

Kurtz, S., and Schleiermacher, C. (1999). REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics *15*, 426–427.

Kurtz, S., Narechania, A., Stein, J.C., and Ware, D. (2008). A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics *9*, 517.

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

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lerat, E. (2010). Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity *104*, 520–533.

Lerat, E., Rizzon, C., and Biémont, C. (2003). Sequence divergence within transposable element families in the *Drosophila melanogaster* genome. Genome Res. *13*, 1889–1896.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. *18*, 1851–1858.

Li, R., Ye, J., Li, S., Wang, J., Han, Y., Ye, C., Wang, J., Yang, H., Yu, J., Wong, G.K.-S., *et al.* (2005). ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. PLoS Comput. Biol. *1*, e43.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., *et al.* (2010). The sequence and *de novo* assembly of the giant panda genome. Nature *463*, 311–317.

Li, Y.J., Satta, Y., and Takahata, N. (1999). Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method. Genes Genet. Syst. *74*, 117–127.

Lisch, D.R., and Kidwell, M.G. (2000). Transposable elements and host genome evolution. Trends Ecol. Evol. *15*, 95–99.

Lockton, S., and Gaut, B.S. (2010). The evolution of transposable elements in natural populations of self-fertilizing *Arabidopsis thaliana* and its outcrossing relative *Arabidopsis lyrata*. BMC Evol. Biol. *10*, 10.

Lucier, J.-F., Perreault, J., Noël, J.-F., Boire, G., and Perreault, J.-P. (2007). RTAnalyzer: a web application for finding new retrotransposons and detecting *L1* retrotransposition signatures. Nucleic Acids Res. *35*, W269–W274.

Macas, J., Neumann, P., and Navrátilová, A. (2007). Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. BMC Genomics *8*, 427.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. *18*, 1509–1517.

Martin, J.A., and Wang, Z. (2011). Next-generation transcriptome assembly. Nat. Rev. Genet. *12*, 671–682.

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764–770.

McClintock, B. (1956). Controlling elements and the gene. Cold Spring Harb. Symp. Quant. Biol. *21*, 197–216.

Novák, P., Neumann, P., and Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics *11*, 378.

Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y., and Okada, N. (2003). Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. Genome Biol. *4*, R74.

Pevzner, P.A., Tang, H., and Waterman, M.S. (2001). An Eulerian path approach to DNA fragment assembly. Proc. Natl. Acad. Sci. U.S.A. *98*, 9748–9753.

Phillippy, A.M., Schatz, M.C., and Pop, M. (2008). Genome assembly forensics: finding the elusive mis-assembly. Genome Biol. *9*, R55.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). *De novo* identification of repeat families in large genomes. Bioinformatics *21* (Suppl 1), i351–i358.

Quesneville, H., Bergman, C.M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., and Anxolabehere, D. (2005). Combined evidence annotation of transposable elements in genome sequences. PLoS Comput. Biol. *1*, 166–175.

Rebollo, R., Horard, B., Hubert, B., and Vieira, C. (2010). Jumping genes and epigenetics: towards new species. Gene *454*, 1–7.

Renny-Byfield, S., Chester, M., Kovařík, A., Comber, S.C.L., Grandbastien, M.-A., Deloger, M., Nichols, R.A., Macas, J., Novák, P., Chase, M.W., *et al.* (2011). Next generation sequencing reveals genome downsizing in allotetraploid *Nicotiana tabacum*, predominantly through the elimination of paternally derived repetitive DNAs. Mol. Biol. Evol. *28*, 2843–2854.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q., *et al.* (2010). De novo assembly and analysis of RNA-seq data. Nat. Methods *7*, 909–912.

Rumble, S.M., Lacroute, P., Dalca, A.V., Fiume, M., Sidow, A., and Brudno, M. (2009). SHRiMP: accurate mapping of short color-space reads. PLoS Comput. Biol. *5*, e1000386.

Sacomoto, G.A.T., Kielbassa, J., Chikhi, R., Uricaru, R., Antoniou, P., Sagot, M.-F., Peterlongo, P., and Lacroix, V. (2012). KISSPLICE: *de-novo* calling alternative splicing events from RNA-seq data. BMC Bioinformatics *13* (Suppl 6), S5.

Saha, S., Bridges, S., Magbanua, Z., and Peterson, D. (2008a) Computational approaches and tools used in identification of dispersed repetitive DNA sequences. Trop. Plant Biol. *1*, 85–96.

Saha, S., Bridges, S., Magbanua, Z.V., and Peterson, D.G. (2008b) Empirical comparison of *ab initio* repeat finding programs. Nucleic Acids Res. *36*, 2284–2294.

Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V., and Gibrat, J.-F. (2012). Mapping reads on a genomic sequence: an algorithmic overwiew and a practical comparative analysis. J. Comput. Biol. *19*, 796–813.

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

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

Siomi, H., and Siomi, M.C. (2008). Interactions between transposable elements and Argonautes have (probably) been shaping the Drosophila genome throughout evolution. Curr. Opin. Genet. Dev. *18*, 181–187.

Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. *8*, 272–285.

Sun, C., Shepard, D.B., Chong, R.A., López Arriaza, J., Hall, K., Castoe, T.A., Feschotte, C., Pollock, D.D., and Mueller, R.L. (2012). LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. Genome Biol. Evol. *4*, 168–183.

Surget-Groba, Y., and Montoya-Burgos, J.I. (2010). Optimization of *de novo* transcriptome assembly from next-generation sequencing data. Genome Res. *20*, 1432–1440.

Swaminathan, K., Varala, K., and Hudson, M.E. (2007). Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. BMC Genomics *8*, 132.

Szak, S.T., Pickeral, O.K., Makalowski, W., Boguski, M.S., Landsman, D., and Boeke, J.D. (2002). Molecular archeology of L1 insertions in the human genome. Genome Biol. *3*, research0052.

Tamura, K., Subramanian, S., and Kumar, S. (2004). Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. Mol. Biol. Evol. *21*, 36–44.

Tang, H. (2007). Genome assembly, rearrangement, and repeats. Chem. Rev. *107*, 3391–3406.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Treangen, T.J., and Salzberg, S.L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. *13*, 36–46.

Tu, Z., Li, S., and Mao, C. (2004). The changing tails of a novel short interspersed element in *Aedes aegypti*: genomic evidence for slippage retrotransposition and the relationship between 3′ tandem repeats and the poly(dA) tail. Genetics *168*, 2037–2047.

Vieira, C., Lepetit, D., Dumont, S., and Biémont, C. (1999). Wake up of transposable elements following *Drosophila simulans* worldwide colonization. Mol. Biol. Evol. *16*, 1251–1255.

Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M.Q., *et al.* (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. Nat. Genet. *40*, 897–903.

Wetzel, J., Kingsford, C., and Pop, M. (2011). Assessing the benefits of using mate-pairs to resolve repeats in *de novo* short-read prokaryotic assemblies. BMC Bioinformatics *12*, 95.

Wicker, T., Schlagenhauf, E., Graner, A., Close, T.J., Keller, B., and Stein, N. (2006). 454 sequencing put to the test using the complex genome of barley. BMC Genomics *7*, 275.

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

Wicker, T., Taudien, S., Houben, A., Keller, B., Graner, A., Platzer, M., and Stein, N. (2009). A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. Plant J. *59*, 712–722.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829.

Zhang, Y., and Mager, D.L. (2012). Gene properties and chromatin state influence the accumulation of transposable elements in genes. PLoS One *7*, e30158.

2

Méthodes de détection de transferts horizontaux, et application aux éléments transposables

1 Le modèle biologique

Drosophila melanogaster fait partie des organismes modèles en biologie et de nombreux travaux sur les éléments transposables (ET) ont été effectués chez cette espèce. Cette drosophile qui affiche maintenant une répartition mondiale est originaire d'Afrique. C'est chez *D. melanogaster* que l'on a découvert le cas historique du transfert horizontal (TH) de l'élément *P*, un transposon à ADN. Cet élément avait été détecté dans des populations naturelles, alors que les lignées de laboratoire échantillonnées 40 ans plus tôt en étaient dépourvues. Ceci a permis de mettre en évidence un transfert horizontal récent de cet élément chez *D. melanogaster* provenant d'une espèce distante de drosophile : *D. willistoni* [200]. C'est aussi chez *D. melanogaster* que la voie de régulation des ET par les piRNA a été caractérisée [201, 115, 202, 116].

Cependant, ce modèle peut être considéré comme une exception parmi les espèces du sous-groupe melanogaster pour ce qui concerne les ET. En effet, *D. melanogaster* est la seule espèce de ce sous-groupe dont l'analyse a révélé un grand nombre de familles d'ET jeunes et actives [162, 203]. L'état des ET dans cette espèce correspond aux premières

étapes du cycle de naissance et de mort décrit dans l'introduction de cette thèse. À l'inverse, chez les autres espèces du sous-groupe melanogaster qui ont été analysées, la majorité des copies d'ET correspond à des insertions anciennes et dégradées, avec peu de copies actives [203].

Ces différences ont été particulièrement étudiées entre *D. simulans* et *D. melanogaster*. *D. simulans* est une autre espèce de drosophile cosmopolite, dont la spéciation avec *D. melanogaster* remonte à 2 à 3 millions d'années (Ma) [204]. Même si ces deux espèces sont très proches phylogénétiquement l'une de l'autre, elles présentent des différences importantes en terme de contenu en ET. Les ET représentent 6,85% du génome de *D. simulans* alors qu'ils constituent près de 16% de celui de *D. melanogaster* [162, 205]. Cette différence peut être expliquée par l'état des copies d'ET dans ces deux génomes. Chez *D. simulans*, la plupart des copies sont anciennes et fortement dégradées, ce qui conduit à une taille plus réduite de leurs copies, alors que chez *D. melanogaster* elles sont plus jeunes et possèdent des séquences d'une longueur plus proche de celle de leur séquence d'origine inférée [203]. De plus, les séquences d'ET chez *D. simulans* sont moins homogènes que chez *D. melanogaster*, et pour de nombreux ET on peut observer deux variants du même élément. Par exemple, la Figure 2.1 présente l'alignement des deux variants de l'élément *412*, un rétrotransposon à LTR, présents dans le génome de *D. simulans*. On peut observer sur la figure que les deux variants sont très proches en terme de séquence, excepté au niveau de leur extrémité 5', correspondant à la région régulatrice de l'élément. Un de ces variants, *412DM*, est aussi présent chez *D. melanogaster*. Les séquences du variant commun aux deux espèces sont quasiment identiques ($\sim$ 99% d'identité nucléotidique), ce qui est bien supérieur à l'identité attendue entre les deux espèces compte tenu de leur temps de divergence ($\sim$ 95%). Cette observation a été montrée pour 10 autres éléments [203]. La grande proximité entre les séquences pourrait correspondre au TH de ces éléments entre ces deux espèces. Ainsi, pour revenir sur l'exemple de l'élément *412*, le variant commun pourrait avoir été impliqué dans un TH avec *D. melanogaster*.

Pour expliquer les différences de contenu en ET observées entre *D. simulans* et *D. melanogaster*, le modèle décrit dans la Figure 2.2 a été proposé par Emmanuelle Lerat. Pour plus de simplicité, ce modèle est présenté ici pour le cas d'un seul ET présent et actif chez l'ancêtre commun à *D. simulans* et *D. melanogaster*. Après la spéciation, cet élément aurait eu une histoire différente entre *D. simulans* et *D. melanogaster*. Chez *D. melanogaster*, une régulation efficace de l'élément aurait été rapidement mise en place conduisant à son inactivation et à la dégradation de ses copies. Chez *D. simulans*, la régulation de cet élément, moins efficace, aurait permis à certaines copies de rester

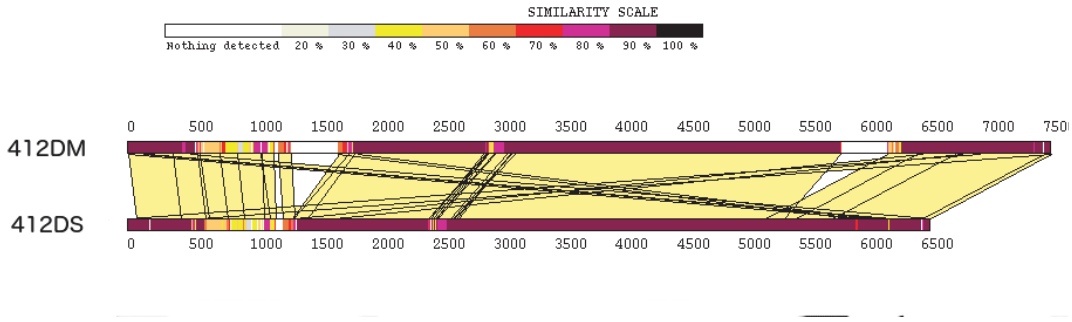


FIGURE 2.1 : Alignement des deux variants de l'élément *412* détecté dans le génome de *D. simulans*. Le variant du haut (*412DM*) correspond à la séquence canonique de cet élément, alors que le variant du bas (*412DS*) semble être spécifique au génome de *D. simulans*. (Figure fournie par E. Lerat)

actives. La séquence de ces copies actives aurait alors continué à évoluer, menant à la formation d'un nouveau variant pour cet ET. Une régulation incomplète de cet élément pourrait donc être la source des variants décrits chez *D. simulans* [203]. Plus tard le TH de ce nouveau variant depuis le génome de *D. simulans* vers celui de *D. melanogaster* aurait conduit à une explosion de transposition dans ce dernier génome, produisant un grand nombre de copies jeunes et actives chez cette espèce.

Ce modèle peut donc expliquer les différences observées entre le contenu en ET chez *D. simulans* avec deux variants pour de nombreux éléments dont l'un correspond à des copies anciennes et dégradées, et un variant encore actif qui présente une grande similarité de séquence avec les copies du même élément chez *D. melanogaster*.

Le scénario décrit dans ce modèle se serait répété pour plusieurs ET chez *D. melanogaster*, ce qui expliquerait que la plupart des ET sont jeunes et actifs chez cette espèce et que la séquence de certains des éléments pourrait correspondre à des variants d'éléments encore actifs chez d'autres espèces proches [203]. Ce scénario est par ailleurs compatible avec les nombreux cas de transferts horizontaux d'ET qui ont été mis en évidence dans la littérature entre *D. simulans* et *D. melanogaster* [206, 207, 208].

Comme nous l'avons vu dans l'introduction, les mécanismes permettant le TH d'ET entre différentes espèces eucaryotes sont toujours mal compris. Cependant, étant donné le nombre important de TH détectés entre *D. simulans* et *D. melanogaster*, la possibilité

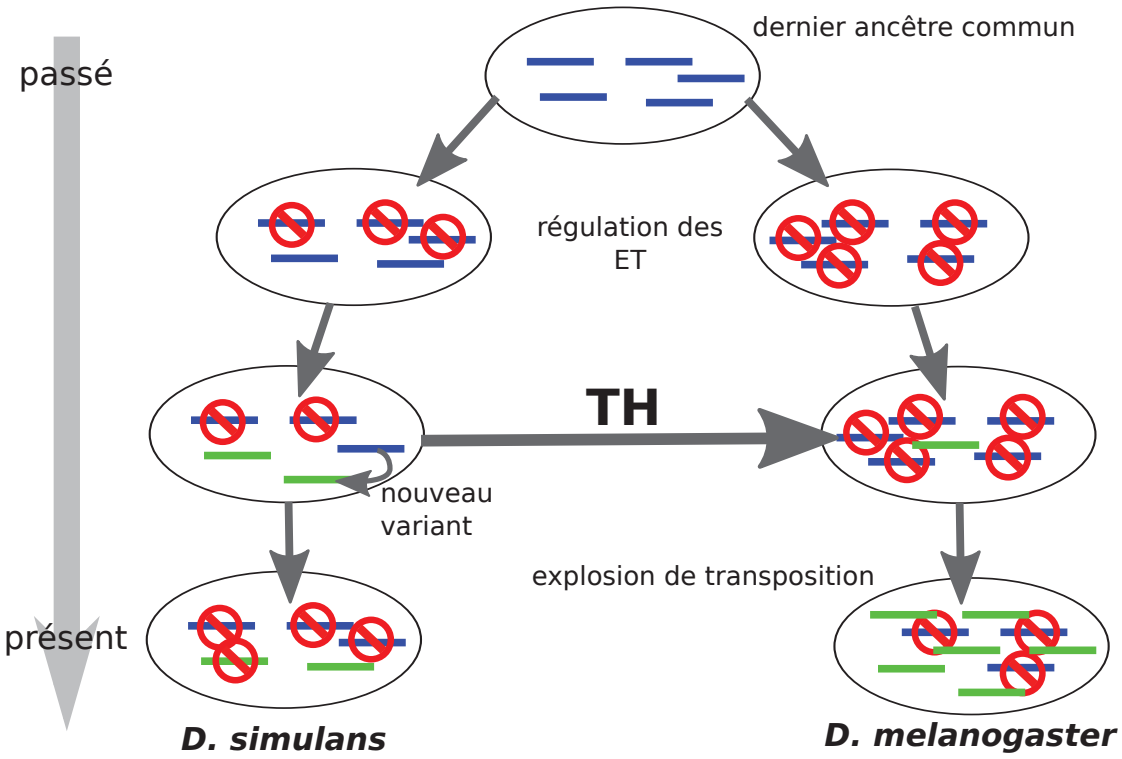


FIGURE 2.2 : Modèle de l'évolution d'un élément transposable chez *D. melanogaster* et chez *D. simulans*. Les copies correspondant au variant ancestral de l'élément sont représentées en bleu, alors que le nouveau variant est représenté en vert. Les copies sous un cercle rouge barré correspondent aux copies désactivées de l'élément.

de l'introgression d'un fragment d'ADN contenant des copies de différentes familles d'ET pourrait être une explication parcimonieuse à ces observations. La datation de la majorité de ces événements de TH les situe entre 1,4 et 2,3 Ma, avant l'expansion mondiale de ces deux espèces il y a 15000 ans [208]. À cette époque, ces deux espèces étaient toutes deux présentes en Afrique et pouvaient présenter un isolement reproducteur moins marqué qu'actuellement. La présence d'hybrides fertiles permettant une introgression aurait donc été possible, sans même nécessité de la présence de mutations permettant de restaurer leur fertilité [209]. Une introgression est caractérisée par le transfert d'un fragment d'ADN indéterminé. Ce fragment peut être constitué de séquences d'ADN intergéniques, comme de séquences de gènes et d'ET. Pour valider cette hypothèse, il faut donc pouvoir détecter le TH de séquences d'ADN indéterminées entre deux espèces.

2 La problématique

Il existe de nombreux exemples de TH d'ET dans la littérature. La majorité de leur détection par des méthodes bioinformatiques repose sur des approches séquence-spécifiques qui permettent de déterminer si une séquence d'ADN particulière a été transférée horizontalement entre deux génomes. Néanmoins, si les approches séquence-spécifiques permettent d'obtenir des résultats intéressants pour l'étude des TH de séquences prédéterminées, elles ne permettent pas d'étudier l'ensemble des séquences d'ADN pouvant avoir été transférées horizontalement entre deux génomes eucaryotes. Ce type d'approche ne permet donc pas d'étudier la possibilité d'une introgression de fragments d'ADN indéterminés entre *D. simulans* et *D. melanogaster*. Bien qu'il existe des approches pour la détection de TH pour des génomes complets [210, 211, 212, 213], leur manque de spécificité et de sensibilité les rend inutilisables pour des applications biologique réelles. Nous pouvons aussi ajouter à cette liste une autre méthode publiée en 2014, mais dont l'application est limitée à la détection de TH entre espèces divergentes et qui pourrait ne pas être adaptée à l'étude des séquences d'ET [214].

Cette absence de solution méthodologique adaptée pour la détection de TH pour des génomes complets m'a conduit à développer une nouvelle méthode pour identifier toutes les séquences, indépendamment de leur type, pouvant avoir été transférées entre deux génomes eucaryotes. Cette méthode, ainsi que son application aux génomes séquencés de *D. melanogaster* et *D. simulans*, a fait l'objet d'un article dans le journal "Genome Biology and Evolution", publié début 2014.

Ce travail a par ailleurs été présenté sous forme de posters ou de présentations orales lors de plusieurs conférences :

- ICTE 2012 à Saint Malo (poster) : On the track of horizontally transferred sequences between two closely related species *Drosophila melanogaster* and *Drosophila simulans*
- SMBE 2012 à Dublin (poster) : On the track of horizontally transferred sequences between two closely related species *Drosophila melanogaster* and *Drosophila simulans*
- GDR Élément transposable 2012 à Paris (présentation) : On the track of horizontally transferred sequences between two closely related species *Drosophila melanogaster* and *Drosophila simulans*
- GDRE 2013 Comparative Genomics à Barcelone (présentation) : A new genome-wide method to track horizontally transferred sequences : application to Drosophila

3 Les résultats principaux

La méthode que j'ai développée pour détecter les TH entre deux génomes complets eucaryotes se décompose en trois parties. La première partie correspond au calcul de la liste des meilleurs alignements locaux des séquences d'un génome contre l'autre. Ce calcul se base sur les résultats d'un BLAST nucléotidique entre ces deux génomes pour identifier ces paires de séquences [215]. Cette étape permet de réduire la liste de paires de séquences à analyser et repose sur le fait que mécaniquement, il ne peut y avoir qu'une séquence transférée horizontalement entre une position donnée de ces deux génomes.

Ensuite une estimation de la divergence nucléotidique attendue entre leurs génomes permet de tester si l'identité nucléotidique de chacune de ces paires de séquences est plus grande que celle attendue entre les deux génomes. Le grand nombre d'hypothèses testées a conduit au développement d'une nouvelle procédure pour la correction des tests multiples unilatéraux, assez sensible pour pouvoir détecter des paires de séquences avec une identité supérieure à celle attendue entre deux espèces proches. Les méthodes statistiques utilisées pour cette phase de correction des tests multiples sont le sujet du chapitre suivant.

Enfin la dernière partie de cette méthode consiste à identifier les séquences ayant été transférées horizontalement parmi la liste des paires de séquences plus identiques que l'attendu entre les deux génomes. En effet, nous nous attendons à retrouver des séquences sous sélection purificatrice et des séquences tranférées horizontalement dans la liste des paires de séquences plus identiques que l'attendu. J'ai donc construit deux

filtres (un pour les séquences non-répétées et un autre pour les ET) afin d'éliminer les paires de séquences conservées au sein des deux génomes qui ne correspondent pas à des TH. Pour les ET, la multiplicité des copies conduit à une plus grande probabilité que deux copies d'un élément soient plus identiques que l'attendu par hasard entre deux espèces. J'ai donc développé une nouvelle mesure appelée "activity track" qui permet d'étudier l'identité des copies d'un ET entre deux espèces. Cette mesure utilise la paire de copies de cet élément la plus identique entre les deux génomes. Ensuite la divergence de chaque copie de cet ET par rapport à cette paire de copies est calculée dans chaque génome. Dans le cas où cette paire de copies a été impliqué dans un TH, nous nous attendons à observer les conséquences d'une explosion de transposition avec un grand nombre de copies très similaires à cette paire chez au moins l'une des deux espèces. Ainsi, l'analyse de l'"activity track" d'un ET permet d'étudier sa dynamique récente par rapport à un hypothétique TH et de déterminer si ce TH a conduit à une explosion de transposition de cet élément.

Cette nouvelle méthode appliquée à *D. melanogaster* et *D. simulans* a permis de retrouver les 24 cas de transferts horizontaux d'ET décrits dans la littérature [206, 207, 208, 203] ainsi que 10 nouveaux cas qui n'avaient pas encore été décrits. Nous n'avons pas pu détecter de trace d'introgression entre les deux génomes séquencés de *D. simulans* et *D. melanogaster*, ce qui pourrait être expliqué par le fait que les allèles correspondants aux fragments d'ADN introgressés ne sont pas présents chez les individus séquencés, ou bien que ces 34 événements de TH sont le résultat de transferts indépendants. Nous pouvons aussi souligner l'absence de détection de transfert de gènes entre *D. simulans* et *D. melanogaster*, ce qui est en accord avec l'hypothèse que les transferts de matériel génétique réussis entre eucaryotes correspondent plus souvent à des ET et plus rarement à des gènes.

Pour conclure, la méthode que j'ai développée nous a permis de détecter de nouveaux cas de TH d'ET entre *D. melanogaster* et *D. simulans* qui sont pourtant des espèces très étudiées. Ces résultats soulignent l'importance de travailler sur des génomes complets et non pas uniquement sur des séquences particulières d'ET pour étudier leur dynamique. Avec ces analyses, nous avons mis à jour une portion du riche réseau de TH d'ET qui semble interconnecter les génomes de drosophiles. Cette méthode peut être appliquée à d'autres paires de génomes séquencés chez les eucaryotes pour avoir une meilleure compréhension du rôle des TH dans la dynamique des ET.

A New Genome-Wide Method to Track Horizontally Transferred Sequences : Application to Drosophila

Laurent Modolo, Franck Picard et Emmanuelle Lerat

Genome Biology and Evolution, 6(2), *416-432.*

Publié en 2014

Genome Biology and Evolution

SMBE

A New Genome-Wide Method to Track Horizontally Transferred Sequences: Application to *Drosophila*

Laurent Modolo[1], Franck Picard[1], and Emmanuelle Lerat[1,*]

[1]Université de Lyon, France, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne, France

*Corresponding author: E-mail: emmanuelle.lerat@univ-lyon1.fr.

Accepted: January 28, 2014

Abstract

Because of methodological breakthroughs and the availability of an increasing amount of whole-genome sequence data, horizontal transfers (HTs) in eukaryotes have received much attention recently. Contrary to similar analyses in prokaryotes, most studies in eukaryotes usually investigate particular sequences corresponding to transposable elements (TEs), neglecting the other components of the genome. We present a new methodological framework for the genome-wide detection of all putative horizontally transferred sequences between two species that requires no prior knowledge of the transferred sequences. This method provides a broader picture of HTs in eukaryotes by fully exploiting complete-genome sequence data. In contrast to previous genome-wide approaches, we used a well-defined statistical framework to control for the number of false positives in the results, and we propose two new validation procedures to control for confounding factors. The first validation procedure relies on a comparative analysis with other species of the phylogeny to validate HTs for the nonrepeated sequences detected, whereas the second one built upon the study of the dynamics of the detected TEs. We applied our method to two closely related *Drosophila* species, *Drosophila melanogaster* and *D. simulans*, in which we discovered 10 new HTs in addition to all the HTs previously detected in different studies, which underscores our method's high sensitivity and specificity. Our results favor the hypothesis of multiple independent HTs of TEs while unraveling a small portion of the network of HTs in the *Drosophila* phylogeny.

Key words: horizontal transfer, genome-wide method, *Drosophila*, transposable elements, FDR.

Introduction

Thanks to next-generation sequencing (NGS) technologies and to recent advances in de novo genome-assembly algorithms, we now have access to an increasing number of complete eukaryotic genomes. This methodological shift toward deep sequencing has changed the scale of investigation for many genomic studies and now allows the study of horizontal transfers (HTs) between eukaryotic species (Gilbert et al. 2010, 2013; Gilbert and Cordaux 2013).

HTs are defined by an exchange of genetic material between two reproductively isolated organisms (Gilbert et al. 2009) or by a movement of genetic information across normal mating barriers between more or less distantly related organisms (Keeling and Palmer 2008). Contrary to prokaryotes, for which HTs are common and well described (Fall et al. 2007; Juhas et al. 2009; Weinert et al. 2009), HTs are thought to be rare in eukaryotes, and their underlying mechanisms remain unknown (Andersson 2005). Proposed hypotheses to explain HTs in eukaryotes range from virus-mediated HTs using direct transfer of episomes (O'Brochta et al. 2009), viral particles, or infection (Kim et al. 1994; Dupuy et al. 2011) to parasite-mediated transfers (Gilbert et al. 2010). Overall, the main difference between eukaryotes and prokaryotes regarding HTs resides in the type of DNA material that is transferred: HTs usually involve genes in prokaryotes (Ochman et al. 2000), whereas in eukaryotes, HTs usually involve noncoding DNA and transposable elements (TEs) (Schaack et al. 2010). Following the availability of complete assembled genomes, more attention has been directed to the detection of HTs in eukaryotes, but most studies rely on similar approaches to the ones used for the detection of HTs in prokaryotes (Doyon et al. 2011). However, the differences in the type of horizontally transferred sequences between prokaryotes and eukaryotes and in the quantity of DNA to be investigated raise specific methodological challenges that need to be addressed to obtain a broader picture of genome-wide HT dynamics in eukaryotes (de Carvalho and Loreto 2012).

A particularity of the detection of HTs in eukaryotes is that it first requires the genome-wide identification of candidate

pairs of sequences that are not necessarily predefined. This point has motivated the development of several approaches, such as the surrogate method, which relies on differences in nucleotide patterns consistent with foreign DNA (Ragan 2001; Putonti et al. 2006). Nevertheless, this type of approach displays such a high rate of false detections that it is not efficient for real case studies (Azad and Lawrence 2011). Other genome-wide approaches start with all-to-all Blast searches between genomes of many species and detect HTs using an arbitrary cutoff using *e*-values (Shi et al. 2005) or a lineage probability index (Podell and Gaasterland 2007). However, although these strategies have given better results than the surrogate method, the lack of statistical framework for the detection of HTs in both methods has limited the interpretation of their results and the precise assessment of their specificity and sensitivity (de Carvalho and Loreto 2012). Overall, the promises of genome-wide approaches have been tempered by these common drawbacks, which explain the prevalence of sequence-specific approaches in HT studies even for genome-wide data sets.

When focusing on sequence-specific approaches to study HTs, we can discriminate between tree-topology-based approaches and sequence-divergence-based approaches. In prokaryotes, the gold standard for detecting HTs relies on the study of incongruences between the phylogeny of the sequences undergoing HTs and the phylogeny of the species. Because the pairwise identity of a horizontally transferred sequence is higher than expected according to the divergence time of the two species (Silva et al. 2004; Loreto et al. 2008), a phylogenetic tree based on this sequence will be discordant from the species tree. Unfortunately, phylogenetic approaches require a large taxonomic sampling of genes to have sufficient power of detection, which is often lacking in eukaryotes. Moreover, this method poorly differentiates HT genes from ancestral gene duplication(s) followed by gene loss(es) (Roger 1999). Phylogenic incongruences can also be produced when two or more variants of the ancestral lineage sequence have been stochastically inherited by the derived lineages (Dias and Carareto 2012). Finally, another pitfall of these approaches is the possibility of phylogenetic reconstruction artifacts, which can lead to strongly supported but false trees and thus to false positives for HT detection.

Studying pairwise sequence divergences constitutes an alternative that is commonly used when working with eukaryotes. It can rely on different divergence metrics, such as the synonymous substitution rates (d_S or K_s), to test the consistency of the number of synonymous differences accumulated between two sequences with the divergence time between the two species. Confounding factors can also decrease the power of d_S-based approaches. Codon usage bias, for instance, can result in a reduced d_S for the reference genes, which can decrease the sensitivity of detection of sequences with low d_S (Wallau et al. 2012). Purifying selection and variable rates of sequence evolution can also lead to spurious HT detections or a lack of power for identity-based methods (Capy et al. 1994; Pace et al. 2008). Finally, a third line of evidence for the detection of HTs is a patchy distribution of the sequences within a group of taxa (as they are not vertically transmitted). However, because of stochastic losses, the lack of coverage of some parts of the genomes and the random sampling of the population alleles in the sequenced strains, this third line of evidence is hardly self-sufficient to infer an HT event (Keeling and Palmer 2008; Schaack et al. 2010).

One strategy to control for spurious HT detections has been to focus on one line of evidence for the detection of HTs and to rely on the two others for validation purposes (Loreto et al. 2008; Gilbert et al. 2010). However, when dealing with eukaryotes, the absence of evidence for phylogenetic incongruences and the absence of a patchy distribution are likely to be poor validating arguments, as they do not constitute strict evidence against the possibility of an HT (Wallau et al. 2012). Another weakness of current sequence-specific approaches is that both tree-topology- and sequence-divergence-based approaches are restricted to coding sequences (CDSs). This represents only a small part of most eukaryote genomes and introduces an important detection bias for the analysis of horizontally transferred sequences.

In eukaryotes, for which HT events involve noncoding DNA and TEs, only 330 cases of horizontally transferred TEs have been described to date (Wallau et al. 2012) compared with rates as high as 30% of lateral gene transfers per phylogenetic branches for prokaryotes (Abby et al. 2012). TEs are DNA segments that are able to replicate and insert themselves into the genome using different mechanisms (Finnegan 1997; Wicker et al. 2007; Jurka et al. 2011). One of the outstanding features of TEs is their ability to cross species boundaries and invade new genomes (Daniels et al. 1990; Pinsker et al. 2001; Ludwig et al. 2008). These elements can represent the most abundant part of large eukaryotic genomes, as is the case of the maize genome (85%) (Schnable et al. 2009) and of the human genome (between 45% and 78% according to the detection method [Lander et al. 2001; de Koning et al. 2011]).

Notably, among the 330 horizontally transferred TEs detected, 178 concern drosophilid species, and from the 101 putative HT events proposed in *Drosophilae* in 2008, only 15% were confirmed by the three lines of evidence we have mentioned (Loreto et al. 2008). Regardless of this overrepresentation of drosophilids, the majority of these 330 HT detections relied on sequence-specific studies of candidate sequences. With this approach, only a small part of the genomes is exploited, which leads to an underestimation of the number of HTs. Our proposed genome-wide approach aims to solve this bias by requiring no prior knowledge concerning the sequences of interest and evaluating all the identifiable pairs of sequences between two genomes with an identity-based approach. Our method addresses the detection of all HTs genome wide as a multiple-testing problem to

handle this large number of identity-based detections and to control the proportion of false positives in the results (Wei et al. 2009). We also propose two new filtering methods to sort out spurious HT detections corresponding to conserved sequences in the results.

We applied our method to the genome-wide detection of all putative HT sequences between two *Drosophila* species: *Drosophila melanogaster* and *D. simulans*. These two cosmopolitan *Drosophila* species have a divergence time estimated between 4.3 and 6.5 Myr (Tamura et al. 2004) and are highly similar on many points, except in their TE content. TEs in *D. melanogaster* represent a large amount of the genome (15% [Dowsett and Young 1982]), with mainly young and active (highly similar) copies (Bowen and McDonald 2001; Kaminker et al. 2002; Lerat et al. 2003). In contrast, the TEs in *D. simulans* are represented mainly by old and degraded copies (Lerat et al. 2011) and only account for 6.85% of the genome (Hu et al. 2013). To explain the differences in the TE landscape between these two species, previous studies based on a restricted number of TEs have shown that numerous HTs were likely to be involved (Bartolomé et al. 2009; Lerat et al. 2011) (see Carareto [2011] for a review). To obtain a broader picture of HT between these two genomes, we performed a whole-genome comparison study between *D. melanogaster* and *D. simulans* assuming that undefined fragments of DNA may have been transferred from one species to the other. These undefined fragments of DNA can contain any types of sequences, such as TEs, nuclear genes, or intergenic DNA, thus removing any detection bias toward CDSs. As a result, we detected 10 new putative horizontally transferred TEs in addition to all the horizontally transferred TEs described by different studies between *D. melanogaster* and *D. simulans*, bringing to light a portion of the rich network of HTs that seems to link together the *Drosophila* species.

Materials and Methods

Our method can be divided into two main parts. For the first part, it relies on a multiple-testing framework to identify with a high sensitivity all the sequences that may have been horizontally transferred between two species at the genome scale. This approach is divided into three different steps described later. Then, we developed a multiple-testing framework to evaluate the output of multiple identity-based detections of HTs while controlling for the expected proportion of false positives in the results. A novelty of our approach is the modeling of the data throughout the genome as candidate sequences that are structured spatially, accounting for their dependency structure with a nonhomogeneous Markov model (NHMM) to increase the power of the multiple-testing correction (Kuan and Chiang 2012). For the second part of our method, we discriminate between putative HTs and other mechanisms, leading to a high pairwise identity to increase our specificity. For this purpose, we propose two novel validation procedures that can be applied for genome-wide studies to control for the numerous sources of spurious detections inherent to the detection of HT.

We will thereafter introduce the software, the algorithms, and the statistical models that we used for the different parts of this approach (supplementary fig. S1, Supplementary Material online). In our application, genome A corresponds to the genome of *D. melanogaster* and genome B to the genome of *D. simulans*.

Description of the Tree Steps for the Detection of Putative Horizontally Transferred Sequences between Two Genomes A and B

Step 1: Selection of the Sequences of Interest

To identify HT events, we define a sequence of interest as part of a pair of sequences with a higher pairwise nucleotidic identity than expected between the two species A and B. This first part of the pipeline aims to delimit such sequences in the two genomes. To achieve this goal, we start by retrieving the list of all the identifiable pairs of sequences between the two species A and B. For this step, we performed a nucleotidic all-to-all `Blast` (version 2.2.26) of one genome against the other (Altschul et al. 1990). The output of such a Blast defines a many-to-many cardinality between sequences from the two species, meaning that a given sequence from one species can be linked to many sequences in the other species, and vice versa. These types of links are complex and represent a large quantity of data to address. Moreover, as we cannot observe two different horizontally transferred sequences at the same locus in the species A, we filter the resulting pairs of sequences to only retain the best match for each position of the genome of A. For the task at hand, we only need the best local alignments of sequences for each position along the genomes because the other alignments would have a lower identity and thus a lower probability to correspond to an HT event.

To parse the Blast output and obtain a one-to-one cardinality from a many-to-many cardinality, we developed in python the program `htdetect.py` (available from the online resources). This program uses the fact that when working on two different genomes, there is always a genome of better quality (genome A) than the other genome (genome B). Our algorithm can be divided into the four following stages (fig. 1):

1. Compute the identity between each pair of sequences and the corresponding *P*-values to account for the identity and the size of the pair of sequences (see unilateral binomial test later) (fig. 1*A*).
2. Order all the pairs of sequences according to their position in genome A (fig. 1*B*).
3. Merge all the overlapping pairs of sequences in genome A to obtain a one-to-many cardinality from the many-to-many cardinality (fig. 1*C*).

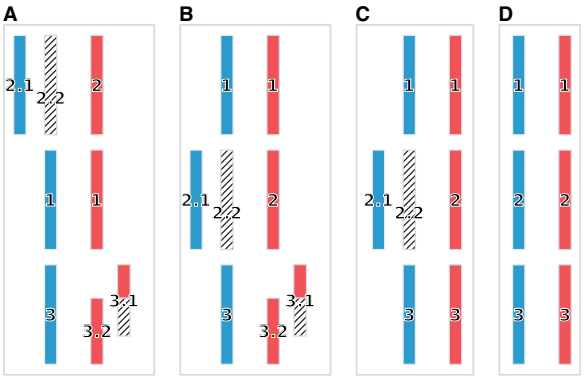


Fig. 1.—Algorithm to reduce the many-to-many cardinality in the results of an all-to-all nucleotidic Blast to a one-to-one cardinality between a genome A (red) and a genome B (blue). (*A*) Compute the identity between each pair of sequences and the corresponding *P* values (see Materials and methods, stage 3), and order all the pairs of sequences according to their position on genome A (the sequence order is 1-2-3). (*B*) Merge all the overlapping pairs of sequences in the genome A to go from a many-to-many cardinality to a one-to-many cardinality (remove the dashed part of the sequence 3.1). (*C*) Keep the sequences with the lowest *P* values from genome B for each pair of sequences that were merged in stage 3 to obtain a one-to-one cardinality (remove the dashed sequence 2.2). (*D*) One-to-one cardinality between the two genomes.

4. Keep the sequences with the lowest *P* values from genome B for each pair of sequences that have been merged in step 3 to obtain a one-to-one cardinality (fig. 1*C* and *D*).

In the hypothetical case where both genomes are of equivalent quality, the above steps will be strictly symmetrical to obtain a one-to-one cardinality.

Step 2: Computation of the Expected Pairwise Identity between the Compared Species

To test H_0 : "the number of differences is greater than or equal to the expected number of differences," for each of the filtered sequences, we compute the expected pairwise nucleotide identity between the species A and B, given their time of divergence. For this purpose, we used a global pairwise alignment of the genome of the species B against the genome of species A. We compute the number of identical nucleotides for each nonoverlapping window of size 1 kb along each chromosome arm of the species A. The size of 1 kb was empirically chosen as a trade-off between the resolution for the identity computation (of 0.01%) and information about the identity variation (a large window size only gives access to the average identity). To compute the nucleotide identity percentage between the species A and B for each of these windows, we removed the unknown nucleotides and the gaps from the computation.

We then used a Gaussian kernel smoothing function of these nonoverlapping window identity scores to obtain the distribution of the nucleotide identity between the two species. As this identity distribution is skewed to the right in the case of our application to *Drosophila* species (fig. 2), we chose to use the highest mode of this distribution as the expected pairwise identity between the two genomes, instead of the mean or a given quantile.

Step 3: Test of the Sequence Pairwise Identity

To model the pairwise identity, for every pair of sequences n, we denote by W_n the number of different nucleotides between the two sequences. The distribution of W_n is $\mathcal{B}(L_n, p_n)$, where L_n is the length of the pair of sequences of interest and p_n is the probability of having a nucleotidic dissimilarity. Our aim is to test $H_0 : \{p_n \leq p_0\}$ accounting for L_n, in which $1 - \widetilde{p_0}$ is the expected identity calibrated using the reference distribution constructed from the global alignment of the two genomes (=95.62% for our application). Thus, we compute for each pair of sequences n the probability $P(w_n^{\text{obs}}) = P\{W_n \geq w_n^{\text{obs}}\}$ of having a number of different nucleotides lower than expected, or unilateral *P*-value.

The number of tests N equals the number of candidate pairs of sequences for each chromosome arm and for the whole genome. Thus, for a given level of type I error (e.g., $\alpha = 0.05$), with a crude estimate under independence of the tests, the number of false positives ($N \times \alpha$) can be larger than the number of positives.

At each position along the genome of species A, we have a *P*-value denoted by $P(w_n)$ that is distributed according to a uniform distribution in [0,1] under H_0. From this *P* value, we want to infer an indicator variable denoted by S_n, such that $S_n = 1$ if H_0 is rejected at position n and $S_n = 0$ otherwise. To proceed, we use the local false discovery rate (ℓFDR) strategy, which consists in assessing the posterior probability that S_n is under H_0 (Efron et al. 2001). Instead of using raw *P*-values, a standard strategy consists in using the inverse probit transform, such that $z_n = \Phi^{-1}(P(w_n))$, which results in centered standard Gaussian variables for the z under H_0, whereas the others follow an unknown density distribution f_1. Then, the posterior probability of being under H_0 is $\ell FDR_n = P(S_n = 1|z_n)$. The decision rule consists in selecting positions $n = 1, \ldots, \ell$, such that $\ell = \max\left\{i : (1/i)\sum_{j=1}^{i} \ell FDR_i \leq \alpha\right\}$, where $\ell FDR_1, \ldots, \ell FDR_N$ is ordered and α is the false discovery rate (*FDR*) level (Benjamini and Yekutieli 2001).

By mapping candidate sequences along the genome of species A, we expect the probability for one locus to have a higher pairwise nucleotide identity than expected to depend on its neighbors. Moreover, with the fragmentation of the candidate sequences due to the nucleotidic Blast, we also could detect small adjacent pieces of this locus instead of a unique DNA fragment, and because of their small sizes, each of these pieces of alignment could be statistically nonsignificant on its own. In the case of dependency, all the

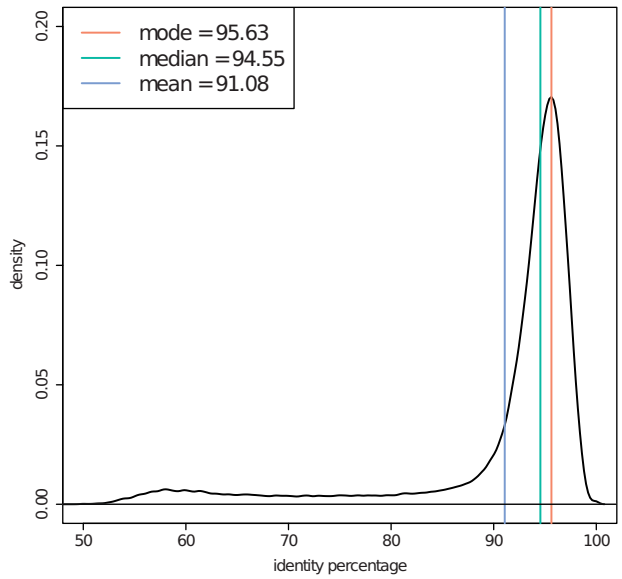


FIG. 2.—Distribution of the pairwise nucleotide identity, genome wide with nonoverlapping windows of size 1 kb, of the *Drosophila simulans* genome alignment on the *D. melanogaster* genome. The vertical bars represent the values of the mean, the median, and the mode of this distribution.

multiple-testing procedures not accounting for the dependency structure are suboptimal (Wei et al. 2009), meaning that if the procedure controls for the *FDR* for a given level, it does not minimize the false nondiscovery rate. Because decision at position n may depend on neighbor tests, we used the local index of significance (*LIS*) to compute $P(S_n = 0|z_1, \ldots, z_N)$ (Sun and Tony Cai 2009).

To proceed, we considered a homogeneous hidden Markov model in which S_n is the hidden states ($S_n \in \{0, 1\}$), which is governed by transition probabilities $P(S_{n+1}|S_n)$. Moreover, we also accounted for the genomic context of each sequences, like GC content or the distance between the sequences that can influence the transition and emission probability of the model, as we do not expect the dependency of a given sequence to its neighbors to be the same between every sequences. We considered a logistic regression to account for covariates $X_1, \ldots, X_N$ characterizing the sequences, such that:

$$P(S_1 = j|X_1 = x) = \frac{\exp\left(\lambda_j + \rho_j^n \times x\right)}{\sum_{k=0}^{1} \exp\left(\lambda_k + \rho_k^n \times x\right)}$$

$$P(S_n = j|S_{n-1} = i, X_n = x) = \frac{\exp\left(\sigma_{ij} + \rho_j^n \times x\right)}{\sum_{k=0}^{1} \exp\left(\sigma_{ik} + \rho_k^n \times x\right)}$$

with $i, j = \{0, 1\}$ (Kuan and Chiang 2012). The model parameters $\Psi = (\kappa, f_1, \lambda_i, \sigma_{ij}, \rho_j)$, with κ being the proportion of *P*-values equal to 1, can be estimated using the EM algorithm. We developed a zero-inflated Gaussian distribution to handle unilateral tests with the appropriate *z*-values transformed. This model is implemented in the `R` package `fdrDEP` available on the `CRAN` for multiple unilateral hypothesis testing.

The *LIS* statistics are computed for each chromosome arm of the species A and concatenated to control for the *FDR* at a level of 10% for the whole genome of A with the Benjamini, Hochberg, and Yekutieli procedure (Wei et al. 2009).

Filtering for True Putative HT Events

With steps 1–3, we could have detected highly similar fragments of sequence alignments that would not have been significant for the whole corresponding sequences, so we first recovered the full length of each annotated DNA fragment detected in the species A. To reconstruct the complete sequences for these results, we used the `bedtools` suite (version 2.17.0, options `intersectBed -a annotations.gff -b results.bed -wa`) (Quinlan and Hall 2010) to extract the annotated sequences corresponding to results with positions intersecting the ones from the species A. Then, we applied the two following filters to sort out conserved sequences from our results for nonrepeated and repeated sequences.

For Nonrepeated Sequences

For CDSs, we expect to observe an effect of selection because nonsynonymous mutations can be deleterious, neutral, or advantageous. Thus, for the CDSs identified with our approach, we can compute their d_S values using orthologous genes. We then performed the same unilateral binomial test as for the nucleotidic identity to determine whether the d_S of a given CDS is significantly lower than the expected identity between the two species considered while controlling for the *FDR* at a level of 10% (Benjamini and Yekutieli 2001).

In addition, to take into account non-CDSs that cannot be used in d_S approaches, we developed a new validation procedure based on sequence conservation, which can be applied to both coding and non-CDSs. In the set of detected sequences, a sequence identified with the same level of significance, both between *D. melanogaster* (the species A) and *D. simulans* (the species B) and between *D. melanogaster* and other *Drosophila* species, would illustrate a conserved sequence across the phylogeny rather than multiple HTs at the same position in *D. melanogaster*. Thus, we performed the same analysis with four other species from the 12 *Drosophila* genomes project: *D. sechellia*, *D. yakuba*, *D. pseudoobscura,* and *D. virilis*, as a gradient of phylogenetically divergent species, before subtracting these results from those of the *D. melanogaster–D. simulans* analysis. We used the `bedtools` suite to subtract the `.bed` tracks of the results of each species along the *D. melanogaster* genome. Figure 3 describes the decision rule used in this subtraction according

to the corresponding phylogenetic tree. This step provided us with a landscape of all the sequences with a pairwise identity higher than expected between *D. simulans* and *D. melanogaster* and not conserved in the other *Drosophila* species. This last filter relies on the strong hypothesis that a pair of sequences absent between a given pair of species is not missing due to random sampling of the population alleles in the sequenced individuals, a lack of genome coverage, or a misassembly.

As TEs and other repeated sequences are present at multiple loci between a pair of genomes, they were excluded from this filtering step and were validated separately, considering that we could not discriminate which TE copy identified between two genomes corresponded to a specific locus in the genome of the species A.

For Repeated Sequences

With our genome-wide approach, the set of TEs detected was not restricted to elements with a coding capacity, preventing us from relying on the d_S metric for their validation. Moreover, for TE family with a large number of copies, we can expect one or more of these copies to be more identical than expected between the two genomes just by chance. To account for the full set of detected TEs and analyze each detected TE family, we developed a new validating procedure based on the recent dynamics of the detected TEs in the genomes of species A and B. We worked under the hypothesis that, after an HT, a TE escapes the host defense mechanisms for a time and quickly replicates itself in the new host genome (Anxolabéhère et al. 1988; Le Rouzic and Capy 2005; Granzotto et al. 2011). Thus, in the case of an identifiable horizontally transferred TE, we expected to observe many highly similar copies of the TE corresponding to this burst of transposition in one or both genomes, in contrast to few conserved TE insertions (Lerat et al. 2011; Dias and Carareto 2012).

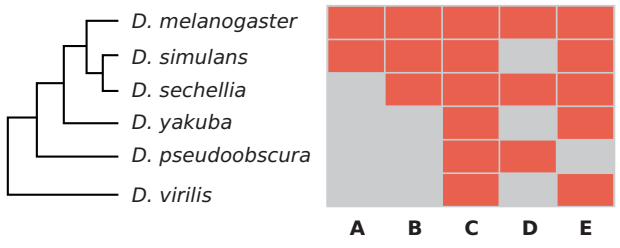


Fig. 3.—Decision rule for the filtering step about selective pressure, with the presence (red) or absence (gray) of a pair of sequences between the corresponding species. (*A*) Putative HT between *Drosophila melanogaster* and *D. simulans*. (*B*) Putative HT between *D. melanogaster* and *D. simulans* prior to the *D. sechellia* speciation event or conserved sequences between *D. melanogaster*, *D. simulans,* and *D. sechellia*. (*C*) Conserved sequences in the *melanogaster* subgroup. (*D*–*E*) Conserved sequences with stochastic loss or ancestral polymorphisms.

To help in the synthetic interpretation of the recent history of each TE in our results, we start by defining the most identical pair of copies between the two genomes, as the last putative horizontally transferred copy in the case of an HT. For each TE family, this most identical pair of copies between the two genomes is defined as the pair of copies with the lowest *P*-values from all the detected copies using the 80-80-80 rule (Wicker et al. 2007). Then, we Blast each of these most identical copies on the genome of A using a nucleotidic `Blast` (version 2.2.26) (Altschul et al. 1990). We built an index of the similarity of each copy of these elements compared with the most identical pair of copies between the two genomes, normalized by the size of the copies. We called this index the activity track. These activity tracks are used to rank between 0 and 1 all the copies of each identified TE according to their divergence from the corresponding most identical pair of copies between the two genomes, with 1 corresponding to a low degree of divergence and a recent activity of this TE and 0 corresponding to old and divergent copies. The activity track corresponds to the probability of having a pairwise nucleotidic identity with the most identical pair of copies less than or equal to the expected identity $1-\widetilde{p_0}$, estimated using the reference distribution constructed from the global alignment of the two genomes. For every pair of TE copies n, we denote by W_n the number of different nucleotides between the two copies. The distribution of W_n is $\mathcal{B}(L_n, p_n)$, where L_n is the length of the alignment between the copies and p_n is the nucleotidic dissimilarity. Our aim was to compute for each pair of sequences n the probability $P\{W_n \leq w_n^{\text{obs}}\}$ corresponding to the activity track. The same analysis is performed with the genome of species B to get an overview of the TE activity in both genomes. We developed in python the program `activity_tracks.py` (available from the online resources) to compute this index.

Finally, we manually inspected the results in `.bed` format on each chromosome arm of *D. melanogaster* to look for cluster of sequences with a higher identity than expected using the `integrative genome viewer` software (Thorvaldsdóttir et al. 2013).

All the statistical analyses in this article were performed using the software `R` (version 3.0.0) (R Core Team 2013).

Data Acquisition

We used the last available versions of the genomes of *D. melanogaster* (species A) (version r5.49), *D. sechellia* (version r1.3), *D. yakuba* (version r1.3), *D. pseudoobscura* (version r2.30), and *D. virilis* (version r1.2) and the corresponding annotation tracks from flybase (http://flybase.org [Marygold et al. 2013]). For *D. simulans* (species B), we did not work at first on the genome sequenced by the 12 *Drosophila* genomes project (Drosophila 12 Genomes Consortium 2007). Indeed,

this genome is a patchwork of six independently derived strains with the assembly of six major chromosome arms representing only 101.3 Mb of the 137.8 Mb expected. Moreover, this genome presents several major misassemblies and has the worst read quality of the 12 *Drosophila* genomes (Hu et al. 2013). This is why we used the *D. simulans* genome that was resequenced in 2012 and assembled from the *w501* strain of the original Sanger data, in addition to a high-coverage Illumina sequencing of iso-females of this same strain (Hu et al. 2013). However, to be able to compare our approach with previous studies, we also conducted a second analysis with the genome of *D. simulans* (version 1.3) available from flybase (http://flybase.org).

The genome pairwise alignments were retrieved from the UCSC website (http://hgdownload.cse.ucsc.edu).

The sequence annotation tracks used to obtain the full length of the corresponding TEs and CDSs and annotate the noncoding DNA were downloaded from flybase (http://flybase.org) in `.gff` format (Marygold et al. 2013).

Instead of computing the d_S of the detected CDSs, we used the d_S data of the 11,000 orthologs from the 12 *Drosophila* genomes, available from the study of Heger and Ponting (2007).

Quality of the TE Content in the Genome of *D. simulans*

We used the software `SeqGrapheR` (Novăk et al. 2010) to analyze the TE content of the *D. simulans* genome directly from a uniform random sample of 900 k reads obtained from the 2012 genome project (`SRA:SRX159034`) (Hu et al. 2013). The assembled repetitions were annotated using `RepeatMasker` (version 3.3.0) (Smit AF, Hubley R, Green P, unpublished data).

Data Access

All the scripts used for our pipeline are available in a `git` repository at: git://dev.prabi.fr/modolo2013.

Results

Genome-Wide Detection of Sequences with a Higher Nucleotidic Identity than Expected

Defining the Set of Candidates for HT Detection

We kept the best local alignments obtained by the Blast search of *D. simulans* against *D. melanogaster* for each position in the genome of *D. melanogaster*, thereby taking into account the repeated content that is often removed from genome-wide alignment (i.e., best global alignment). The cumulative size of the filtered sequences decreased with the divergence time between a given species and *D. melanogaster*, which is consistent with the nucleotidic Blast algorithm (table 1). For example, we retrieved approximately 112 Mb of sequences between *D. melanogaster* and *D. simulans* (divergence time of 5.4 ± 1.1 Myr), compared with only 13 Mb between *D. melanogaster* and *D. virilis* (divergence time of 42.9 ± 8.7 Myr). However, such a trend was not observed for the number of filtered sequences, which can be explained by the fragmentation of the retrieved sequences, which increased with the phylogenetic distance (table 1). With this set of candidates, we used our method to determine whether the observed pairwise nucleotidic identity for each of these pairs of sequences was higher than expected between the considered species and *D. melanogaster*.

Assessing the Reference Distribution for Nucleotidic Identity

We computed a reference nucleotidic identity distribution with the analysis of the global alignment of the genome of *D. simulans* along the genome of *D. melanogaster* (fig. 2). This distribution accounted for the variations in nucleotidic identity along the two genomes, in contrast to the common mutation rate of $1.1 \pm 0.2 \times 10^{-8}$ mutations per site per year per lineage for the *Drosophila* phylogeny that has been computed on a limited number of nuclear genes (Tamura et al. 2004). Consequently, this mutation rate based on the molecular clock hypothesis (Weir and Schluter 2008) may not be representative of the pairwise nucleotidic identity between the whole genomes of *D. melanogaster* and *D. simulans* (Drosophila 12 Genomes Consortium 2007) and is not suitable for a genome-wide analysis. For the detection of HTs between *D. melanogaster* and *D. simulans*, we were only interested in the expected nucleotide identity corresponding to the accumulation of mutations between these two species since their time of divergence. Thus, we choose the highest mode of identity distribution as a reference to compute the unilateral *P*-values of our tests, which quantified the probability of each candidate to have a nucleotidic identity exceeding 95.63%, while accounting for the size of the alignment (fig. 2).

Controlling for False Positives in the Context of Genomic Dependencies

As in many genomic studies, the number of statistical tests to perform was large (168,325 pairs of sequences for the comparison *D. melanogaster* vs. *D. simulans*). If no multiple-testing procedure is applied, we can roughly expect to declare an average of 10% of the tests (16,832) to be false positives by retrieving all the *P*-values below 0.1, which can be higher than the number of true positives (Finner and Roters 2002). By applying the standard Benjamini–Hochberg multiple-testing correction with an *FDR* level of 10% (Benjamini and Hochberg 1995), without taking into account the dependency structure between the tests, we only retrieved 605 CDSs, 934 TE insertions, and 2,345 intergenic DNA fragments. Thus, we used our method to assess the probability that each pair of sequences has a higher pairwise identity than expected while accounting for its dependency to its neighbors, adjusted to

Table 1
Results of the Filter of the All-to-All Nucleotidic Blast between *Drosophila melanogaster* and the Corresponding Species

Species	Sequence Size (kb)			Number of Sequences			Divergence Time to *D. melanogaster* (Myr)
	Row	Filtered	Significant[a]	Row	Filtered	Significant[a]	
D. simulans	550,226	112,748	9,012	4,468,121	168,325	11,927	5.4
D. sechellia	1,219,599	111,909	5,452	7,947,377	170,394	7,025	5.4
D. yakuba	1,972,352	91,584	977	23,960,790	239,011	3,185	12.8
D. pseudoobscura	102,146	22,241	593	1,431,447	213,790	11,323	30.0
D. virilis	184,640	13,463	298	2,186,411	117,831	6,305	42.0

NOTE.—Row, results corresponding to a many-to-many cardinality; filtered, results corresponding to a one-to-one cardinality.
[a]Results corresponding to the significative identity-based tests after multiple-testing correction.

other genomic covariates to increase our sensitivity. Indeed, the GC content of the sequences as well as the distance between a pair of sequences and the next on a chromosome arm and the presence of TEs are likely to be proxies of the similarity of a pair of sequence to its neighbors. We also expected the recombination rate to be an important factor, but no significant correlation was found between the recombination data available for the genome of *D. melanogaster* and the *P*-values of our tests. With this correction applied to the 168,325 tests, we retrieved 7.3 Mb of sequences, including 2,651 fragments from CDSs (2.46 Mb), 3,967 fragments from insertions of 28 different TE families (201 kb), and a large number of intergenic DNA fragments (13,806 sequences corresponding to 4.68 Mb), between *D. melanogaster* and *D. simulans*.

Distinction between "True" HT Events and Biological False Positives

HT Sequences in the Light of Other Drosophila *Species*

We detected a set of sequences with an identity higher than expected between the genomes of *D. simulans* and *D. melanogaster* that was not reduced to HT sequences, thus we started by retrieving the full-length sequences of each annotated fragment from the genome of *D. melanogaster*. Then, we discriminated putative HT sequences from the sequences displaying a signature of functional constraints. We tested whether the d_S of the 2,651 detected CDSs was significantly lower than expected in the *D. melanogaster-D. simulans* analysis, and we finally retained 26 CDSs.

To discriminate between conserved and horizontally transferred sequences for the full set of detected nonrepeated sequence (i.e., both coding and non-CDSs), we used the comparative analysis between *D. melanogaster* and *D. simulans*, and between *D. melanogaster* and other *Drosophila* species. This subtraction allowed us to remove approximately 40% of the base pairs for the intergenic DNA (thus keeping 2.79 Mb of the 4.68 Mb), with a consistently high pairwise identity with *D. melanogaster* in this phylogeny. This result is consistent with the results from Casillas et al. (2007) where 38.6% of the noncoding DNA in *D. melanogaster* display the signature of functional constraints. We also retained 28 of the 2,651 CDSs with this second approach.

The intersection of the results from the d_S analysis with the ones from this subtraction led to the detection of 11 CDSs annotated from RNA-Seq data but of unknown function (Marygold et al. 2013). These 11 CDSs were sparsely distributed along all the major chromosome arms of *D. melanogaster* and found in clusters of CDSs with significant pairwise nucleotide identity but nonsignificant d_S. Thus, these 11 CDS in our results could be biological false positives caused by the dependency model used in the multiple-testing correction, lowering their probability of being under the null hypothesis due to their conserved neighbors, which does not support the hypothesis of their HT. For the detected noncoding DNA, we were not able to use the *D. melanogaster* annotations to retrieve the full-length sequences of the DNA fragments. This class of fragmented DNA, representing 63.91% of the detected DNA in our results, was annotated based on the *D. melanogaster* annotation tracks (supplementary table S1, Supplementary Material online) but was only analyzed as neighboring sequences of the detected CDSs and TE sequences.

Horizontally Transferred TEs

For the repeated sequence, we used the activity track to study the recent activity of the detected TE family. According to the activity track distributions, most of the detected TE families in our results presented a recent period of activity in *D. melanogaster* (supplementary figs. S2 and S3, Supplementary Material online), with a large number of copies highly similar to the most identical pair of copies between the two genomes (see e.g., the diver element, fig. 4*A*). However, for some elements such as the ancient element *INE-1*, described as having invaded the ancestor lineage of *D. melanogaster* and *D. simulans* (Kapitonov and Jurka 2003), the activity track showed a majority of divergent copies with only few ones close to the most identical copy between the two genomes (fig. 4*B*), as expected by chance for a large number of

Genome Biology and Evolution

SMBE

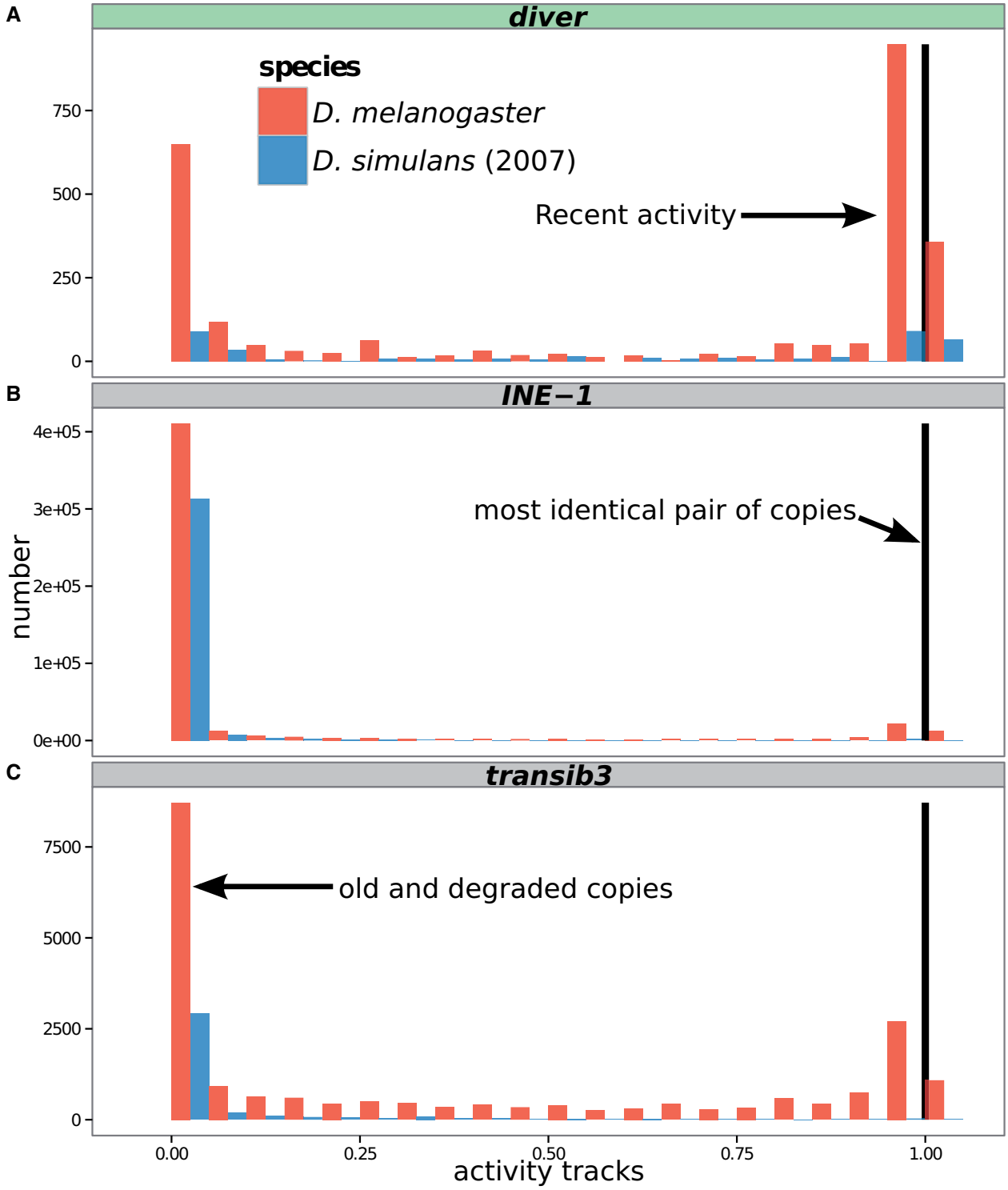


Fig. 4.—Density distributions of the activity tracks computed with the 2007 version of the genome of *Drosophila simulans*. The rightmost black bar corresponds to the most identical pair of copies between the two genomes, whereas the colored bars represent the number of copies ranked according to their similarity to this most identical pair of copies for a given TE. The red bars represent the activity tracks in *D. melanogaster,* whereas the blue bars represent the activity tracks in *D. simulans* 2007. (*A*) Example of a TE family presenting a recent period of activity corresponding to a putative HT from *D. simulans* toward *D. melanogaster*. (*B*) Example of a TE family with an activity not consistent with an HT between *D. simulans* and *D. melanogaster*. (*C*) Example of a TE family with different waves of activity.

old and degraded copies. With the activity track ranking the TE copies according to the most identical pair of copies between the two genomes which can be seen as the last putative horizontally transferred copy in the case of an HT, we were able to balance the direction of the transfers, which is crucial to understand the horizontally transferred TE history and dynamics. In the case of a horizontally transferred TE from a first species toward a second species, we can expect the TE to be present in a small number of highly similar and potentially active copies in the first species and in a large number of highly similar copies in the second species (supplementary fig. S3, Supplementary Material online). We observed this pattern for elements such as *diver*, which had a large number of copies with an activity track close to 1 in *D. melanogaster* and few copies in *D. simulans* (fig. 4*A*). This pattern was consistent with its HT from *D. simulans* toward *D. melanogaster*, even with few copies that were statistically significant in the two genomes. In the genome of *D. simulans*, most of the activity track distributions were bimodal, with few TE copies close to 1 and a large number of copies close to 0 corresponding to old and degraded copies (supplementary fig. S3, Supplementary Material online), which was consistent with the observations made for some of these TE families in the genome of *D. simulans* (Lerat et al. 2011). In contrast, in *D. melanogaster*, most of the TE copies had an activity track close to 1, which was representative of young and active TE populations. These differences of TE landscape between these two species support the hypothesis of multiple horizontally transferred TEs from *D. simulans* (fig. 4*A*) and from other species (fig. 4*C*) toward *D. melanogaster*.

With the 2012 version of the genome of *D. simulans*, we were able to identify 21 TE families with 10 new cases of horizontally transferred TEs, which were not previously identified as horizontally transferred between these two species (supplementary fig. S2, Supplementary Material online). However, 11 TEs were missing from the 24 horizontally transferred TEs previously described by different studies between *D. simulans* and *D. melanogaster* (de la Chaux and Wagner 2009; Bartolomé et al. 2009; Carareto 2011; Lerat et al. 2011). From these 11 TEs, the elements were only present in a few noncomplete copies in the 2012 version of the genome of *D. simulans*, which explain their absence from our results. For the elements *F*, *copia*, *gypsy5,* and *gypsy10*, the TE copies were highly divergent from those present in the genome of *D. melanogaster* and displayed a nonsignificant nucleotidic identity. To confirm the absence of the *412* element, known to be active in some populations of *D. simulans* (Vieira and Biémont 1997), we performed a de novo assembly of the TEs directly from the reads of the 2012 *D. simulans* genome project. The reads corresponding to this 7,566 bp element represented 50 kb of the 137.8 Mb genome of *D. simulans*, with the majority of the reads matching the long terminal repeats and few reads mapping within the element, which was concordant with the 2012 assembly. Therefore, the absence of these 11 horizontally transferred TEs from our results was likely the result of their absence from the assembled strain in the 2012 version of the genome of *D. simulans* rather than a lack of sensitivity of our method. Using the genome of *D. simulans* from the 12 *Drosophila* genomes project (Drosophila 12 Genomes Consortium 2007) (supplementary fig. S3, Supplementary Material online), we were able to recover in one analysis the 24 HTs previously described in the literature, including the 11 families missing from our analysis with the *D. simulans* genome of 2012. The 10 new and the 24 previously described TEs all presented activity track distributions consistent with the after effect of a horizontally transferred TE from *D. simulans* toward *D. melanogaster* (supplementary figs. S2 and S3 and table S2, Supplementary Material online). Thus, given the number of horizontally transferred TEs detected between *D. melanogaster* and *D. simulans* in the short time since their divergence, a parsimonious hypothesis could be the introgression of one or more fragments of DNA containing different TEs instead of multiple independent HTs.

Introgression versus Multiple HT Events

To obtain a broader view of the HTs between *D. melanogaster* and *D. simulans*, and discriminate between introgression and multiple independent HTs, we manually inspected the 11 CDSs, the TE insertions from the 21 families left and the 10,232 fragments of noncoding DNA in the final results along each chromosome arm of *D. melanogaster* and with the 2012 genome of *D. simulans*. In the case of introgression, we expected to observe the simultaneous transfer of these three types of sequences in one large DNA fragment. However, we found no sequence containing three or even two of these different types of sequences in the final results. This absence of completely introgressed fragments could be a consequence of the fragmentation of the detected sequences between the two genomes. However, we also did not find any obvious clusters of these different types of DNA along the chromosome arms of *D. melanogaster*. Overall, the types of detected sequences in our study support the prevalence of TEs and noncoding DNA in HTs between these two species. However, the informations contained in the genome of the sequenced individuals are not sufficient to support the hypothesis of multiple independent horizontally transferred TEs toward *D. melanogaster* rather than introgression events.

To better understand the horizontally transferred TEs involving *D. melanogaster*, we performed the same horizontally transferred TE analysis with the data from our comparison with the four other *Drosophila* species (*D. sechellia*, *D. yakuba*, *D. pseudoobscura,* and *D. virilis*) (supplementary figs. S4–S7 and table S2, Supplementary Material online). We detected numerous horizontally transferred TEs in this restricted window of time starting 5.4 ± 1.1 Ma, corresponding to the expected identity threshold between *D. melanogaster*

and *D. simulans*. The comparison between *D. melanogaster* and *D. sechellia* provided evidence for HTs of 21 elements between these two species (supplementary fig. S4, Supplementary Material online). *Drosophila sechellia* is the only species with a divergence time to the ancestor that it shares with *D. simulans* within the time window of our analysis, so for this species we can discriminate between horizontally transferred TEs involving *D. melanogaster* and the ancestor of *D. simulans* and *D. sechellia*, and horizontally transferred TEs involving *D. melanogaster* and *D. sechellia* (fig. 3). For the element (with an activity track nonconsistent with an HT in the 2007 version of the genome of *D. simulans* and absent from the 2012 version), the results rather indicate recent activity in both species, which suggest the existence of a third donor species, such as another *D. simulans* strain than those sequenced in 2007. We also observed the same pattern for 17 elements between *D. melanogaster* and *D. yakuba* (supplementary fig. S5, Supplementary Material online). However, we did not find any clear evidence of recent horizontally transferred TEs or a burst of transposition in the analysis of the TEs detected between *D. melanogaster* and *D. pseudoobscura*, or between *D. melanogaster* and *D. virilis*, which could be explained by the degree of fragmentation of the corresponding sets of sequences (supplementary figs. S6 and S7, Supplementary Material online). Finally, recent activity of the *Roo* element was detected in *D. melanogaster*, *D. simulans*, *D. sechellia,* and *D. yakuba*, which could be the result of independent HTs of this element toward these four species after their divergence (de la Chaux and Wagner 2009).

Discussion

In eukaryotes, the study of horizontally transferred sequences is confined to CDSs and often focuses on specific TEs. Thus, there are two systematic biases in the detection of HTs in eukaryotes: the candidate HT must be known and must have a coding capacity (Keeling and Palmer 2008). We propose a new genome-wide approach that aims to bypass these biases inherent to sequence-specific approaches by considering all the best local alignments of one genome to another as possible horizontally transferred sequences. Then, we test each of these sequences to retrieve those with a higher nucleotidic identity than expected between the two species while accounting for the multiplicity of the tests and their dependency structure throughout the target genome. We detected 2,651 CDSs, 3,967 insertions from 28 different TE families, and a large number of intergenic DNA fragments (13,806) more identical than expected from the 4,468,121 pairs of sequences identifiable between *D. melanogaster* and the 2012 version of the genome of *D. simulans*. Finally, we discriminated between spurious HT detection and putative HTs in our results with two novel validation procedures for genome-wide HT detection. And after manual inspection of the results, we retained 21 TE families as horizontally transferred between these two species, validating the prevalence of TE sequences in HTs between these two species without detection bias toward this type of sequence.

Genome-Wide Identification of Putative Horizontally Transferred Sequences

Previous genome-wide approaches used a wide range of procedures to infer sequences more identical than expected given the phylogeny of the species to detect HTs in eukaryotes (Loreto et al. 2008; Wallau et al. 2012). However, none of these procedures relied on a statistical testing framework to validate their sensitivity and specificity. This explains why sequence-specific approaches are still used: their particular reliability despite the limited set of sequences considered (Wallau et al. 2012). The collegial tests for the identity-based detection of horizontal transferred sequences in eukaryotes rely on the synonymous substitution rate, often in the form for a codon-based *Z*-test (Pace et al. 2008; Gilbert and Cordaux 2013). In our study, the set of candidates sequences was not restricted to the small coding portion of eukaryote genomes, and this justified the use of a binomial test to retrieve the sequences with a higher pairwise nucleotidic identity than expected between two species without any codon information while accounting for the size of each candidate. This simple model for codon substitution is sensitive enough to detect recent HTs for which we can expect a small saturation between sequences. The saturation corresponds to the occurrences of multiple mutations at a single nucleotide (or site), which leads to an underestimation of the nucleotidic divergence between two sequences because we can only observe the last mutation in the case of multiple mutations per site. A pair of sequences with saturation is expected to have more single mutations per site than multiple mutations per site. Thus, the complex cases, ill-defined by the model, will also correspond to the sequences in the "uninteresting" side of our unilateral hypothesis and will be correctly assigned to the set of nonsignificant sequences.

In genome-wide analyses, we often face multiple-testing issues, and our results underscore the importance of working with a well-defined statistical framework to control the number of incorrect detections and increase the power of the study. We also took advantage of the fact that when comparing two genomes, we always have a genome of better quality to map the detected candidate sequences, to greatly reduce the dimensionality of the data to be analyzed, thus increasing the power of our study (Storey and Tibshirani 2003). For our analysis, the now standard Benjamini–Hochberg *FDR* (Benjamini and Yekutieli 2001) procedure had a too low specificity to produce relevant results. This was caused by the dependency between the tests in our analysis, which was taken into account with the *LIS* framework to increase the specificity of our approach (Sun and Tony Cai 2009). Modeling this dependency between each pair of

candidates along the chromosome arms of *D. melanogaster* with a homogeneous Markov model (HMM) was not sufficient to retrieve all the horizontally transferred TEs described in the literature with the genome of *D. simulans* from the 12 *Drosophila* project (Drosophila 12 Genomes Consortium 2007). For our purposes, an HMM would have tended to homogenize the *LIS* statistics according to the information provided by the adjacent without taking into account any information about the type of sequences or the distance between them (i.e., the nonsignificant pairs of sequences surrounding a TE copy), which can explain these missing horizontally transferred TEs. With an NHMM, we were able to enrich the standard Markovian dependency according to covariates, such as the distance between the statistical tests along the genome and the presence of TEs, and to detect the 24 horizontally transferred TEs described in the literature (Carareto 2011) with an *FDR* level of 10% and the 2007 version of the genome of *D. simulans*, in addition to 10 new ones with the 2012 version of the genome (supplementary table S2, Supplementary Material online).

Our approach can be applied to any pair of sequenced species in which one species has an assembled genome, into which candidate sequences will be placed, to model the dependency structure between the tests with a NHMM. The specificity of this method is high enough to detect sequences more identical than expected between closely related species while controlling for the *FDR* in the results. This procedure could also be used with any other unilateral tests for different biological problems or to model the nucleotidic differences in ancient HTs or between more divergent species with a greater prevalence of saturation.

Two New Methods to Confirm HTs in Genome-Wide Studies

Most of the methods for HT detection in eukaryotes use sequence-specific approaches and rely on strong d_S evidence to infer putative HTs (Bartolomé et al. 2009; Lerat et al. 2011). In the remaining studies, the candidate sequences generally involve distantly related species and recent HT events where the identity line of evidence can be self-sufficient (Loreto et al. 2008), for example, the case of the TEs *SPIN* and *OC1* (Gilbert et al. 2010, 2013). This can also be the case for recent HTs, such as the well-known example of the *P* element transferred from *D. willistoni* to *D. melanogaster* less than 100 years ago, for which the nucleotidic identity is almost of 100% between the two species (Daniels et al. 1990). We were able to retrieve sequences with an identity percentage higher than 99% between *D. melanogaster* and *D. yakuba* for the elements *Doc*, *jockey,* and *transib3*, which was unexpected and could be sufficient to infer their HT. However, the number of obvious cases was small, and we needed to confirm the other HTs by other lines of evidence.

When studying the pattern of sequence divergence between genomes to infer HTs, we can encounter a large number of confounding factors that need to be checked (Siepel et al. 2005; Pollard et al. 2010). These factors range from natural turnover (gain or loss of functional elements) to the effect of purifying and positive selection, which can act on entire sequences or on parts of sequences, canceling the effect of divergence. For the study of HTs, we can add to this list the effect of different evolutionary rates for the sequence under consideration or the effect of stochastic losses in the phylogeny of the candidate sequence(s) (Loreto et al. 2008). Moreover, we have to rely on orthologs and sequence identification, which is nontrivial and can lead to numerous false positives (Gronau et al. 2013). The possibility of misplaced DNA sequences in the genomic database, polymerase chain reaction mispriming, contamination, incomplete sequence data, and poorly rooted trees can also be technical sources of errors for HT detection (Lisch 2008). Therefore, to differentiate between putative HT events and the possibility of vertical transmission, we need to investigate other lines of evidence (Loreto et al. 2008; Gilbert et al. 2010). In genome-wide studies of HT, in contrast to sequence-specific approaches, all the candidate sequences are not assumed to have been horizontally transferred from one species to the other, and the procedures need to include validation steps to produce sound results.

Validation of the Nonrepeated Content

Our approach follows an identity-based line of evidence to detect HTs, so we would need phylogenetic clues to validate them. In the case of *D. melanogaster* and *D. simulans*, which are almost at a terminal node of the *Drosophila* phylogenic tree, phylogenetic incongruences would mostly consist of nonsignificant differences in branch lengths compared with those expected. Even if incomplete lineage sorting could remain a problem, for a sequence-specific identity-based approach, the validation procedures mainly consist of showing evidence that the high observed nucleotidic identity is not the result of other mechanisms than HT, such as purifying selection or a mutational cold spot (Pace et al. 2008; Casillas et al. 2007). When dealing with genome-wide data, tools such as SCONE or the ones from the PHAST package can produce conservation tracks from multiple genome-alignment between different species (Asthana et al. 2007; Hubisz et al. 2011). However, these conservation tracks consist of quantitative scores to measure the departure from neutrality for each nucleotide, and these scores are difficult to incorporate into a statistical test to determine whether a given detected fragment is conserved or horizontally transferred.

We thus developed a more conservative approach that also accounted for non–CDSs, by subtracting the results of the *D. melanogaster–D. simulans* comparison from those retrieved

for the comparison between *D. melanogaster* and other *Drosophila* species (*D. sechellia*, *D. yakuba*, *D. pseudoobscura,* and *D. virilis*). With this comparative analysis, we discriminated between putative horizontally transferred sequences and sequences under purifying selection that are expected to be conserved across the phylogeny genome-wide for coding and noncoding DNA. The use of those four other species strengthened our results by preventing the detection of stochastic loss or ancestral polymorphisms. These two mechanisms could lead to the detection of conserved sequences between *D. melanogaster* and *D. simulans* that are absent from a third species, which is unlikely to occur simultaneously in the five analyzed species.

Finally, as this line of evidence is easily accessible and should always be considered for the study of the HT of CDSs, we also checked for CDS d_S values lower than expected given the time of divergence of the considered species. Overall, our results show an absence of HTs involving CDS between *Drosophila* species (Schaack et al. 2010), which is not caused by detection bias toward these types of sequences. Because this validation procedure is restricted to the nonrepeated content of our results, we also developed a second validation procedure to assess the TEs identified.

Validation of Horizontally Transferred TEs

To identify horizontally transferred TEs among the set of TEs with an identity higher than expected between *D. melanogaster* and *D. simulans*, we analyzed their recent dynamics since their last putative HT (Dias and Carareto 2012). The TE dynamics and maintenance in the host genome can be described as a birth-and-death processes (Schaack et al. 2010; Le Rouzic et al. 2013). The death of a TE corresponds to the inactivation of all its copies by the host defense mechanisms or the accumulation of disabling mutations (Jurka et al. 2012). On the other hand, the birth of a TE corresponds to an active copy colonizing a novel host devoid of specific transposition controls against this TE, which immediately leads to the burst of transposition of the founder copy in the new genome (Le Rouzic and Capy 2005; Naito et al. 2009). Bursts of transposition have been recorded for different TEs in numerous *Drosophila* species (García Guerreiro 2012) and can be easily identifiable because all the resulting TE copies are almost identical to each other. Afterward, most of the copies accumulate stochastic mutations and are lost over time by attrition, except for a minority of copies that can become exapted and can be identified as DNA segments conserved across species (Margulies et al. 2003; Siepel et al. 2005; Pace et al. 2008). Because TEs are likely to evolve neutrally after their insertion, we could use the neutral rate of substitution to compute the timing of a burst of transposition by calculating the pairwise divergence between all the TE copies and their consensus as an approximation of the founder copy as described in the literature (Pace and Feschotte 2007; Schaack et al. 2010; Le Rouzic et al. 2013). However, the consensus is not always a good approximation of the ancestral copy. Thus, instead of studying the complete history of a TE family with a consensus approach, our method focuses on the period of time surrounding its last putative HT between the considered species and ignores the events older than the divergence time between *D. melanogaster* and *D. simulans*. Thus, this change in the time scale provided us with a better temporal resolution for the study of the last bursts of transposition. In *D. melanogaster*, where the TE activity was recent (Bowen and McDonald 2001; Lerat et al. 2003), we were not able to clearly discriminate between the different activity periods of the TE families with an approach based on an estimated neutral mutation rate between all the copies of a TE family and their consensus sequences (Ray et al. 2008). Moreover, for the TEs with different waves of activity, such as the element *transib3* (fig. 4*C*) in the studied species, a consensus would correspond to a hypothetical copy dated in the middle of the waves of activity rather than to the ancestral copy. Our approach solves these drawbacks of consensus-based TE analysis and accounts for highly variable lengths of copies between TE families.

Another important point concerning HTs is to determine the direction of these transfers. In the cases of horizontally transferred TEs, we could expect a species with a high number of TE copies to have a higher probability to horizontally transmit one of its copies to another species, resulting in numerous identical copies in one species and few in the other. For this scenario to be valid, the transferred TEs would need to be almost instantly regulated in the receiver species to stay at a low copy number or for the receiver species to be sequenced before their burst of transposition. For both cases, these TE insertions would not have a high frequency in the species and would most likely not be observed in the sequenced strains. In an opposite scenario (a horizontally transferred TE from a species with few putatively active but controlled TE copies), a TE is transferred to a species where this TE is unknown for the host regulation system, which would lead to a burst of transposition and a quick fixation of this TE in the receiver species. Consequently, we are more likely to observe the results of this second scenario in the sequenced individuals, and we can use it to decipher the direction of detected horizontally transferred TEs (Dias and Carareto 2012).

Overall, our results show that different waves of activity seem to have occurred for different TE families and that their dynamics can be used to describe the numerous horizontally transferred TEs between *D. melanogaster* and *D. simulans*. After a horizontally transferred TE and a burst of transposition, we expect to observe a unique wave of activity before the control of the element, so the presence of other waves seems to be indicative of a complex history of the TE dynamics in *Drosophila*.

Drosophila melanogaster as a Target of Multiple Independent Horizontally Transferred TE Events

Exchange of TEs with D. simulans

In regard to the number of horizontally transferred TEs that have been detected, a parsimonious hypothesis would be their simultaneous transfer by introgression instead of independent HTs. Thus, we can wonder why no traces of introgression were detected between *D. melanogaster* and *D. simulans*, when hybrids are known to have been possible between these two species (Sawamura et al. 2004; Barbash 2010). A first genomic explanation could be that due to mutations and recombinations, the signal of an introgressed fragment of DNA has been lost over time. In this case, we can wonder why this DNA fragment would have undergone such high recombination and mutation rates, when most of the DNA is still identifiable between *D. melanogaster* and *D. simulans*.

As "nothing in evolution makes sense except in light of population genetics" (Lynch 2007), we can try to answer this question at a population level. For TEs, many steps are necessary for an HT to be successful (Le Rouzic and Capy 2005). After passing through the new host barriers, the TE must transfer itself into the germ line to be transmitted to the descendants. Then, the TE needs to have a sufficient transposition rate to propagate into the host genome and to increase in frequency in the species by vertical transmission. For TEs, which are able to actively colonize genomes, most of the TE insertions in natural populations are absent from the sequenced genome, as shown by the study of 113 *D. melanogaster* strains isolated from natural population (Kofler et al. 2012).

In the case of an introgression, all the cells in the progeny of the backcross with the hybrid will have a copy of the introgressed DNA fragment, so the first step of contaminating the germ line is always successful. Afterward, this introgressed DNA fragment has to increase in frequency in the species to be likely to be observed in the individuals actually sequenced. In contrast to the active TE copies, the introgressed fragment cannot actively replicate itself in the new genome, and its probability of fixation is simply its frequency in the population, at least in diploid organisms with a large effective population size, such as *D. melanogaster* (Nolte and Schlötterer 2008). As a result, the frequency of this introgressed fragment would be almost null in comparison to the effective population size of *D. melanogaster*, and even with the carrier subpopulation hypothesis (Jurka et al. 2011), where the population is divided into demes in each of which we can observe an effect of genetic drift that favors the fixation of low-frequency alleles, the introgressed fragment would have a low probability to be transmitted to the other demes and to be fixed in the species (Ghosh et al. 2012). Therefore, we would need to use *D. melanogaster* and *D. simulans* population-genetic data to be able to detect any traces of introgression events, as in the recent study of introgression events between *D. simulans* and *D. sechellia* from Brand et al. (2013).

This population-genetic aspect of the genomic data needs to be taken into account, as it can explain other aspects of our TE-based results. For example, the differences in the detection of horizontally transferred TEs between *D. melanogaster* and *D. simulans* found between the 2012 version of the *D. simulans* genomes sequenced from one strain (Hu et al. 2013) and the version from the 12 *Drosophila* genomes project sequenced from five different strains (Drosophila 12 Genomes Consortium 2007) can be explained by the variability of TE insertions between the populations of *D. simulans* (Vieira and Biémont 2004).

With Other Drosophila *Species*

Overall, the extensive evidence of horizontally transferred TEs detected in *D. melanogaster* seems to indicate that the fixation of new TEs could be facilitated in this genome. The timing of most HTs involving *D. melanogaster* was estimated between 1.4 and 2.3 Ma, before the worldwide expansion of *D. melanogaster* and *D. simulans* that happened 15,000 years ago (Stephan and Li 2007; Cararetto 2011). The *melanogaster* subgroup is endemic to Afrotropical regions, with the proto-melanogaster founder dated between 17 and 20 Ma from the oriental region of Africa (Lachaise et al. 1988). Thus, a parsimonious hypothesis for the numerous horizontally transferred TEs detected among *D. melanogaster*, *D. simulans*, *D. sechellia,* and *D. yakuba* would place them at a time when these species were all living in Africa, before the worldwide expansion of *D. melanogaster* and *D. simulans*. In this scenario, there would have been fewer geographical barriers to hamper the fixation of horizontally transferred TEs into sympatric populations with a smaller repartition area than the worldwide populations of *D. melanogaster*. The arrival of these new TE copies in the genome *D. melanogaster* may have been a springboard for the worldwide expansion of this species, as the load of TEs can be correlated with the colonization of new territory (Vieira et al. 1999, 2002). In contrast, a stronger population structure in *D. simulans* (Mousset and Derome 2004) could explain the polymorphisms of TE insertion that have resulted in different TE loads between populations (Vieira and Biémont 2004) and that may have independently favored the worldwide expansion of this species, even if in both cases the cause of such a mechanism is not yet understood.

Conclusions

We have developed a novel approach for the genome-wide detection of all putative HT sequences independently of their coding capability between two genomes. Our method relies on a well-defined testing framework to approach this genome-wide problem as a multiple-testing problem. We successfully applied this method between the genomes of *D. melanogaster* and *D. simulans*, underscoring the sensitivity

of our approach to detect HTs between closely related species. Like previous studies of HTs in eukaryotes, we validated these results with other lines of evidence. We also proposed two novel approaches to remove bias due to the detection of conserved sequences, by a comparative analysis with phylogenetically related species in the case of CDS and non–CDSs and by an analysis of their recent activity in the case of TEs. After these validation steps, we retrieved all the horizontally transferred TEs previously described in different studies (see Cararetto [2011] for a review) and very few spurious CDS, attesting to the sensitivity and the specificity of our approach.

By a manual analysis of our results along each chromosome arm of *D. melanogaster*, we did not detect any trace of introgression between *D. melanogaster* and *D. simulans*, even if this does not completely rule out this hypothesis. We also detected a large number of horizontally transferred TEs involving *D. melanogaster* and other *Drosophila* species with our assessment steps, bringing to light a small portion of the network of horizontally transferred TEs in this phylogeny. This large number of HTs for different TE families also supports the model of birth and death, where HT events are a vital part of the TE life cycle that prevents their extinction (Schaack et al. 2010). We are just beginning to understand the complex horizontally transferred TE network in eukaryotes, and our approach could be applied to any pair of sequenced species to increase our knowledge of the dynamics of these sequences, which seem to jump both within and between species.

Supplementary Material

Supplementary figure S1–S7 and tables S1 and S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank C. Cararetto for her critical reading of the manuscript. The English of the manuscript has been edited by the American Journal Experts company. This work was supported by the ANR-09-BLANC-0103-01.

Literature Cited

Abby SS, Tannier E, Gouy M, Daubin V. 2012. Lateral gene transfer as a support for the tree of life. Proc Natl Acad Sci U S A. 109:4962–4967.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Andersson JO. 2005. Lateral gene transfer in eukaryotes. Cell Mol Life Sci. 62:1182–1197.

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

Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S. 2007. Analysis of sequence conservation at nucleotide resolution. PLoS Comput Biol. 3:e254.

Azad RK, Lawrence JG. 2011. Towards more robust methods of alien gene detection. Nucleic Acids Res. 39:e56.

Barbash DA. 2010. Ninety years of *Drosophila melanogaster* hybrids. Genetics 186:1–8.

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

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc. 57:289–300.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. Ann Statist. 10:467–498.

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

Brand CL, Kingan SB, Wu L, Garrigan D. 2013. A selective sweep across species boundaries in *Drosophila*. Mol Biol Evol. 30:2177–2186.

Capy P, Anxolabéhère D, Langin T. 1994. The strange phylogenies of transposable elements: are horizontal transfers the only explanation? Trends Genet. 10:7–12.

Cararetto CM. 2011. Tropical Africa as a cradle for horizontal transfers of transposable elements between species of the genera *Drosophila* and *Zaprionus*. Mob Genet Elem. 1:179–186.

Casillas S, Barbadilla A, Bergman CM. 2007. Purifying selection maintains highly conserved noncoding sequences in *Drosophila*. Mol Biol Evol. 24:2222–2234.

Daniels SB, et al. 1990. Evidence for horizontal transmission of the *P* transposable element between *Drosophila* species. Genetics 124: 339–355.

de Carvalho MO, Loreto ELS. 2012. Methods for detection of horizontal transfer of transposable elements in complete genomes. Genet Mol Biol. 35:1078–1084.

de Koning AP, et al. 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 7:e1002384.

de la Chaux N, Wagner A. 2009. Evolutionary dynamics of the LTR retrotransposons roo and rooA inferred from twelve complete *Drosophila* genomes. BMC Evol Biol. 9:205.

Dias ES, Cararetto CMA. 2012. Ancestral polymorphism and recent invasion of transposable elements in *Drosophila* species. BMC Evol Biol. 12: 119.

Dowsett AE, Young MW. 1982. Differing levels of dispersed repetitive DNA among closely related species of *Drosophila*. Proc Natl Acad Sci U S A. 79:4570–4574.

Doyon JP, Ranwez V, Daubin V, Berry V. 2011. Models, algorithms and programs for phylogeny reconciliation. Brief Bioinform. 12:392–400.

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

Dupuy C, et al. 2011. Transfer of a chromosomal Maverick to endogenous bracovirus in a parasitoid wasp. Genetica 139:489–496.

Efron B, Tibshirani R, Storey JD, Tusher V. 2001. Empirical Bayes analysis of a microarray experiment. J Am Statist Assoc. 96:1151–1160.

Fall S, et al. 2007. Horizontal gene transfer regulation in bacteria as a "spandrel" of DNA repair mechanisms. PLoS One 2:e1055.

Finnegan DJ. 1997. Transposable elements: how non-LTR retrotransposons do it. Curr Biol. 7:R245–R248.

Finner H, Roters M. 2002. Multiple hypotheses testing and expected number of type I. errors. Ann Statist. 30:220–238.

García Guerreiro MP. 2012. What makes transposable elements move in the *Drosophila* genome? Heredity 108:461–468.

Ghosh A, Meirmans PG, Haccou P. 2012. Quantifying introgression risk with realistic population genetics. Proc Biol Sci. 279:4747–4754.

Gilbert C, Cordaux R. 2013. Horizontal transfer and evolution of prokaryote transposable elements in eukaryotes. Genome Biol Evol. 5: 822–832.

Gilbert C, Pace JK, Feschotte C. 2009. Horizontal spinning of transposons. Commun Integr Biol. 2:117–119.
Gilbert C, Waters P, Feschotte C, Schaack S. 2013. Horizontal transfer of OC1 transposons in the Tasmanian devil. BMC Genomics 14:134.
Gilbert C, et al. 2010. A role for host-parasite interactions in the horizontal transfer of transposons across phyla. Nature 464:1347–1350.
Granzotto A, Lopes FR, Vieira C, Carareto CMA. 2011. Vertical inheritance and bursts of transposition have shaped the evolution of the BS non-LTR retrotransposon in *Drosophila*. Mol Genet Genomics. 286:57–66.
Gronau I, Arbiza L, Mohammed J, Siepel A. 2013. Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. Mol Biol Evol. 30:1159–1171.
Heger A, Ponting CP. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes. Genome Res. 17:1837–1849.
Hu TT, Eisen MB, Thornton KR, Andolfatto P. 2013. A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. Genome Res. 23:89–98.
Hubisz MJ, Pollard KS, Siepel A. 2011. PHAST and RPHAST: phylogenetic analysis with space/time models. Brief Bioinform. 12:41–51.
Juhas M, et al. 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev. 33:376–393.
Jurka J, Bao W, Kojima KK. 2011. Families of transposable elements, population structure and the origin of species. Biol Direct. 6:44.
Jurka J, et al. 2012. Distinct groups of repetitive families preserved in mammals correspond to different periods of regulatory innovations in vertebrates. Biol Direct. 7:36.
Kaminker JS, et al. 2002. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biol. 3: RESEARCH0084.
Kapitonov VV, Jurka J. 2003. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. Proc Natl Acad Sci U S A. 100:6569–74.
Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet. 9:605–618.
Kim J, et al. 1994. Rodent *BC1* RNA gene as a master gene for *ID* element amplification. Proc Natl Acad Sci U S A. 91:3607–3611.
Kofler R, Betancourt AJ, Schlötterer C, Schlo C. 2012. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. PLoS Genet. 8: e1002487.
Kuan PF, Chiang DY. 2012. Integrating prior knowledge in multiple testing under dependence with applications to detecting differential DNA methylation. Biometrics 68:774–783.
Lachaise D, et al.. 1988. Historical biogeography of the *Drosophila melanogaster* species subgroup. In: Hecht MK, Wallace B, Prance GT, editors. Evolutionary biology. Springer. p. 159–225.
Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.
Le Rouzic A, Capy P. 2005. The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. Genetics 169:1033–1043.
Le Rouzic A, Payen T, Hua-Van A. 2013. Reconstructing the evolutionary history of transposable elements. Genome Biol Evol. 5:77–86.
Lerat E, Burlet N, Biémont C, Vieira C. 2011. Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. Gene 473:100–109.
Lerat E, Rizzon C, Biémont C. 2003. Sequence divergence within transposable element families in the *Drosophila melanogaster* genome. Genome Res. 13:1889–1896.
Lisch D. 2008. A new *SPIN* on horizontal transfer. Proc Natl Acad Sci U S A. 105:16827–16828.
Loreto ELS, Carareto CMA, Capy P. 2008. Revisiting horizontal transfer of transposable elements in *Drosophila*. Heredity 100:545–554.
Ludwig A, Valente VL, Loreto EL. 2008. Multiple invasions of Errantivirus in the genus *Drosophila*. Insect Mol Biol. 17:113–124.
Lynch M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci U S A. 104:8597–8604.
Margulies EH, Blanchette M, Haussler D, Green ED. 2003. Identification and characterization of multi-species conserved sequences. Genome Res. 13:2507–2518.
Marygold SJ, et al. 2013. Flybase: improvements to the bibliography. Nucleic Acids Res. 41:751–757.
Mousset S, Derome N. 2004. Molecular polymorphism in *Drosophila melanogaster* and d. simulans: what have we learned from recent studies? Genetica 120:79–86.
Naito K, et al. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature 461: 1130–1134.
Nolte V, Schlötterer C. 2008. African *Drosophila melanogaster* and *D. simulans* populations have similar levels of sequence variability, suggesting comparable effective population sizes. Genetics 178:405–412.
Novăk E, Neumann P, Macas J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11:378.
O'Brochta DA, et al. 2009. Transpositionally active episomal *hAT* elements. BMC Mol Biol. 10:108.
Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304.
Pace JK, Feschotte C. 2007. The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. Genome Res. 17:422–432.
Pace JK, Gilbert C, Clark MS, Feschotte C. 2008. Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. Proc Natl Acad Sci U S A. 105:17023–17028.
Pinsker W, Haring E, Hagemann S, Miller WJ. 2001. The evolutionary life history of *P* transposons: from horizontal invaders to domesticated neogenes. Chromosoma 110:148–158.
Podell S, Gaasterland T. 2007. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. Genome Biol. 8:R16.
Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20:110–121.
Putonti C, et al. 2006. A computational tool for the genomic identification of regions of unusual compositional properties and its utilization in the detection of horizontally transferred sequences. Mol Biol Evol. 23: 1863–1868.
Quinlan AR, Hall IM. 2010. Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.
R Core Team. 2013. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Ragan MA. 2001. Detection of lateral gene transfer among microbial genomes. Curr Opin Genet Dev. 11:620–626.
Ray DA, et al. 2008. Multiple waves of recent DNA transposon activity in the bat, *Myotis lucifugus*. Genome Res. 18:717–728.
Roger A. 1999. Reconstructing early events in eukaryotic evolution. Am Nat. 154:S146–S163.
Sawamura K, Karr TL, Yamamoto MT. 2004. Genetics of hybrid inviability and sterility in *Drosophila*: dissection of introgression of *D. simulans* genes in *D. melanogaster* genome. Genetica 120:253–260.
Schaack S, Gilbert C, Feschotte C. 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol Evol. 25:537–546.
Schnable PS, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115.
Shi SY, Cai XH, Ding DF. 2005. Identification and categorization of horizontally transferred genes in prokaryotic genomes. Acta Biochim Biophys Sin. 37:561–566.
Siepel A, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.

Silva JC, Loreto EL, Clark JB. 2004. Factors that affect the horizontal transfer of transposable elements. Curr Issues Mol Biol. 6: 57–71.

Stephan W, Li H. 2007. The recent demographic and adaptive history of *Drosophila melanogaster*. Heredity 98:65–68.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 100:9440–9445.

Sun W, Tony Cai T. 2009. Large-scale multiple testing under dependence. J Roy Statist Soc. 71:393–424.

Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. Mol Biol Evol. 21: 36–44.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. Brief Bioinform. 14:178–192.

Vieira C, Biémont C. 1997. Transposition rate of the *412* retrotransposable element is independent of copy number in natural populations of *Drosophila simulans*. Mol Biol Evol. 14:185–188.

Vieira C, Biémont C. 2004. Transposable element dynamics in two sibling species: *Drosophila melanogaster* and *Drosophila simulans*. Genetica 120:115–123.

Vieira C, Lepetit D, Dumont S, Biémont C. 1999. Wake up of transposable elements following *Drosophila simulans* worldwide colonization. Mol Biol Evol. 16:1251–1255.

Vieira C, et al. 2002. Evolution of genome size in *Drosophila*.. Is the invader's genome being invaded by transposable elements? Mol Biol Evol. 19:1154–1161.

Wallau GL, Ortiz MF, Loreto ELS. 2012. Horizontal transposon transfer in eukarya: detection, bias, and perspectives. Genome Biol Evol. 4: 689–699.

Wei Z, Sun W, Wang K, Hakonarson H. 2009. Multiple testing in genome-wide association studies via hidden Markov models. Bioinformatics 25: 2802–2808.

Weinert LA, Welch JJ, Jiggins FM. 2009. Conjugation genes are common throughout the genus *Rickettsia* and are transmitted horizontally. Proc R Soc B. 276:3619–3627.

Weir JT, Schluter D. 2008. Calibrating the avian molecular clock. Mol Ecol. 17:2321–2328.

Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973–982.

Associate editor: Josefa Gonzalez

3

Correction de tests multiples unilatéraux pour des applications génomiques

Comme nous l'avons vu dans le chapitre précédent, les analyses bioinformatiques de données génomiques sont souvent liées à des problèmes statistiques. Par exemple, pour mesurer la significativité d'un résultat, la notion de test statistique est essentiellei en science. Pour définir un test statistique, il faut commencer par définir une hypothèse nulle, ou hypothèse d'absence d'effet. Ensuite les données observées pourront être confrontées aux valeurs attendues sous cette hypothèse nulle pour déterminer si l'on peut considérer raisonnablement qu'elles correspondent à cette hypothèse ou qu'elles en diffèrent.

1 Problématique des tests multiples

Pour une mesure observée, on peut calculer une statistique de tests associée z qui est la réalisation d'une variable aléatoire continue Z sur $\mathbb{R}^+$ dont on connaît la distribution sous l'hypothèse nulle $\mathcal{H}_0$. On peut donc calculer la probabilité d'observer une valeur au moins aussi extrême que z sous $\mathcal{H}_0$, $x = \Pr(Z \geq z|\mathcal{H}_0)$. Cette probabilité est appelée probabilité critique ou p-value. La probabilité critique peut aussi être décrite comme la proportion de fois où des valeurs plus extrêmes ou aussi extrêmes que z seraient observées si l'hypothèse nulle était vraie et si on répétait l'expérience un grand nombre de fois. Pour décider si une mesure observée est probable sous $\mathcal{H}_0$, il suffit de comparer la p-value

associée à un seuil α, par exemple $\alpha = 0,1$. Si la p-value est inférieure au seuil choisi, on rejète $\mathcal{H}_0$ en considérant que cette valeur est improbable sous cette hypothèse. En effet, la probabilité d'observer des valeurs au moins aussi divergentes de l'hypothèse nulle est inférieure ou égale à 0,1. On note que sous $\mathcal{H}_0$, nous avons bien une probabilité de 0,1 d'observer des valeurs au moins aussi divergentes de l'hypothèse nulle. En prenant la décision de rejeter $\mathcal{H}_0$, nous avons donc un risque de 0,1 de prendre cette décision à tort (erreur de type I). On dit que le risque de rejeter $\mathcal{H}_0$ à tort est contrôlé à un niveau $\alpha = 0,1$.

La notion de contrôle du risque dans les tests d'hypothèse est au cœur de la problématique des tests multiples. En effet, dans le cas d'un test statistique, si l'hypothèse nulle est bien définie, nous pouvons choisir un seuil α qui va contrôler notre probabilité de rejeter $\mathcal{H}_0$ à tort. Cependant, si l'on effectue plusieurs tests simultanément, ce risque va être multiplié par le nombre de tests m. Ainsi, en testant m hypothèses et en rejetant $\mathcal{H}_0$ pour les hypothèses dont la p-value associée est inférieure à α, la probabilité de rejeter par erreur $\mathcal{H}_0$ au moins une fois devient supérieure à α. Il faut donc définir le risque que l'on veut controler. Si l'on veut maintenir un contrôle de la probabilité de faire moins de un rejet de $\mathcal{H}_0$ à tort pour m hypothèses à un niveau α, il faut changer de règle de décision. On peut ainsi appliquer la procédure de Bonferroni en rejetant toutes les p-values inférieures ou égales à α/m et maintenir ce contrôle. Bien que parfaitement valide, ce type de contrôle, appelé "family wise error rate" ou *FWER*, conduit à ne rejeter qu'un petit nombre restreint d'hypothèses pour ne pas prendre le risque d'en rejeter une à tort. Cette stratégie peut vite sembler déraisonnable quand le nombre d'hypothèses testées est grand. Ce cas de figure est souvent rencontré en bioinformatique, notamment avec le traitement de données de séquençage de nouvelle génération (NGS), que nous avons présenté dans le chapitre 1. Par exemple, ce type de contrôle peut conduire à ne pas déclarer comme exprimés de façon différentielle entre différentes conditions, des centaines voire des milliers de gènes, pour ne pas prendre le risque de se tromper une fois pour l'un d'eux. Par conséquent, si le contrôle du *FWER* peut correspondre à différentes problématiques, comme pour certaines études médicales pour lesquelles on veut être sûr de ne pas se tromper, ce contrôle peut sembler trop strict pour nombre d'analyses exploratoires en biologie. Il pourrait sembler plus raisonnable par exemple d'utiliser une règle de décision permettant de tolérer plus de faux positifs quand le nombre de positifs est grand.

2 Contrôle du *FDR*

Ce type de contrôle a été introduit par Benjamini et Hochberg [216], avec une procédure permettant le contrôle du nombre moyen de taux de faux positifs (*FDR* pour l'anglais "False Discovery Rate"). Ainsi, quand le nombre de rejets de $\mathcal{H}_0$ est grand, on accepte de rejeter à tort en moyenne au plus un certain pourcentage de ces hypothèses. On peut remarquer que dans le cas où une seule hypothèse est rejetée, un contrôle du *FDR* à un niveau de 0,1 contrôle aussi le *FWER* à un niveau de 0,1. Par contre, à mesure que le nombre d'hypothèses rejetées augmente, la différence entre contrôle du *FDR* et du *FWER* grandit, avec de plus en plus de faux positifs (mais en pourcentage constant en moyenne du nombre total d'hypothèses rejetées) pour un contrôle du *FDR*.

La problématique des tests multiples a été un sujet de recherche très actif ces 20 dernières années, notamment concernant l'amélioration du contrôle du *FDR*. En effet, même si la procédure proposée par Benjamini et Hochberg [216] est valide pour le contrôle du *FDR*, il est possible d'obtenir de meilleures performances en terme de nombre de vrais positifs détectés pour un même contrôle du *FDR*. Par exemple, quand le nombre d'hypothèses est assez grand, il est possible d'utiliser des approches dites de contrôle du *FDR* local (ℓ*FDR*) qui permettent une meilleure discrimination entre les p-values associées avec l'hypothèse nulle $\mathcal{H}_0$ et celles associées avec l'hypothèse alternative $\mathcal{H}_1$ pour lesquelles on souhaite pouvoir rejeter $\mathcal{H}_0$ [217]. Ce type de méthode se base sur le fait que la distribution des p-values sous $\mathcal{H}_0$ est connue et suit une loi uniforme entre 0 et 1 dans le cas où Z est continue sous $\mathcal{H}_0$. En effet, avec $x = \Pr(Z \geq z|\mathcal{H}_0) = 1 - \Pr(Z < z|\mathcal{H}_0)$ et $x = 1–F_0(z)$ avec F_0 la fonction de distribution de fréquence cumulée (cdf pour l'anglais "cumulative density function") de z sous $\mathcal{H}_0$. La cdf est la fonction qui présente, pour chaque valeur possible d'une variable aléatoire, la probabilité que cette variable aléatoire prenne une valeur inférieure ou égale à celle-ci. Comme une cdf est croissante et monotone, nous pouvons écrire :

$$\Pr(Z \geq z|\mathcal{H}_0) = \Pr(F_0(Z) \geq F_0(z)) = 1 - \Pr(F_0(Z) < F_0(z)) \tag{3.1}$$

Par conséquent, nous avons $\Pr(F_0(Z) < F_0(z)) = F_0(z)$ ce qui veut dire que $F_0(z)$ est une distribution uniforme, et que $1 - F_0(z)$ est aussi une distribution uniforme. Les p-values suivent donc une distribution uniforme sous l'hypothèse nulle.

Comme nous l'avons vu, la p-value peut être considérée comme une mesure de l'écart de l'observation avec ce que l'on attend sous $\mathcal{H}_0$. Ainsi, les observations moins probables sous $\mathcal{H}_0$ sont associées avec une p-value plus petite. Par conséquent, les p-values associées avec l'hypothèse alternative vont suivre une distribution avec une forte densité

proche de zéro et dont la densité décroit avec les p-values. Cette distribution est cependant inconnue puisque seule la distribution de Z sous $\mathcal{H}_0$ est définie dans le test. Les approches de type ℓFDR tirent profit de ces deux distributions pour traiter la problématique des tests multiples comme un problème de classification non supervisée où les p-values peuvent provenir de deux distributions différentes, une uniforme et une inconnue. Ce type de méthode permet de calculer la probabilité pour chaque p-value de correspondre à l'hypothèse nulle ou à l'hypothèse alternative en prenant en compte la forme de la distribution de toutes les p-values sous chacune de ces deux hypothèses. Ce type de procédure conduit à une meilleure classification entre les p-values associées à $\mathcal{H}_0$ et celles associées à $\mathcal{H}_1$ [218].

La figure 3.1 représente les résultats de la réalisation de multiples tests statistiques bilatéraux. Le cas présenté est un cas simple de tests pour lesquels z suit une loi Normale centrée réduite sous $\mathcal{H}_0$. Dans le graphique du haut on peut voir la transformation de z-values (en abscisse) en p-values (en ordonnées) par la cdf des z-values sous $\mathcal{H}_0$. Le graphique du bas est un histogramme des p-values observées sur cet axe des abscisses.

3 Les tests multiples unilatéraux

S'il existe maintenant une littérature importante sur le contrôle du FDR pour les tests multiples que nous avons décrit, la problématique des tests multiples unilatéraux reste peu étudiée. Il existe cependant des différences fondamentales entre le comportement d'un test bilatéral, où la p-value mesure un écart à une valeur théorique, et un test unilatéral, où la p-value mesure un écart dans une seule direction par rapport à une valeur théorique. Ainsi pour les tests unilatéraux, les p-values associées à $\mathcal{H}_0$ peuvent aussi correspondre à des écarts par rapport à une valeur théorique dans une autre direction que celle étudiée. Comme la distribution sous l'hypothèse nulle du test ne décrit pas plus la distribution de ces mesures que celle correspondant à un écart dans le sens d'intérêt, nous avons donc deux distributions inconnues de p-values pour les tests multiples unilatéraux.

Comme pour les tests bilatéraux, les p-values associées à $\mathcal{H}_1$ vont suivre une distribution inconnue avec une forte densité proche de zéro dont la densité décroit avec les p-values. Les p-values correspondant à des mesures dont l'écart à la valeur théorique sous $\mathcal{H}_0$ n'est pas dans la direction étudiée vont quant à elles, suivre une distribution inconnue avec une forte densité près de 1 qui croît avec les p-values. Dans le chapitre précédent, concernant la problématique des tests multiples unilatéraux, les tests ont été définis pour mesurer une divergence nucléotidique plus élevée que celle attendue entre deux séquences d'ADN entre *Drosophila melanogaster* et *D. simulans*. Dans ce cas, toutes les paires de

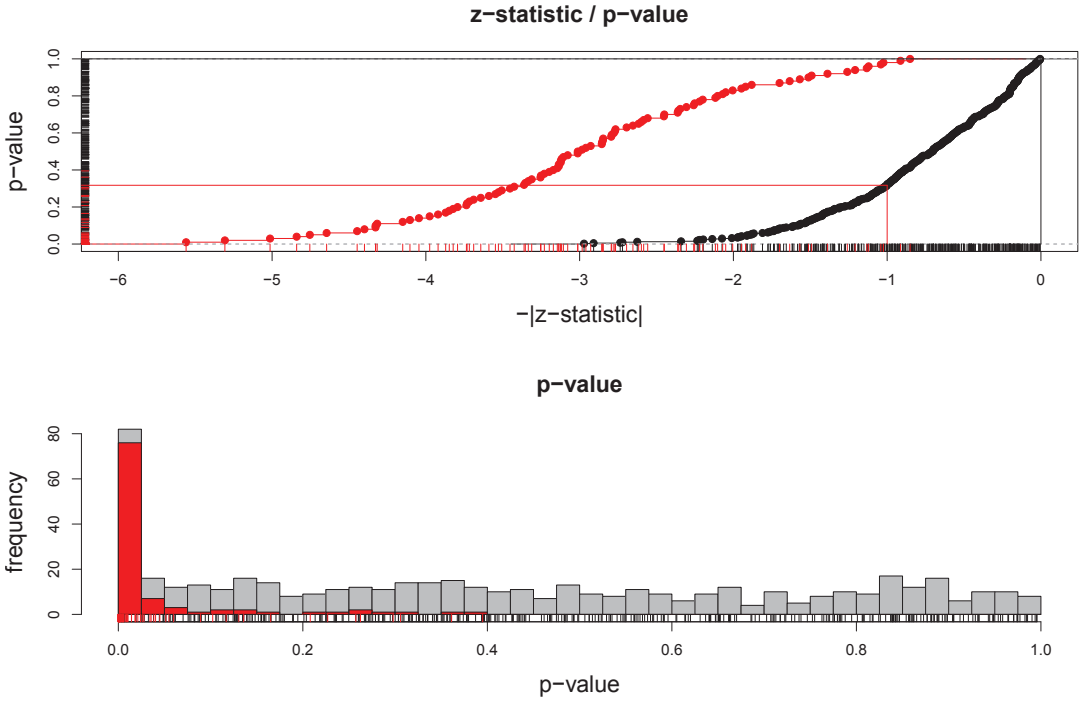


FIGURE 3.1 : Représentation de tests bilatéraux multiples. Les statistiques z et les p-values associées à l'hypothèse nulle $\mathcal{H}_0$ sont représentées en noir alors que celles associées à $\mathcal{H}_1$ sont représentées en rouge. Pour le graphique du haut, les trais représentent la cdf des différentes populations de z et les traits en abscisse et en ordonnée représentent les valeurs prises par la statistique et les p-values. Le graphique du bas représente la contribution des p-values associées à $\mathcal{H}_0$ en gris, ainsi que celles associées à $\mathcal{H}_1$ en rouge, à la distribution des p-values.

séquences entre ces deux espèces ayant accumulé plus de mutations que l'attendu vont représenter un excès de densité pour des valeurs faibles d'identité nucléotidique, ce qui va se traduire par un excès de p-values proches de 1.

La Figure 3.2 représente les résultats de la réalisation de multiples tests statistiques unilatéraux. Cette figure est similaire à la Figure 3.1 avec l'ajout de la contribution des statistiques z mesurant un écart à $\mathcal{H}_0$ dans une autre direction que celle étudiée. On peut voir que la présence de ces mesures provoque un excès de p-values proche de 1 sur la distribution des p-values.

S'il n'est pas pris en compte, cet excès de p-values proches de 1 peut conduire à de mauvaises estimations des paramètres des modèles utilisés par les procédures de contrôle du *FDR* [219]. C'est pourquoi au cours de ma thèse j'ai travaillé à l'élaboration d'une procédure permettant d'appliquer les approches de classification non supervisée au contrôle du *FDR* aux tests multiples unilatéraux. Ce travail a été effectué sous la supervision de Franck Picard avec la collaboration d'Alain Celisse (Laboratoire Paul Painlevé, UMR 8524 CNRS-Université Lille 1) et a conduit à la rédaction d'un article, dont le manuscrit correspond au corps de ce chapitre, intitulé :

- The unilateral side of multiple-testing: an l*FDR* application

Ce travail a par ailleurs été présenté sous la forme d'un poster lors de la conférence internationale SMPGD à Amsterdam en 2013 intitulé :

- Zero-inflated Gaussian hidden Markov models for multiple testing under genomic dependencies

La procédure que nous avons développée permet d'estimer la position d'un point μ dans une distribution de p-values unilatérales au delà duquel les p-values associées avec $\mathcal{H}_0$ ne suivent plus une distribution uniforme. Cette estimation est faite par une approche de validation croisée, pour sélectionner le meilleur histogramme décrivant la distribution des p-values, qui a été développée par Alain Celisse. Cette étape permet de circonvenir la difficulté d'une approche de classification non supervisée pour un modèle à 3 compartiments (les p-values associées à $\mathcal{H}_1$ et les deux types de p-values associées à $\mathcal{H}_0$), dont deux suivent des distributions inconnues, ce qui est un problème difficilement identifiable.

J'ai ensuite utilisé l'estimation de ce point μ pour proposer une transformation de la distribution des p-values en z-values dont la densité peut être décrite comme un modèle de mélange à trois compartiments donc deux sont connus. Ce modèle de mélange permet donc de calculer la probabilité pour chaque z-value d'être associée à $\mathcal{H}_0$, et de contrôler

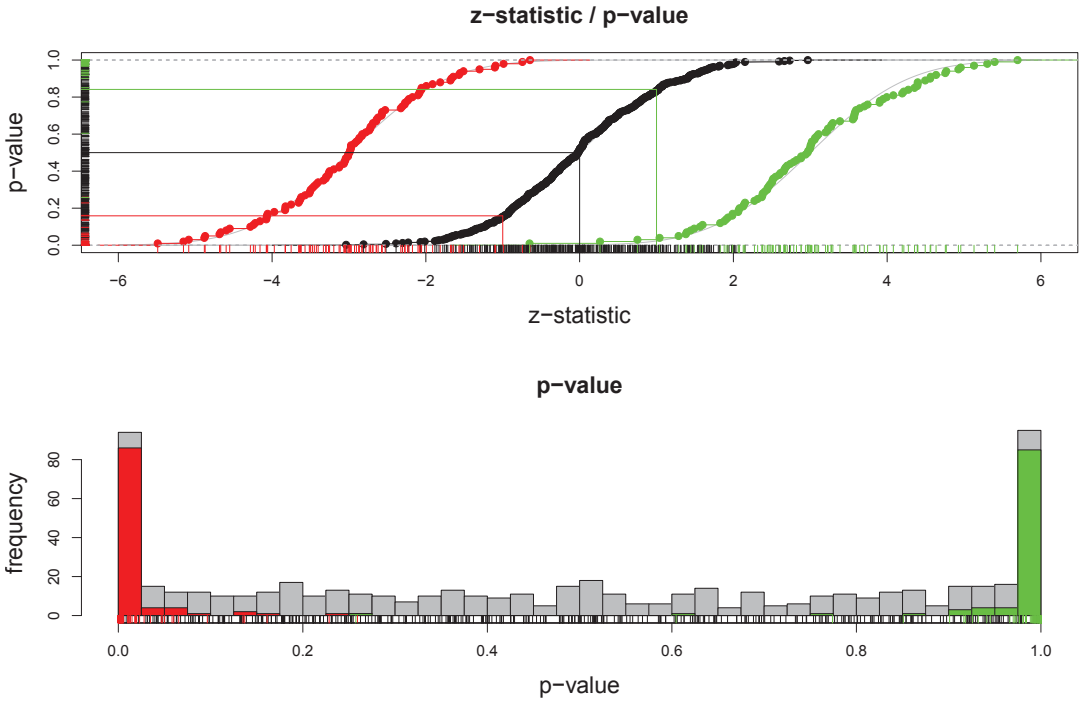


FIGURE 3.2 : Représentation de tests unilatéraux multiples. Les statistiques z et les p-values associées à l'hypothèse nulle $\mathcal{H}_0$ sont représentées en noir alors que celles associées à $\mathcal{H}_1$ sont représentées en rouge. La sous-population des statistiques et p-values correspondant à un écart à $\mathcal{H}_0$ dans une autre direction que celle étudiée est présentée en vert. Pour le graphique du haut, les trais représentent la cdf des différentes population de z et les traits en abscisse et en ordonnée représentent les valeurs prises par la statistique et les p-values. Le graphique du bas représente la contribution des p-values associées à $\mathcal{H}_0$ en gris et vert, ainsi que celles associées à $\mathcal{H}_1$ en rouge, à la distribution des p-values.

le *FDR* à un niveau α donné pour des tests multiples unilatéraux. J'ai implémenté cette procédure dans un package R, uni*FDR*, pour différents modèles de dépendance entre les hypothèses testées, qui sont respectivement un modèle sans dépendance, avec une dépendance Markovienne homogène et avec une dépendance Markovienne non-homogène. Comme nous l'avons vu dans le chapitre précédent, la prise en compte de la dépendance qui existe entre les hypothèses testées peut permettre d'augmenter le nombre de vrais positifs pour un même contrôle du *FDR*.

Pour conclure, les gains apportés par l'utilisation de notre procédure pour la correction des tests multiples unilatéraux pourraient conduire à de nouvelles découvertes en biologie, pour les cas où la problématique testée est fondamentalement unilatérale comme nous l'avons vu dans le chapitre précédent, ou pour les cas où le test utilisé est unilatéral comme pour les tests de taux de vraisemblance par exemple.

The unilateral side of multiple-testing : an ℓFDR application

Laurent Modolo, Alain Celisse, Emmanuelle Lerat et Franck Picard

en préparation

The unilateral side of multiple-testing: an ℓFDR application

Laurent Modolo, Alain Celisse, Emmanuelle Lerat and Franck Picard

October 17, 2014

1 INTRODUCTION

In the era of next generation sequencing (NGS) data, we often deal with hundreds even thousands of simultaneous hypothesis testings. For example in differential gene expression analyses, expression levels of thousands of genes are compared under different conditions. Genome-wide association studies constitute another example where the association of a subset of genetic markers associated with a disease or a trait is investigated among hundreds of thousands makers (Storey and Tibshirani, 2003). Therefore, such analyses are closely linked to multiple-testing (Dudoit and van der Laan, 2008). Classically, the goal of statistical testing has been to avoid making one or more type-I errors (wrongly rejecting the null hypothesis or false positive) and this type of control has been naturally extended to multiple-testing with the family wise error rate control ($FWER$) proposed by Bonferroni (Hochberg, 1988). However, the $FWER$ control leads to a drastic loss in power and such procedure only returns a small subset of true positives. To increase the power of multiple-testing correction, Benjamini and Hochberg (1995) proposed a control of the false discovery rate (FDR), that is controlling the proportion of false positives (FP) for a potentially large number of rejections (R). The FDR is defined by:

$$FDR = \mathbb{E}\left[\frac{FP}{R} \,\middle|\, R > 0\right] \Pr\left(R > 0\right). \tag{1}$$

Another metric for multiple-testing procedure is the false non-discovery rate (FNR), introduced by Genovese and Wasserman (2002) and defined by:

$$FNR = \mathbb{E}\left[\frac{FN}{S} \,\middle|\, S > 0\right] \Pr\left(S > 0\right). \tag{2}$$

The different outcomes of multiple-testing procedures are summarized in Table 1. A FDR procedure is valid if it provides a strong control of the FDR at a nominal level α, and optimal if it has the smallest FNR among all the valid FDR procedures (Sun and Tony Cai, 2009). Similarly to the draw-backs of $FWER$ control, an FDR procedure that is valid but non-optimal will only return a subset of true positives potentially hiding important results from the analysis.

One of the main advantage of mutliple-testing procedures is their hability to work directly with p-values regardless of the underlying test statistic. However, when confronted with a

Table 1: Classification of tested hypotheses

Hypothesis	Claimed non-significant	Claimed significant	Total
Null	TN	FP	m_0
Non-null	FN	TP	m_1
Total	S	R	m

multiple-testing problem the tests can be either two-sided, when the tested hypothesis is bilateral, or one-sided, when the tested hypothesis is unilateral. In practice, there is a wide range of common applications, such as the likelihood ratio test (Anders and Huber, 2010) or when the problematic is fundamentally unilateral (Modolo et al., 2014), where the hypotheses tested are unilateral. However, while there is a large literature for the control of the FDR for two-sided multiple-testing, the one-sided multiple-testing problem has been mostly overlooked (Pounds and Cheng, 2006).

There are some fondamental differences between one-sided and two-sided hypothesis testing that render the assumption made by bilateral multiple-testing procedures invalid and could lead to poor performance of the FDR procedures when applied to one-sided hypothesis testing (Pounds and Cheng, 2006). To intruduce theses differences, let us first consider the general framework introduced by Efron et al. (2001) for two-sided multiple-testing. The p-values are a random vector X of m random variables. The first assumption is that their are two population of p-values in proportion π_0 and $(1-\pi_0)$ with π_0 the proportion of p-values associated with the null hypothesis. The probability density distribution (pdf) of X, $g(x_i), i = \{1, \ldots, m\}$, can be modeled as the following two-component mixture:

$$g(x) = (1-\pi_0)\, g_1(x) + \pi_0 g_0(x), \forall x \in [0,1], \tag{3}$$

with $g_0(x)$ the pdf the p-values associated with the null hypothesis and $g_1(x)$ the pdf of the p-values associated with the alternative hypothesis. Most of the bilateral multiple-testing procedures rely on the key assumption that $g_0(x) = \mathcal{U}(]0,1])$ when X is continuous (Casella and Berger, 1990). However, observations show that for unilateral multiple-testing this key assumption do not hold. In one-sided multiple-testing the p-values associated with the null hypothesis can arise from a uniform distribution, or can be stochastically higher than uniform, if they correspond to the non-tested hypothesis.

Figure 1 presents a comparison between simple multiple z-tests in the bilateral case ($\mathcal{H}_0$: the mean is equal to zero) and in the unilateral case ($\mathcal{H}_0$: the mean is higher or equal to zero). In the bilateral case, we have $g_0(x) = \mathcal{U}(]0,1])$ for the p-values associated with the non-tested hypotheses, while we clearly don't have an uniform distribution of the p-values associated with the non-tested hypotheses in the unilateral case. Therefore, when building a multiple-testing procedure for unilateral hypothesis, one has to account for the excess of p-values corresponding to the non-tested hypothesis that do not follow a uniform distribution (Han et al., 2011).

One widely used multiple-testing procedure developed for bilateral multiple-testing is the Benjamini & Hochberg procedure (BH) (Benjamini and Hochberg, 1995). With $x_{(1)}, \ldots, x_{(m)}$ the ordered m p-values and $\mathcal{H}_{(1)}, \ldots, \mathcal{H}_{(m)}$ the corresponding hypotheses, the BH procedure is valid for a FDR level α (Benjamini and Yekutieli, 2001; Genovese and Wasserman, 2002):

$$k = max\left\{ i : x_{(i)} \leq \frac{i}{m}\frac{\alpha}{\pi_0} \right\}, \text{ then reject all } \mathcal{H}_{(i)}, 1, \ldots, k. \tag{4}$$

The original BH procedure provides a conservative control for the FDR at a level $\alpha\pi_0$ for bilateral multiple-testing (by setting $\pi_0 = 1$). This control level can be closer to α with plug-in procedures using an estimate of π_0 (Benjamini and Yekutieli, 2001; Genovese and Wasserman, 2002).

To illustrate the problem of the procedure developed for bilateral multiple-testing with the BH procedure, we can rewrite (4) under the framework (3) (Liu et al., 2014) :

$$k = max\left\{ i : \frac{\pi_0 F_0(T_i)}{\widehat{F(T_i)}} \leq \alpha \right\}, \text{ then reject all } \mathcal{H}_{(i)}, 1, \ldots, k, \tag{5}$$

with $i/m = \widehat{F(T_i)}$ the empirical estimate of the cumulative density function (cdf) for the test statistic associated with $\mathcal{H}_{(i)}$ and x_is the estimate of $F_0(T_i)$ (p-values are the cdf of the test

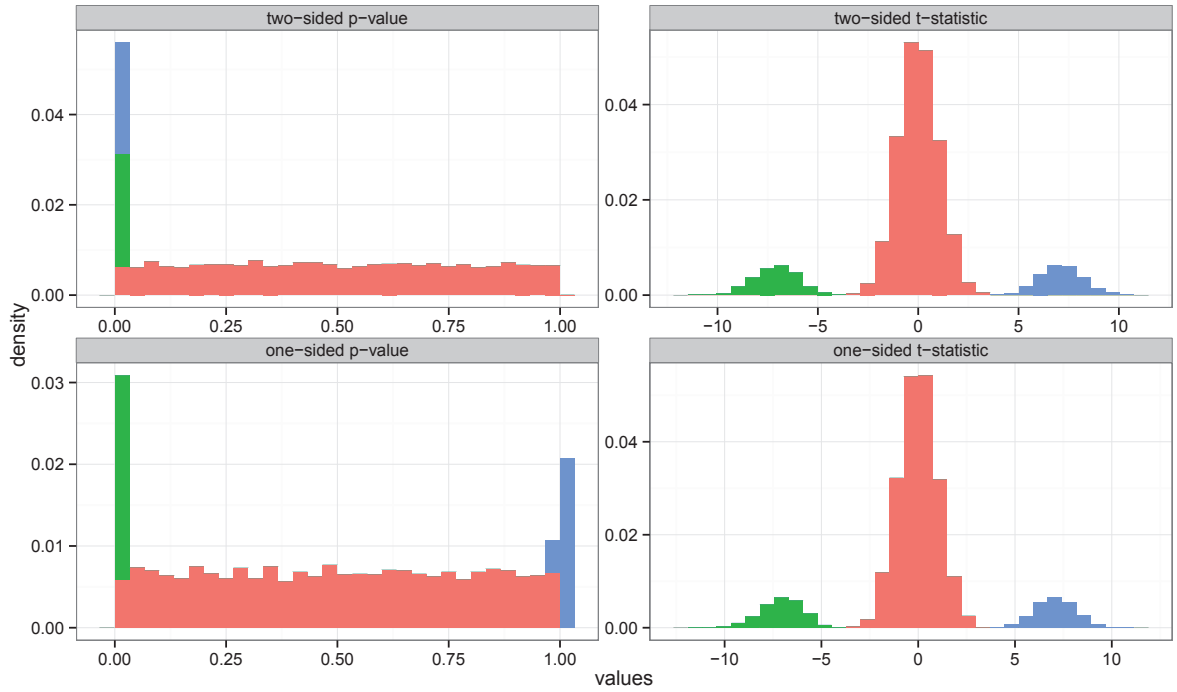


Figure 1: Comparison of bilateral versus unilateral multiple hypotheses testing for 10,000 student tests. For the bilateral tests, the t-statistics and p-values associated with the alternative hypothesis $\mathcal{H}_1$ are in green and blue depending on the side of the departure from the null hypothesis, while the values associated with the null hypothesis $\mathcal{H}_0$ appear in red. For the unilateral tests, the t-statistics and p-values associated with $\mathcal{H}_1$ appear in green, while the values associated with $\mathcal{H}_0$ appear in red or in blue if they correspond to the non-tested hypothesis.

statistic under the null distribution). In the case of unilateral multiple-testing, this estimation of F_0 does not take into account the p-values associated with the non-tested hypotheses and underestimate the real density of F_0. Moreover, every procedure that aims at increasing the power of the FDR control by using a plug-in estimator of π_0 that do not take into account this excess of p-values under the null hypothesis will over estimate the density of the uniform distribution, thus increasing the number of FP.

We propose in this paper a new procedure to handle unilateral multiple-testing using the framework developed by Sun and Tony Cai (2009) to efficiently control the FDR. The rest of the paper is structured as follow. We introduce the compound decision framework developed by Sun and Tony Cai (2009) to control the FDR as a nominal level α in section 2. We present a new mixture distribution to model unilateral p-values distribution under conditional independence for this framework in section 3. Finally in section 4 we present simulation results of our procedure and the performance of its implementation in an R package (R Core Team, 2014).

2 Optimal procedure for False Discovery Rate control

When dealing with a large number of simultaneous hypotheses testing, Sun and Tony Cai (2009) showed that the multiple-testing problem is equivalent to the weighted classification problem under mild conditions. Their work extend the framework developed by Efron et al. (2001) to different form of dependency between the tests and prove its optimality.

Let $\mathbf{x} = (x_1, \ldots, x_m)$ be a vector of observed p-values associated with a vector of unknown

states $\mathbf{H} = (H_1, \ldots, H_m) \in \{0,1\}^m$ such that:

$$X_i | H_i \sim (1 - H_i) g_0 + H_i g_1, \tag{6}$$

with g_0 the conditional probability density function (cpdf) of the p-values corresponding to null-hypotheses ($\mathcal{H}_0$) and g_1 the cpdf of the p-values corresponding to the non-null hypotheses ($\mathcal{H}_1$), denote by $\pi_0 = \Pr(H_i = 0)$ the proportion of null-hypotheses. The goal of the multiple-testing or the weighted classification problem is to choose between $\{H_i = 0\}$ and $\{H_i = 1\}$.

They define the local index of significance (LIS) defined as the following statistic:

$$LIS_i(\mathbf{x}) = \Pr(H_i = 0 | \mathbf{x}), \tag{7}$$

With the LIS and a class of tests statistic that satisfy the monotone ratio condition (MRC), Sun and Tony Cai (2009) define the following optimal multiple-testing procedure for a FDR level α given in their Theorem 4:

With $LIS_{(1)}, \ldots, LIS_{(m)}$ the ordered LIS statistics and $\mathcal{H}_{(1)}, \ldots, \mathcal{H}_{(m)}$ the corresponding hypotheses, the following testing control procedure (the LIS procedure) is valid at an FDR level α:

$$\text{let } k = \max \left\{ i : \frac{1}{i} \sum_{j=1}^{i} LIS_{(j)}(\mathbf{x}) \leq \alpha \right\}, \text{ then reject all } \mathcal{H}_{(i)}, i = 1, \ldots, k. \tag{8}$$

Theorem 5 and 6 of Sun and Tony Cai (2009) show that a local index of significance testing procedure where LIS_i in (8) is replaced by its plug-in statistic $\widehat{LIS}_i$ is optimal and asymptotically valid for a level α.

In the independent case where $H_i \sim \mathcal{B}(\pi_0)$, the LIS reduces to the local false discovery rate (ℓFDR) (Efron et al., 2001):

$$\ell FDR(x_i) = \frac{\Pr(x_i | H_i = 0)}{\Pr(x_i | H_i = 0) + \Pr(x_i | H_i = 1)}. \tag{9}$$

Moreover, Sun and Cai (2007) also showed that the $lFDR(x_i)$ is optimal and asymptotically valid compared to other FDR procedures for a level α in the independent case.

More generally, multiple-testing procedures consist in first ranking the m tests by importance and then thresholding accordingly to a metric that we want to control for. When multiple-testing procedures are based on p-values the general decision rule is simple and we can solve the m component problems separately, when, where based on LIS_i or ℓFDR_i the general decision rule is compound (the classification of H_i depends on other $H_j, j \neq i$) which allows for a better ranking of the hypotheses (Sun and Cai, 2007).

3 p-values distribution for unilateral multiple-testing

The general framework developed in the previous section is valid for unilateral and bilateral multiple-testing. However, for unilateral tests, we need to account for the differences of the shape of the p-value distributions compared to bilateral multiple-testing. Here, we present a new mixture model for one-sided p-values distribution that take into account the p-values associated with the non-tested hypothesis. This new model brings the power of the LIS multiple-testing procedure (8) to unilateral multiple-hypothesis testing.

For bilateral multiple-hypothesis testing, most estimators rely on the following assumption on the shape of the conditional p-values distribution (Celisse and Robin, 2008):

$$\exists \lambda \in]0,1] / \forall i \in \{1, \ldots, m\}, X_i \in [\lambda, 1] \implies X_i | H_i \sim \mathcal{U}([\lambda, 1]). \tag{B}$$

The assumption can be rationalized by the fact that for multiple hypotheses testing, we generally are in the case where the number of null hypothesis is larger than the number of alternative hypothesis (*i.e.* $\pi_0 \gg (1-\pi_0)$), thus with the MRC (Sun and Tony Cai, 2009) there is a point λ above which there is almost no contribution of g_1 to the mixture model such that:

$$\forall i \in \{1,\dots,m\}, x_i \in [\lambda,1] \implies \mathbb{E}[H_i = 0|\mathbf{x}] \gg \mathbb{E}[H_i = 1|\mathbf{x}]. \tag{10}$$

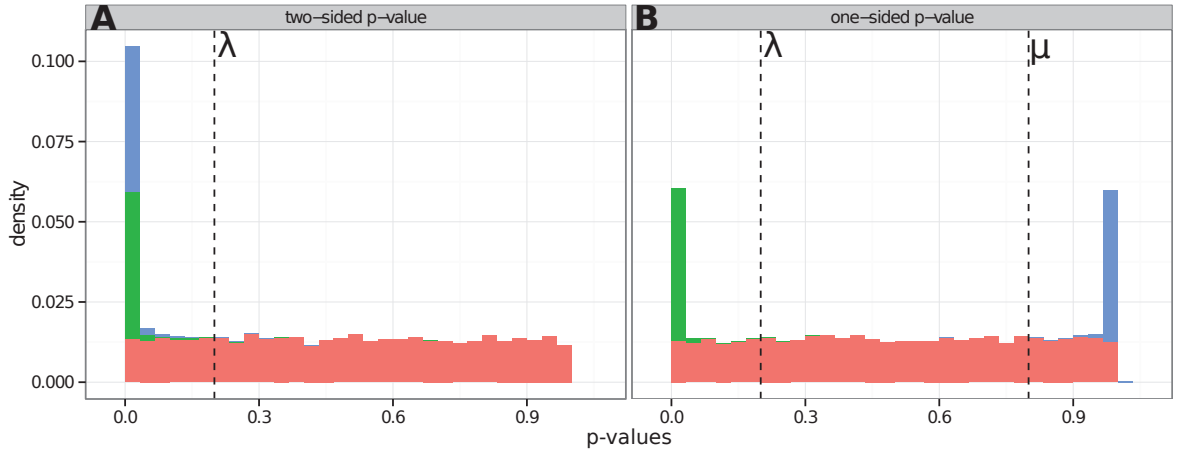


Figure 2: Comparison of bilateral versus unilateral multiple hypotheses testing p-values distributions. For the bilateral tests, the p-values associated with the alternative hypothesis $\mathcal{H}_1$ are in green and blue depending on the side of the departure from the null hypothesis, while the values associated with the null hypothesis $\mathcal{H}_0$ appear in red. For the unilateral tests, the p-values associated with $\mathcal{H}_1$ appear in green, while the values associated with $\mathcal{H}_0$ appear in red or in blue if they correspond to the non-tested hypothesis. The dashed line correspond to the position of λ and μ.

However, for some cases assumption (B) do not hold (Pounds and Cheng, 2006). Unilateral hypothesis testing is one of these cases, where the p-values under the null hypothesis can arise from a uniform distribution, or can be stochastically higher than uniform, if their associated hypothesis is the non-tested hypothesis. Our approach extends the mixture model presented in equation (3) to account for the characteristics of a distribution of unilateral p-values (see Figure 2).

Like for two-sided multiple-hypothesis testing, our first assumption is that a p-values distribution is a mixture of two populations in proportion π_0 and $(1-\pi_0)$ with π_0 the proportion of p-values associated with the null hypothesis. However, by opposition to two-sided multiple hypothesis testing, we don't have only one, but two sources of p-values associated with the null hypothesis: we have p-values computed from statistics that follow the model defined by the test and p-values computed from statistics that do not follow the model defined by the test and whose departure from this model is not tested. To describe the p-values distribution of multiple unilateral tests, our second assumption is that the population of p-values associated with the null hypothesis is subdivided in two sub-populations in proportion π_1 and $(1-\pi_1)$ with π_1 the proportion of p-values corresponding with the null hypothesis because their associated hypothesis is the untested hypothesis.

Like two-sided multiple hypothesis testing, for one-sided multiple hypothesis testing, the p-values can be seen as a random vector X of m random variables. We propose to model the probability density distribution of X, $g(x_i), i = \{1,\dots,m\}$, as the following three-component mixture:

$$g(x) = (1-\pi_0)\, g_1(x) + \pi_0 (\pi_1 g_2(x) + (1-\pi_1)\, g_0(x)) \tag{11}$$

with $g_0 = \mathcal{U}(]0,1])$ and g_1 and g_2 two unknown probability density function. If the test statistic satisfies the MRC (Sun and Tony Cai, 2009), $g_1(x)$ is decreasing with high density near 0 and $g_2(x)$ is increasing with high density near 1 (Han et al., 2011).

To control the FDR at a given level, we need to discriminate between p-values associated with the null hypothesis and p-values associated with the alternative hypothesis. For this we need to introduce a new random variable $H \in \{0,1\}^m$. H is hidden and H_i is equal to 1 when the hypothesis associated with X_i is the alternative hypothesis or equal to 0 otherwise. We could introduce another hidden random variable $U \in \{0,1\}^m$ such that when $U = 1$, X_i follow the uniform distribution and g_2 otherwise. However, a major challenge with this framework is to discriminate between the three compartments of the p-values distribution, two of which are unknown (Han et al., 2011). To simplify this problem and control the FDR for unilateral multiple-testing where assumption (B) do not hold, Celisse and Robin (2008) proposed the following milder assumption (see Figure 2 B):

$$\exists \Lambda = [\lambda, \mu] \subset (0,1] / \forall i \in \{1, \dots, m\}, X_i \in \Lambda \implies X_i | H_i \sim \mathcal{U}(\Lambda), \tag{C}$$

In this setting, the distribution $g(x)$ is only supposed uniform on the interval $[\lambda, \mu]$. With the MRC (Sun and Tony Cai, 2009) this assumption can be rationalized such that there is a point μ under which there is almost no contribution of g_2 to the mixture model, since we expect to be in the case where we have more hypotheses under the true null-hypothesis than the non-tested hypotheses. Otherwise we may want to perform another test that better defines the null-hypothesis.

We emphasize that for the FDR control, we are only interested in discriminating between the two sources of hypothesis: correctly assigning the labels H. Thus from Assumption (C), with an estimate of μ, we propose to simplify this problem by working with a vector of p-values right-censored in μ.

3.1 Right censoring

Here we present our censored model of unilateral p-values distribution. We propose to work with the right-censored random variable Y instead of X such that:

$$Y_i = X_i \mathbb{1}_{[0,\mu[}(X_i) + \mu \mathbb{1}_{[\mu,1]}(X_i). \tag{12}$$

By opposition to truncated data, where some observations are missing, with a censored variable we have access to all the observations. With Y being right censored in μ, all Y observations corresponding with X observations greater than to μ are set to μ. Therefore, the probability density function of Y, h is similar to the pdf of X with all the mass of $g(x)$ on the interval $[\mu, 1]$ shifted to μ. We can write the pdf of Y is as the following mixture:

$$h(y) = (1-\pi_0)\tilde{g}_1(y) + \pi_0[(1-\pi_1)\tilde{g}_0(y) + \pi_1 \tilde{g}_2(y)]$$

with :

$$\forall k \in \{0,1,2\}, \tilde{g}_k(y) = g_k(y)\mathbb{1}_{(y<\mu)} + \delta_\mu(y)[1 - G_k(\mu)]$$

were $\delta_\mu(y)$ is the Dirac density on μ.

Under the Assumption (C) we have $[1 - G_1(\mu)] = 0$ and $G_2(\mu) = 0$. Moreover, with $\int_0^\lambda g_1(x)\,dx = 1$ we can drop $\mathbb{1}_{(y<\mu)}$. Thus we have:

$$\tilde{g}_1(y) = g_1(y) \text{ and } \tilde{g}_2(y) = \delta_\mu(y)$$

To simplify our model further, we consider the density point in μ as one unique component of the mixture. Therefore the pdf of Y is:

$$h(y) = (1-\pi_0)g_1(y) + \pi_0(1-\pi_1)g_0(y)\mathbb{1}_{(y<\mu)} + [\pi_0\pi_1 + \pi_0(1-\pi_1)(1-G_0(\mu))]\delta_\mu(y)$$

Let κ be the weight of the Diract distribution in the mixture, we can rewrite $h(y)$ without π_1 the proportion of untested alternative hypotheses, such that:

$$h(y) = (1-\pi_0)\, g_1(y) + \pi_0 \left[\left(1 - \frac{\kappa}{\pi_0}\right) g_0(y)\, \mathbb{1}_{(y<\mu)} + \frac{\kappa}{\pi_0} \delta_\mu(y) \right] \tag{13}$$

Note that $g_0(y)\, \mathbb{1}_{(y<\mu)}$ the pdf of the uniform distribution on $]0,1]$ set to be equal to zero on $[\mu,1]$ is equal to the pdf of the uniform distribution on $]0,\mu[$. Moreover, with $\forall y \geq \lambda, g_1(y) = 0$ and $\forall y = \mu, g_0(y)\, \mathbb{1}_{(y<\mu)} = 0$, we set implicitly $H_i = 0, \forall y_i = \mu$. Thus our model corresponds to Assumption (C) by splitting the p-values into three segments:

- $]0,\lambda[$, with a mixture of p-values associated with the alternative hypothesis and p-values associated with the null hypothesis that follow an uniform distribution:

$$\forall y_i \in]0,\lambda[\,, h(y) = (1-\pi_0)\, g_1(y) + \pi_0 \left(1 - \frac{\kappa}{\pi_0}\right) g_0(y)\, \mathbb{1}_{(y<\mu)},$$

- $[\lambda,\mu[$, with only p-values associated with the null hypothesis that follow an uniform distribution:

$$\forall y_i \in [\lambda,\mu[\,, h(y) = \pi_0 \left(1 - \frac{\kappa}{\pi_0}\right) g_0(y)\, \mathbb{1}_{(y<\mu)},$$

- and $[\mu,1[$, with only p-values associated with the null hypothesis,

$$\forall y_i \in [\mu,1[\,, h(y) = \kappa \delta_\mu(y).$$

With our model of censored p-values and an estimate of μ, we can easily estimate the other model parameters $\theta = \{\pi_0, \kappa, g_1\}$ and most importantly the probability that the hypothesis associated with the i^{th} p-values is the null hypothesis.

With the hidden variable H, the complete-data log-likelihood of our model is:

$$\begin{aligned} \log \mathcal{L}(y, \mu, \theta, \mathbf{H}) &= \sum_{i=1}^{m} \mathbb{1}(H_i = 1) \log\left((1-\pi_0)\, g_1(y)\right) \\ &+ \sum_{i=1}^{m} \mathbb{1}(H_i = 1) \log\left(\pi_0 \left[\left(1 - \frac{\kappa}{\pi_0}\right) g_0(y)\, \mathbb{1}_{(y<\mu)} + \frac{\kappa}{\pi_0} \delta_\mu(y) \right]\right) \end{aligned}$$

With the indicator function $\mathbb{1}(H_i = a)$ equal to 1 when $H_i = a$ and zero otherwise.

We highlight that our model do specify the point λ and only rely on the posterior probability $\Pr(H = 0|\,\mathbf{y})$ to discriminate between null and alternative hypothesis. Therefore, we expect less FN in our results by not making an error in the estimation of λ in our model.

To be able to use this right censoring of the p-values, we need to obtain an estimate of μ.

3.2 Estimation of the censoring point μ

Celisse and Robin (2008) proposed to obtain a histogram estimation of the p-value density distribution of multiple unilateral tests by exact leave-p-out cross validation (LPO). With this approach we can select an irregular histogram with a wide central column corresponding to the interval $[\lambda,\mu]$ that minimize the quadratic risk of the model. We propose to use this framework to retrieve an estimator of μ as a byproduct of the LPO procedure.

Let us consider $\mathcal{M}_D$ the set of all possible partitions on $[0,1]$ in D segments with equal length:

$$\mathcal{M}_D = \left\{ \mathsf{m}, \mathsf{m} = (I_d)_{d=1}^{D}, I_d = \left] \frac{d-1}{N}, \frac{d}{N} \right] \right\}.$$

Let us denote $\mathcal{S}$ the collection of estimators we consider, such that:

$$\mathcal{S}_{N,D} = \left\{\widehat{s}_{D,N}(x;\widehat{\lambda},\widehat{\mu})\right\} \text{ and } \mathcal{S} = \bigcup_{N,D} \mathcal{S}_{N,D},$$

with

$$\begin{aligned}\widehat{s}_{D,N}(x;\widehat{\lambda},\widehat{\mu}) &= \sum_{d=1}^{N\widehat{\lambda}} \left[\sum_{i=1}^{m} \mathbb{1}\{x_i \in I_d\}\right] \mathbb{1}\{x \in I_d\} + \left[\sum_{i=1}^{m} \mathbb{1}\{x_i \in]\widehat{\lambda},\widehat{\mu}]\}\right] \mathbb{1}\{x \in]\widehat{\lambda},\widehat{\mu}]\} \\ &+ \sum_{d=N\widehat{\lambda}+2}^{D} \left[\sum_{i=1}^{m} \mathbb{1}\{x_i \in I_d\}\right] \mathbb{1}\{x \in I_{N\widehat{\mu}+1+d-(N\widehat{\lambda}+2)}\}\end{aligned}$$

with the constraint that $N\widehat{\mu} + 1 + D - (N\widehat{\lambda} - 2) = N$.

Celisse and Robin (2008) derived the following closed formula to compute the quadratic risk of the model, with leave-p-out cross-validation for every $p \in \{1, \ldots, n-1\}$, for risk estimation Celisse (2014) proves the leave-one-out optimality ($p = 1$), when the family of models to explore is not too large:

$$\widehat{R}_1\left(\widehat{s}_{D,N}(x;\widehat{\lambda},\widehat{\mu})\right) = \frac{1}{(m-1)(m-1)} \sum_{d=1}^{N} \frac{1}{|I_d|} \left[(2m-1)\frac{m_d}{m} - m^2\left(\frac{m_d}{m}\right)^2\right] \tag{14}$$

where $m_i = \sum_{j=1}^{m} \mathbb{1}\{x_j \in I_i\}$

If Assumption (C) is fulfilled, by minimizing $\widehat{R}_1\left(\widehat{s}_{N,\widehat{\lambda},\widehat{\mu}}\right)$ we expect to select an histogram estimator $\widehat{s}_{D,N}(x;\widehat{\lambda},\widehat{\mu})$ with a wide central interval $\left[\widehat{\lambda},\widehat{\mu}\right]$ close to $[\lambda,\mu]$ which can used to estimate μ.

3.3 Estimation of $\Pr(H = 0|\mathbf{y})$

By correctly handling the excess of p-values near 1 that is the trademark of one-sided multiple-testing we can compute the probability of a p-values x_i to be associated with the null hypothesis.

While it is possible to work directly with the mixture density h (16), most of the density of g_1 is concentrated near 0 while the density g_0 is spread between 0 and $\widehat{\mu}$. Thus, from a numerical point of view, working with h will favor the estimation of g_0 over g_1 (Guedj et al., 2009).

For estimation purposes, Efron (2005) proposed work with $z-values$ such that $z = \Phi^{-1}(x)$ instead of p-values, with $\Phi(\bullet)$ the standard Gaussian cumulative density distribution. This transform spreads the z-values on $\mathbb{R}$ and "zooms" on the near zero p-values A refinement was proposed by Sun and Cai (2007), with a transform that maintains the notion of departure from the null hypothesis carried by the p-values. With this transform, the highest density point of the distribution of the z-values associated with the null-hypothesis, corresponds to the less significant p-values (equal to 1) and the density of this distribution decreases with the distance to zero.

We propose to adapt this second transform to unilateral multiple hypothesis testing with the following transform:

$$z = \Phi^{-1}\left(1 - \frac{y}{2\mu}\right) \in \Phi^{-1}[0.5, 1] \tag{15}$$

with y as defined in equation (12). We use y/μ to work in the common Gaussian framework on $\mathbb{R}^+$ instead of the interval $\left[\Phi^{-1}(1-\mu/2), \infty\right[$.

With this transform we obtain the following mixture density distribution for the z-values on $\mathbb{R}^+$:

$$f(y) = (1-\pi_0) f_1(y) + \pi_0 \left[\left(1 - \frac{\kappa}{\pi_0} \right) f_0(y) \mathbb{1}_{(z>0)} + \frac{\kappa}{\pi_0} \delta_0(y) \right] \tag{16}$$

with $f_0(z) = \Phi^{-1}(1 - g_0(y)/2\mu)$ the pdf of the standard folded Gaussian density, $\delta_0(z) = \Phi^{-1}(1 - \delta_\mu(y)/2\mu)$ and $f_1(z) = \Phi^{-1}(1 - g_1(y)/2\mu)$ an unknown pdf with a null density in zero.

For applications, the shape of the f_1 distribution is often not known and must be estimated. For the LIS procedure Sun and Tony Cai (2009) used a Gaussian mixture density distribution which is a simple and flexible tool for such estimation. However, for multiple hypothesis-testing if the observations under the null hypothesis often follow a unique known distribution, the observation under the alternative hypothesis can potentially follow a number of distributions equal to TP. Thus f_1 is no longer a simple parametric distribution. We can use a nonparametric estimation of f_1 via kernel density estimation as described in (Liu et al., 2014) with the constraint that $f_1(0) = 0$. This approach is similar to the one proposed by Guedj et al. (2009) under independence assumption of the hypotheses, which yield good performance for estimating f_1 (Nguyen and Matias, 2012, 2013).

With our LPO estimator of μ, the remaining of the model parameters can then be estimated using the EM-algorithm (Dempster et al., 1977).

We developed this model for three different dependency structures (See supplementary file S1). The first one reduces to the ℓFDR (Efron et al., 2001), with independence between the tests:

$$P(H_i = j|\mathbf{x}) = P(H_i = j|x_i), j \in \{0,1\} \tag{17}$$

The second model assumes homogeneous Markovian dependency between the tests (Sun and Tony Cai, 2009):

$$P(H_i = j|\mathbf{x}) = P(H_i = j|x_i, H_{i-1}), j \in \{0,1\} \tag{18}$$

The third model assumes non-homogeneous Markovian dependency between the tests (Kuan et al., 2012):

$$P(H_i = j|\mathbf{x}) = P(H_i = j|x_i, H_{i-1}, \mathbf{Z}), j \in \{0,1\} \tag{19}$$

This model assume that we have a vector $\mathbf{Z}_{1:m} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_m)$ of covariables with $\mathbf{Z}_i$ a vector of D covariables associated with x_i. Two assumptions are made:

$$\begin{aligned} P(H_i|H_{1:m}, \mathbf{z}_{1:m}, x_{1:m}) &= \begin{cases} P(H_i|H_{i-1}, \mathbf{z}_i) & i \geq 2 \\ P(H_1|\mathbf{z}_1) & i = 1 \end{cases} \\ P(x_i|H_i, \mathbf{z}_{1:m}, x_{i:m}) &= P(x_i|H_i) \end{aligned}$$

In this model the value of the emission probability and the transition matrix are function of $\mathbf{Z}$

$$\begin{aligned} \pi_k(\mathbf{z}) &= P(H_1 = k|\mathbf{Z}_1 = \mathbf{z}) \\ a_{jk}(\mathbf{z}) &= P(H_i = k|H_{i-1} = j, \mathbf{Z}_i = \mathbf{z}) \end{aligned}$$

As we work with probability (i.e. defined in $[0,1]$) Hughes et al. (1999) chose to employ multinomial logistic regression to parametrize the hidden state transition.

$$\begin{aligned} \pi_k(\mathbf{z}) &= \frac{\exp(\lambda_k + \boldsymbol{\rho}_k^i \times \mathbf{z})}{\sum_{\ell=1}^{\{0,1\}} \exp(\lambda_\ell + \boldsymbol{\rho}_\ell^i \times \mathbf{z})} \\ a_{jk}(\mathbf{z}) &= \frac{\exp(\sigma_{jk} + \boldsymbol{\rho}_k^i \times \mathbf{z})}{\sum_{\ell=1}^{\{0,1\}} \exp(\sigma_{j\ell} + \boldsymbol{\rho}_\ell^i \times \mathbf{z})} \end{aligned}$$

with λ_k, $\sigma_{jk} \in \mathbb{R}$ and $\boldsymbol{\rho}_k \in \mathbb{R}^D$. With $\boldsymbol{\omega}_k$ the set of transition parameters for the state k we have to set $\boldsymbol{\omega}_0 = 0$ to guarantee the uniqueness of the parameters. A homogeneous HMMs can be seen as a particular case of NHMM where $\boldsymbol{\rho} = 0$.

$$\boldsymbol{\rho} = 0 \Leftrightarrow P\left(H_i = k | H_{i-1} = j, \mathbf{Z}_i = \mathbf{z}\right) = P\left(H_i = k | H_{i-1} = j\right)$$

4 RESULTS

To evaluate the performances of the proposed procedure for the control of the FDR in the case of unilateral multiple-testing we performed different numerical simulations. As the whole procedure relies on the estimate of the cut-point μ, we first tested the consistency of this estimator from the LPO procedure with a large number of shapes for the p-values distribution. Then we performed simulation of our LIS, in order to study the control of the FDR. The results in this section show good estimates for μ, and a strong control of the FDR as for the benefit taking into account the dependency structure of the data by decreasing the FNR.

4.1 LPO simulations for the estimation of μ

To study the estimate of μ provided by LPO, we simulated p-values according to the following model:

$$g(x) = (1 - \pi_0)\, Beta_{[0,\lambda]}\left(1, \ell\right) + \pi_0 \pi_1 \mathcal{U}_{[0,1]}\left([0,1]\right) + \pi_0\left(1 - \pi_1\right) Beta_{[\mu,1]}\left(r, 1\right), \tag{20}$$

with $Beta_{[a,b]}\left(i, j\right)$ the Beta distribution function with a support on $[a, b]$ and parameters i, j. In this model, the two beta distributions correspond respectively to $g^1_{[0,\lambda]}$ and $g^2_{[\mu,1]}$, $g^0_{[0,1]}$ being the uniform distribution on $[0, 1]$. Therefore, there is no contribution of g_1 and g_2 on the interval $[\lambda, \mu]$ which fulfills condition (C). This framework allows us to have access to the true value of μ.

Simulations with this model were computed 30 times for a vector of p-values of size $m = 10000$ with the following parameters:

- $p = \{1, n/10, n/2\}$
- $\pi_0 = \{0.8, 0.9\}$
- $\pi_1 = \{0.2, 0.3, 0.4, 0.5, 0.6\}$
- $\lambda = 0.2$
- $\mu = \{0.5, 0.6, 0.7, 0.8, 0.9\}$
- $\ell = 5$
- $r = \{2.5, 3, 5\}$

The set of parameters $\{\lambda, \mu, \pi\}$, with $\pi = \pi_0\pi_1$, was estimated from the parameters of the irregular histogram minimizing the LPO risk estimator in the histogram collection $\mathcal{S}$ (**??**). By default, we explore regular grids of size $n = 3^i$ wih $i \in \{3, 4, 5, 6, 7\}$, on which are built irregular histogram defined by jth columns of width $1/n$ followed by a central column of width $(k-j)/n$ and kth columns of width $1/n$. For each size n, j ranges from 2 to $n - 3$ and k ranges from $j + 2$ to $n - 1$.

With this framework, more than 9% of the simulation provides wrong estimates of π (where $\pi > 1$). These 9% of the simulation seem to be linked with an estimation of λ too close to the one of μ. These errors with the estimation of both λ and μ seem to be explained by two factors. The first one is the choice of regular grids to explore: there are large jumps in the model

dimension to explore with $n = 3^i$ and $i \in \{3, 4, 5, 6, 7\}$ (Figure 3 left). Thus, if the increase in model complexity is linked to a large decrease in the risk, we can select the new model only because it is more complex than the previous one (*i.e.* over-fitting), which would not have been the case if all the sizes of regular grid were explored (but this would require huge computational resources). We note that this problem decreases with the value of p, but we need very large values of p for this problem to disappear (of the order $m/2$) with also increase the bias of the LPO estimators. The second one is the shape of the p-values distribution where we have a large number of p-values near 0 and 1. Therefore, we can select the right-most or left-most central column which capture an high number of p-values (Figure 3 right).

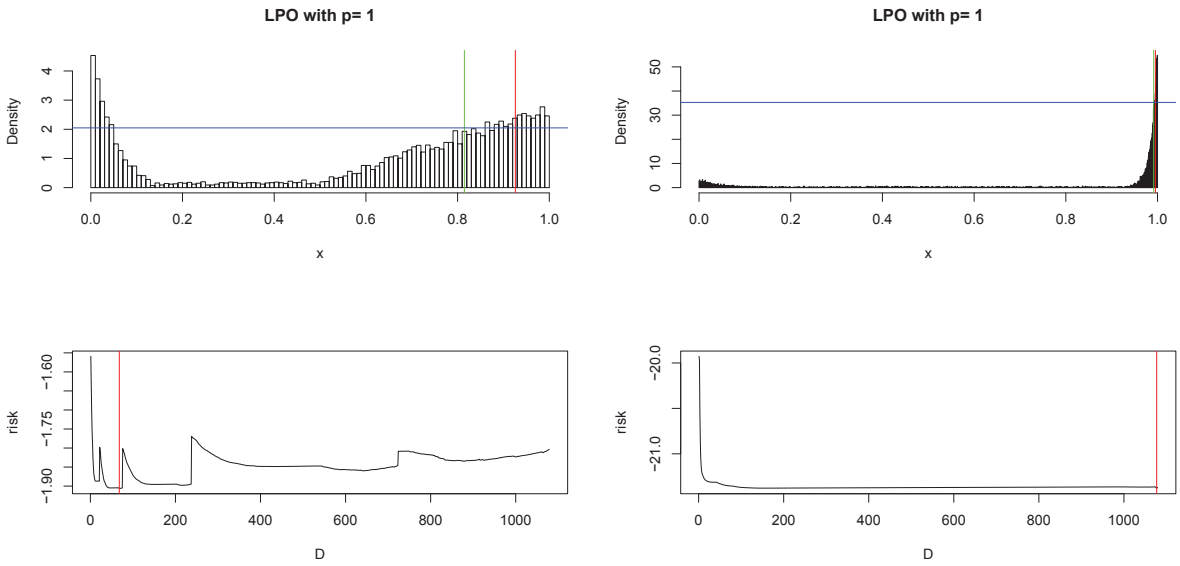


Figure 3: Three examples where LPO provides wrongs estimates (π being superior to 1). For each case the top histogram represents the p-values distribution with the green bar displaying the value of μ, the red bar the one of λ and the blue bar the one of π.

To circumvent these problems, we set the width of the central columns to be at minimum a percentage of the central column selected for the previous grid size (by default 50%). Moreover, for multiple-testing we are typically in cases where $\pi_0 \gg (1 - \pi_0)$, we fix j to be at maximum a given percentage of the grid (by default 20%). These two constraints prevent the above problems and straighten the curve of the risk (Figure 4).

Figure 5 displays the effect of these constraints on the estimation of μ for all the sets of parameters. From this figure we can see, in the cases without constraints, that there seems to be a second population of error where μ is clearly overestimated. This overestimation corresponds to the phenomenon described previously, where the model selected is the irregular histogram with the smallest central column and high value of μ and λ. However, this problem disappears with the constrains.

Globally, the values of $\widehat{\mu}$ seems to be skewed to higher value than the real μ and this tendency is magnified by the value of r (see Figure 6). The slope of the Beta density function $Beta_{[\mu,1]}(r, 1)$ near μ is governed by parameter r, with a smaller slope for higher values of r. Thus, for small slope the density of $Beta_{[\mu,1]}(r, 1)$ near μ is essentially uniform. The difficulty of estimating μ also increase when μ is small, which can also be explained by a larger support for $Beta_{[\mu,1]}(r, 1)$ and a larger interval where this Beta distribution is near uniform. At last there is also a strong correlation between the overestimation of μ and the proportion π_1. The less weight $Beta_{[\mu,1]}(r, 1)$ as in the model the harder it will be to identity its contribution to

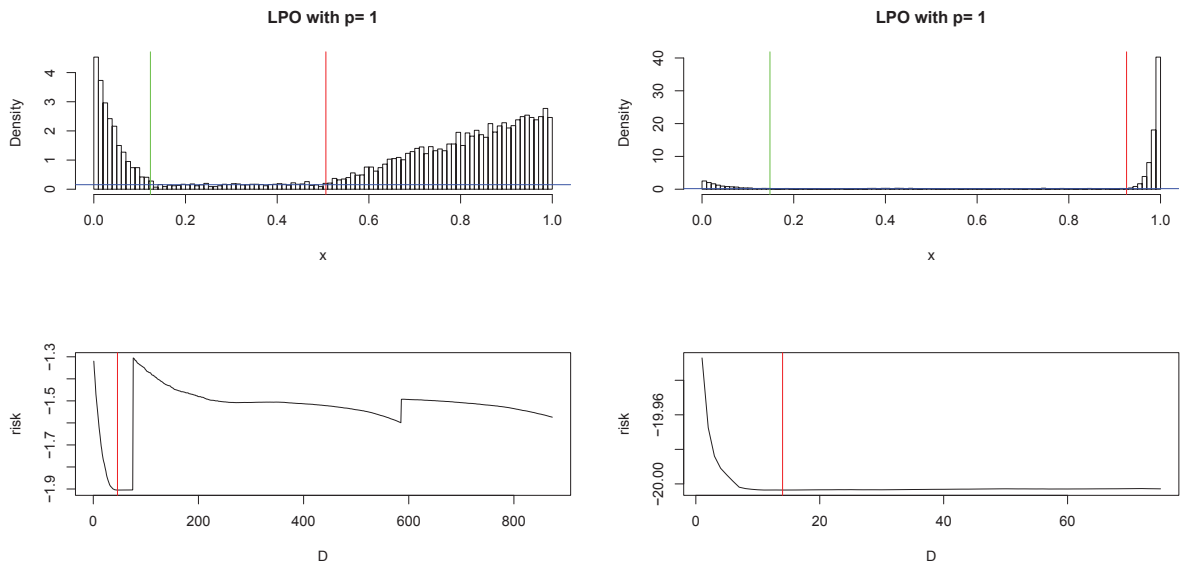


Figure 4: Results of LPO with constraint for the same three cases as the one from Figure 3 For each cases the top histogram represent the p-values distribution with the green displaying the value of μ, the red bar the one of λ and the blue bar the one of π.

the mixture $g(x)$.

To sum-up our approach seems to overestimate μ, in cases where the contribution of $Beta_{[\mu,1]}(r,1)$ near μ is almost uniform. Thus for these cases, there is almost no differences between the true model $g(x)$ and another with a wider central uniform disitribution between λ and $\widehat{\mu}$. We emphasize, that for real p-values distributions there is no true value of μ, and that our procedure seems to produce sound result for the estimation of a point $\widehat{\mu}$ above which the p-values distribution is no longer uniform.

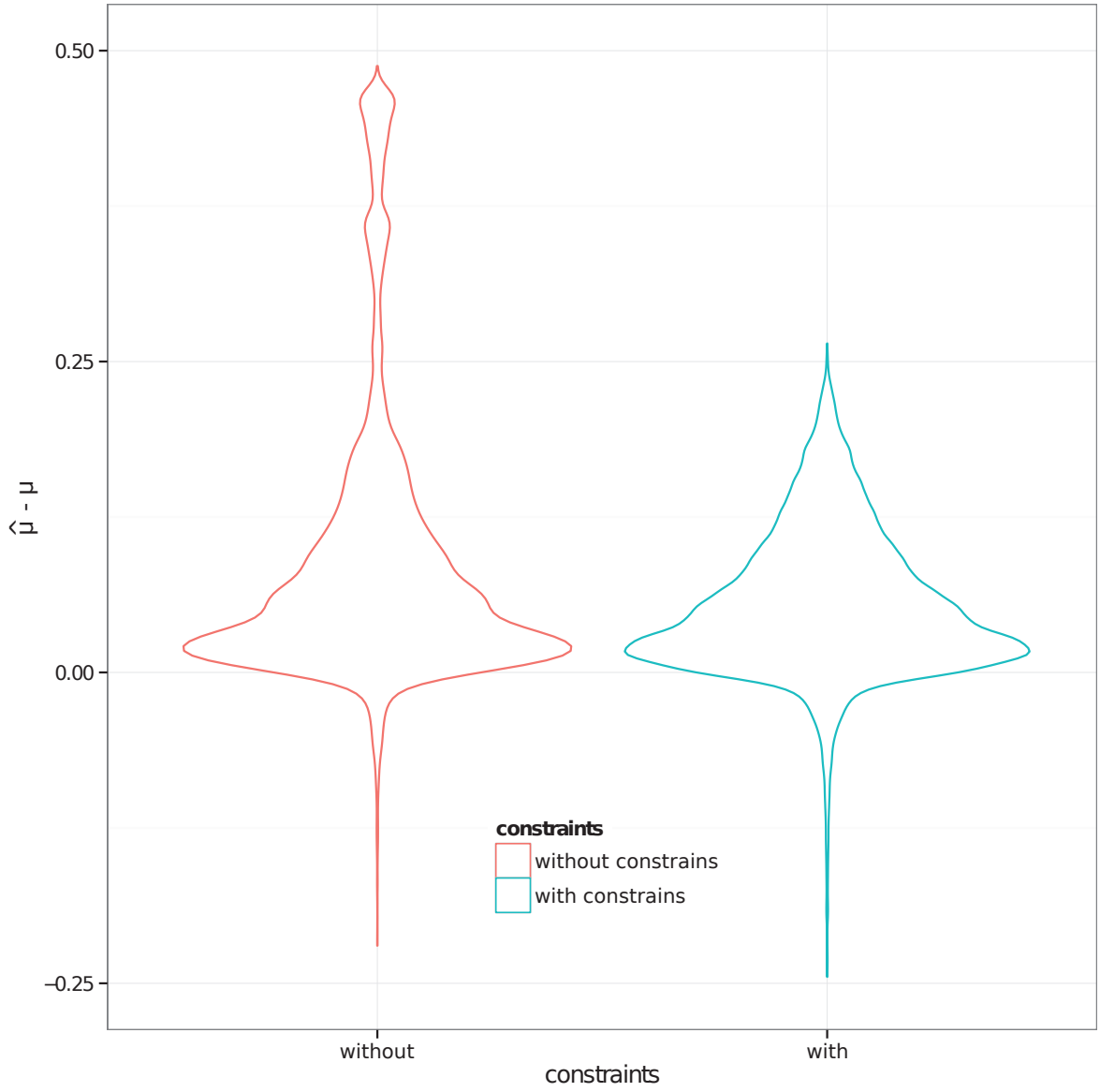


Figure 5: Violin plot of the density of the distribution of the errors made by estimating μ. The errors are simply computed by $\widehat{\mu} - \mu$ to display the skew of their distribution. The violin in red represent the distribution of errors without constrains, while the one in blue represent the distribution of errors with constrains.

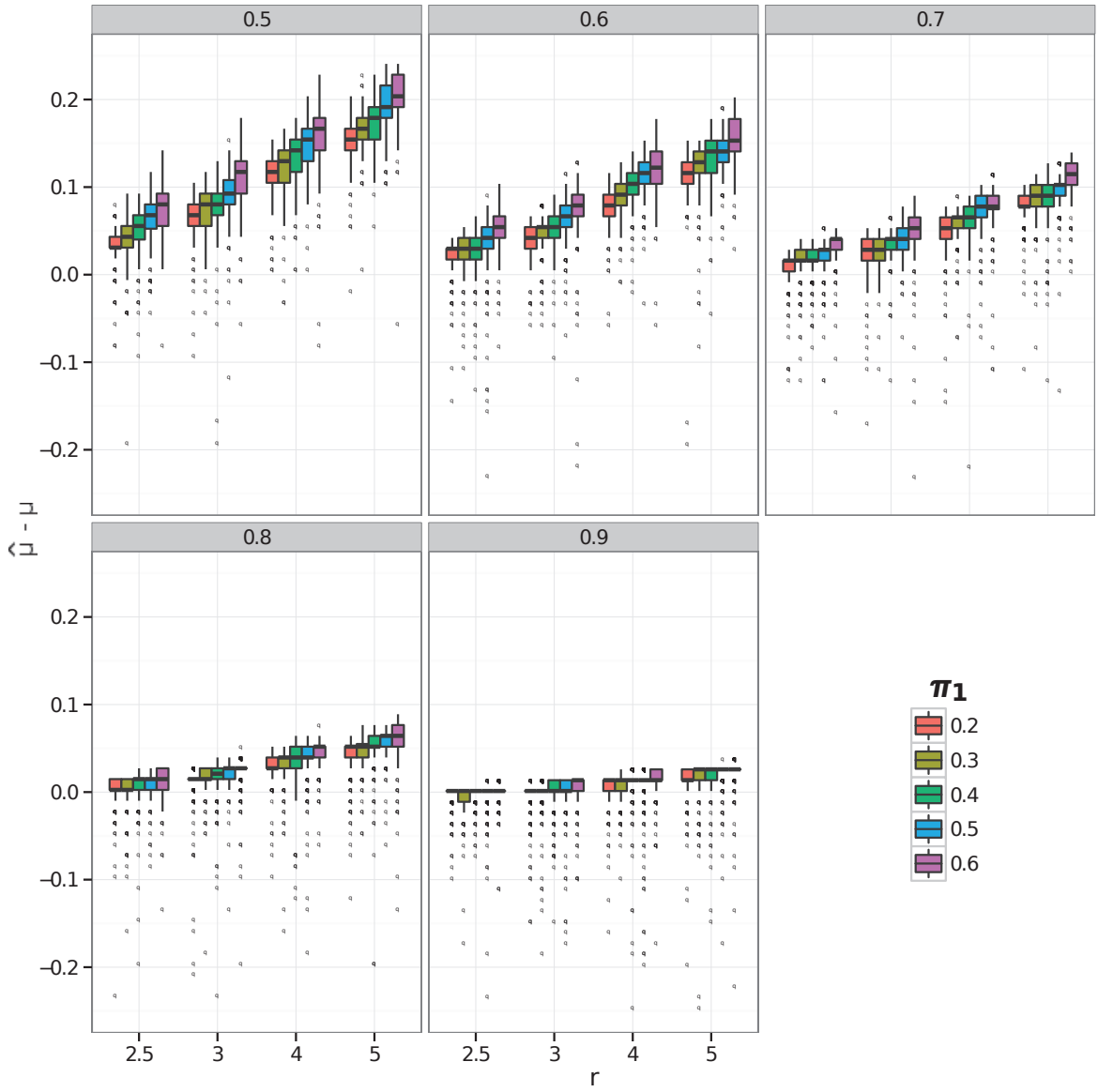


Figure 6: Boxplot of the errors made by estimating μ. The errors are simply computed by $\widehat{\mu} - \mu$ to display the skew of their distribution. Each plot correspond to a different values of μ, while the boxplot color correspond to different values of π_1 and the y-axis to different value of r

4.2 FDR simulation under independence

In this section we present a simulation study of the performance of different multiple-testing procedures for multiple-unilateral tests. We simulate unilateral z-values distribution according to the following model:

$$f(z) = (1-\pi_0)\mathcal{N}(\mu_1,1) + \pi_0\pi_1\mathcal{N}(0,1) + \pi_0(1-\pi_1)\mathcal{N}(\mu_2,1) \quad (21)$$

$$g(x) = \Phi(f(z)), \quad (22)$$

with $\mathcal{N}(\mu_1,1)$ the distribution of the test statistic under the alternative hypothesis, $\mathcal{N}(0,1)$ the distribution of the test statistic under the null hypothesis defined by the test and $\mathcal{N}(\mu_2,1)$ the distribution of the test statistic under the null hypothesis corresponding to the non-tested hypotheses. With $\Phi(x)$, the cumulative density distribution of a standard Gaussian distribution, $g(x)$ is the distribution of the p-values corresponding to the z-statistic $f(z)$ when $\mu_1 < 0$ and $\mu_2 > 0$.

Moreover, the data where generated under the NHMM model (19) to evaluate the gain in taking into account the dependency structure of the data. Simulations with this model were computed 10 times for a vector of p-values of size $m = 10000$ with the following parameters:

- $\pi_0 = \{0.8, 0.9, 0.95\}$
- $\pi_1 = \{0.8, 0.7, 0.6, 0.5, 0.4\}$
- $\mu_1 = \{-1, -1.5, -2, -2.5, -3, -3.5, -4\}$
- $\mu_2 = \{1, 2, 3, 4\}$

For the parameter π_0, we manually set the vector of parameters $\boldsymbol{\omega}$ of the NHMM to values corresponding to $\pi_0 = \{0.8, 0.9, 0.95\}$ in the data.

For each simulation, the results of four procedures where recorded for a FDR level of 0.1. The Benjamini-Hochberg procedure Benjamini and Hochberg (1995), and our procedure with the independence model (17), the homogeneous Markovian dependence model (18) and the non-homogeneous Markovian dependence model (19) between the hidden states H. Due to the unilateral nature of the p-values simulated, we could not use other procedures based on the ℓFDR, as these procedure do not take into account the excess of p-values near 1. This leads to bad estimates for the models parameters, that are of no interest of a comparative study. Step-up procedures like BH, avoid the problem linked to this misspecification of the model for the p-values distribution simply by stopping before it happens. In this case, the step-up procedure consists in rejecting all p-values, starting form the smallest one, until the average of the rejected p-values reach the selected FDR threshold α. In this case, except for unrealistic threshold (*i.e.* $\alpha \geq \mu$), the p-values associated with the non-tested hypothesis will never be considered by the procedure.

Figure 7 compares the average FDR for the four multiples-testing procedures with different values of μ_1 and π_1. The value of μ_1 can be seen as a measure of the difference between the p-values associated with the alternative hypothesis and the one associated with the null-hypothesis. Thus, for values of μ_1 close to zero, the task of discriminating between $H = 0$ and $H = 1$ is harder. Globally we can observe a strong control of the FDR for all procedures in all conditions. However, there seems to be an increasing difference between the BH procedure and the ones taking into account p-values corresponding to the untested hypothesis when their proportion increase (*i.e.* the value of π_1 decreases). While our procedure seems stable according to the different values of π_1, the FDR of the BH procedure increase with π_1. In the BH procedure, the estimate of g by i/m (4), leads to an overestimation of the density of g_0 as the contribution of g_1 to m is not taken into account. Thus, the procedure will produce less rejections for low value of π_1. Because, only extremely low p-values are rejected in this case the probability of false positive decreases.

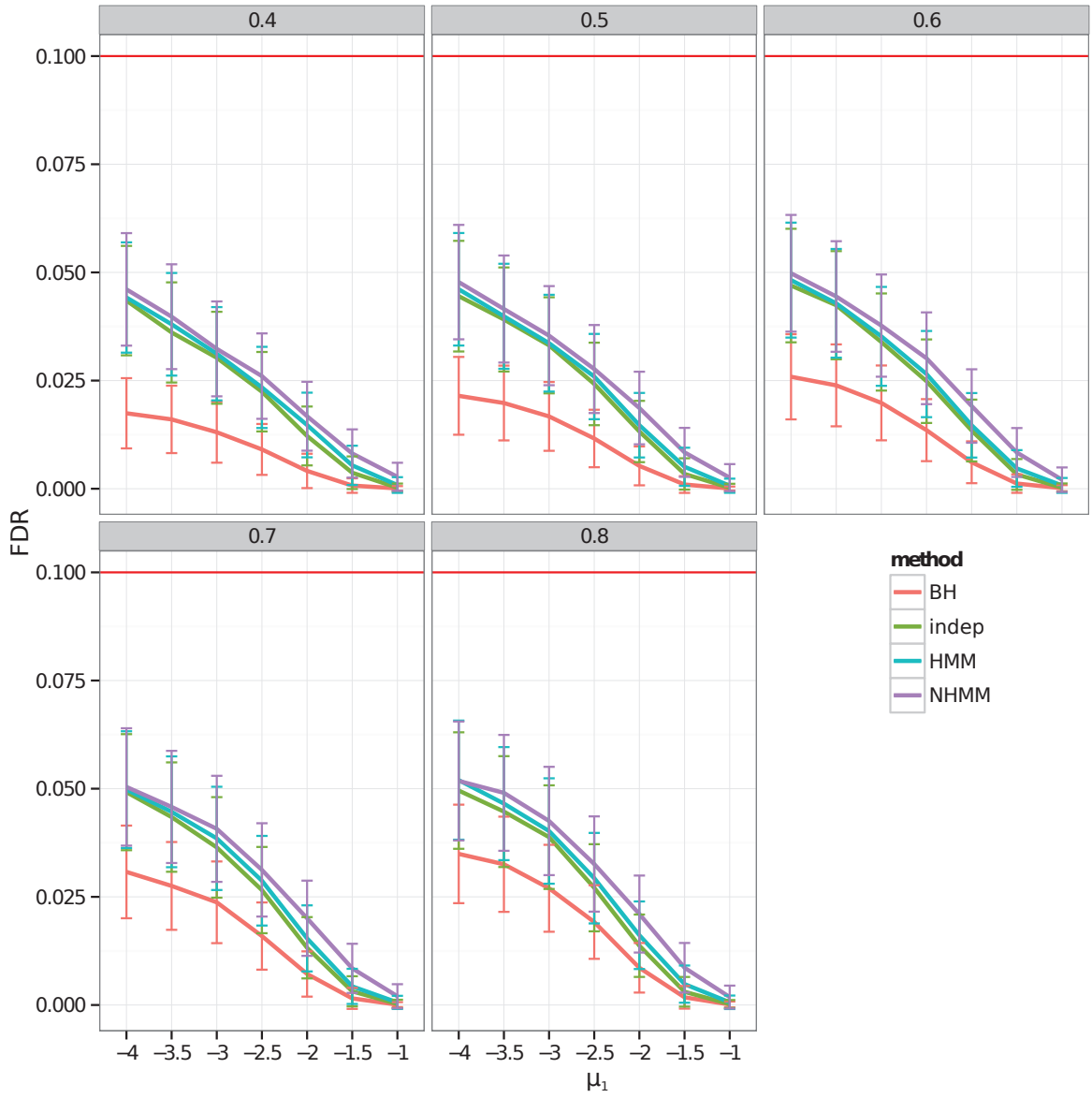


Figure 7: Curve of the FDR mean for different choices of μ_1. Each plot corresponds to different values of π_1 and each line correspond to one of the four following procedures: Benjamini-Hochberg (BH), and our procedure with the independence model (indep), the homogeneous Markovian dependence model (HMM) and the non-homogeneous Markovian dependence model (NHMM). The red horizontal line correspond to the FDR level we control for (0.1)

Figure 8 compares the average FNR for the four multiples-testing procedures with different values of μ_1 and π_1. While all the procedure are valid in term of FDR control at a level 0.1, methods accounting for the dependency structure of the data have lower FNR. The FNR decreases with the complexity of the model, with the NHMM model corresponding to the model under which the data where generated. Like for the FDR, the differences between the BH procedure and our procedures increases with the proportion of untested hypothesis. This, difference in FNR, can be crucial for real applications, as it could correspond to the differences between the detection of crucial differences and overlooking them.

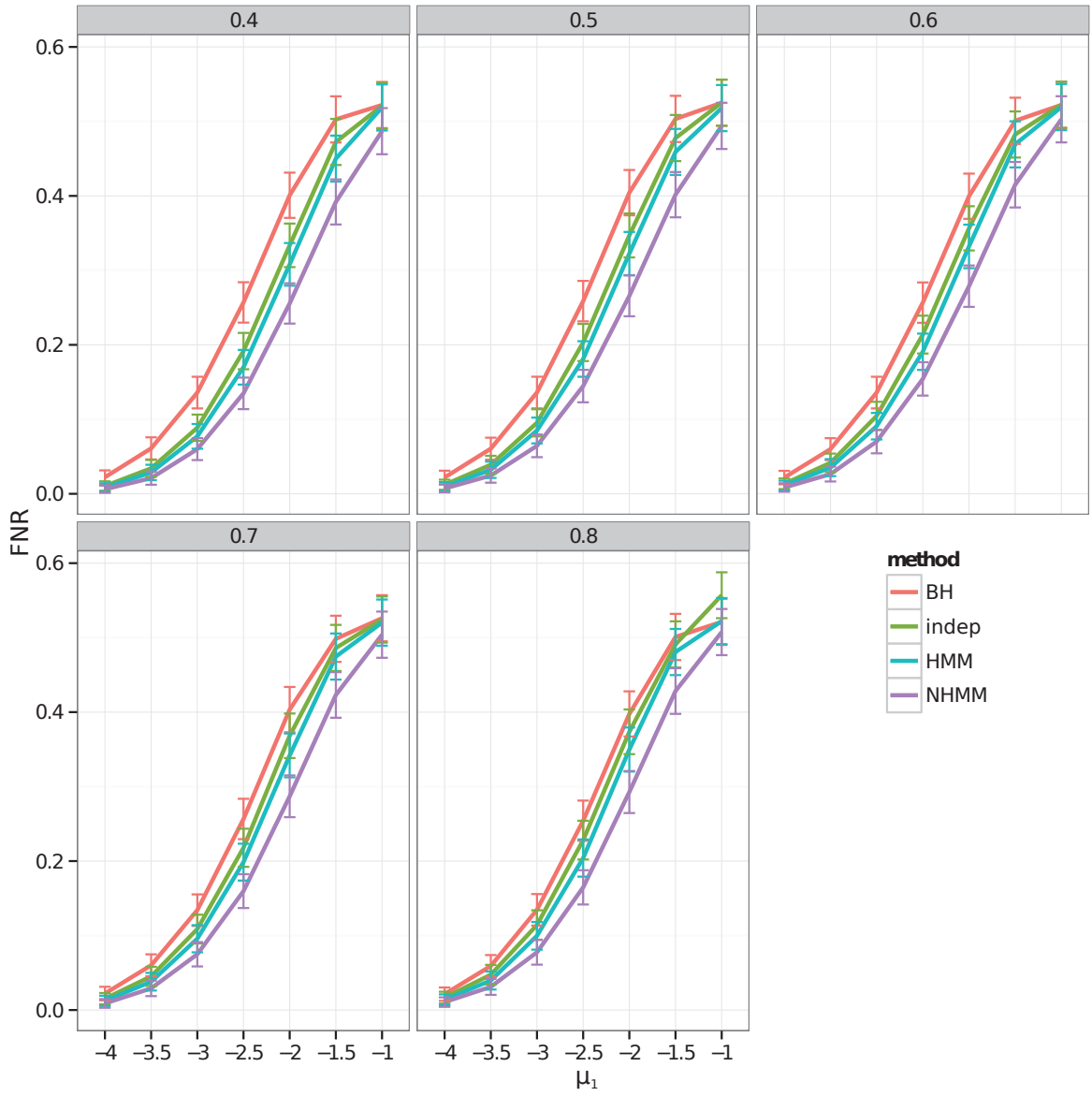


Figure 8: Curve of the FNR mean for different choices of μ_1. Each plot corresponds to different values of π_1 and each line correspond to one of the four following procedures: Benjamini-Hochberg (BH), and our procedure with the independence model (indep), the homogeneous Markovian dependence model (HMM) and the non-homogeneous Markovian dependence model (NHMM).

5 DISCUSSION

This article presents a new flexible framework for unilateral multiple-testing application. This new framework handles the excess of p-values close to 1, seen in unilateral multiple-testing applications, and our transform provide a simple probabilistic mixture model that can be easily be implemented in other procedure. Here, extend the LIS framework developed by Sun and Tony Cai (2009), given access to the better ranking of hypotheses provided by the LIS statistics for unilateral hypothesis testing. We also adapted the model proposed by Kuan et al. (2012) to integrate prior knowledge about the dependency structure of the data. Taking, into account the dependency between the different hypotheses tested can dramatically decrease the number

of false negative and represent the difference between non-significant and significant results.

This method and its implementation are freely available as an R package, uniFDR, with a fast C++ implementation of the most computationally demanding functions. Currently, the model for independent hypothesis testing, similar to the ℓFDR approaches, the model with homogeneous Markovian dependency and non-homogeneous Markovian dependency are available. However, the functions for the p-values transform are easily accessible and can be used for other procedure with few modifications to implement the mixture model (**??**) with the constraint that $\Pr(H = 1|z = 0) = 0$. For example, this could easily be implemented in the covariate-modulated ℓFDR procedure developed by Zablocki et al. (2014), that can incorporate the prior information about hypothesis provided by covariates.

Overall, with our approach, we could revisit many results of unilateral multiple-testing applications and find new positives with the power of ℓFDR and LIS based procedures.

References

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome biol.* 11.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*. 57:289–300.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*. 10:467–498.

Casella G, Berger RL. 1990. Statistical inference, the wadsworth & brooks/cole statistics/probability series, wadsworth & brooks.

Celisse A. 2014. Optimal cross-validation in density estimation with the L2-loss. *Submitted to the Annals of Statistics*. .

Celisse A, Robin S. 2008. A leave-p-out based estimation of the proportion of null hypotheses. *arXiv preprint arXiv:0804.1189*. .

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*. 39:1–38.

Dudoit S, van der Laan M. 2008. Multiple Testing Procedures with Applications to Genomics by Sandrine Dudoit, Mark J. van der Laan. *International Statistical Review*. 76:309–310.

Efron B. 2005. Local false discovery rates. Division of Biostatistics, Stanford University.

Efron B, Tibshirani R, Storey JD, Tusher V. 2001. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*. 96:1151–1160.

Genovese C, Wasserman L. 2002. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 64:499–517.

Guedj M, Robin S, Celisse A, Nuel G. 2009. Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation. *BMC bioinformatics*. 10:84.

Han B, Dalal SR, McCaffrey DF. 2011. Simultaneous One-Sided Tests With Application to Education Evaluation Systems. *Journal of Educational and Behavioral Statistics*. 37:114–136.

Hochberg Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 75:800.

Hughes JP, Guttorp P, Charles SP. 1999. A non-homogeneous hidden Markov model for precipitation occurrence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 48:15–30.

Kuan PF, Carolina N, Hill C, Chiang DY. 2012. Integrating Prior Knowledge in Multiple Testing under Dependence with Applications to Detecting Differential DNA Methylation. *Biometrics*. pp. 1–9.

Liu J, Zhang C, Burnside E, Page D. 2014. Multiple Testing under Dependence via Semiparametric Graphical Models. In: Proceedings of The 31st International Conference on Machine Learning (ICML-14). pp. 955–963.

Modolo L, Picard F, Lerat E. 2014. A new genome-wide method to track horizontally transferred sequences: application to Drosophila. *Genome biology and evolution*. 6:416–32.

Nguyen VH, Matias C. 2012. On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. *arXiv preprint arXiv:1205.4097*. .

Nguyen VH, Matias C. 2013. Nonparametric estimation of the density of the alternative hypothesis in a multiple testing setup. Application to local false discovery rate estimation. *ESAIM: Probability and Statistics*. .

Pounds S, Cheng C. 2006. Robust estimation of the false discovery rate. *Bioinformatics*. 22:1979–87.

R Core Team. 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*. 100:9440–9445.

Sun W, Cai TT. 2007. Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control. *Journal of the American Statistical Association*. 102:901–912.

Sun W, Tony Cai T. 2009. Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 71:393–424.

Zablocki RW, Schork AJ, Levine Ra, Andreassen Oa, Dale AM, Thompson WK. 2014. Covariate-Modulated Local False Discovery Rate for Genome-wide Association Studies. *Bioinformatics*. pp. 1–8.

4

Régulation post-transcriptionnelle des éléments transposables

Depuis les dix dernières années, des progrès importants ont été faits sur la compréhension des mécanismes de régulation de la transposition des éléments transposables (ET). Comme nous l'avons vu dans l'introduction de cette thèse, les effecteurs de ces mécanismes agissent au niveau épigénétique pour réguler l'activité de transposition des ET dans les génomes. Cette régulation repose entre autre sur des mécanismes d'interférence par ARN comme ceux de la voie de la régulation par les "piwi-associated RNA" (piRNA) [220, 125]. Chez la drosophile, les piRNA sont des composants majeurs de la régulation de la transposition des ET [115, 116] et la perturbation de cette voie est associée à une mobilisation des ET [122].

La régulation de la transposition par la voie des piRNA chez la drosophile est complexe et semble étroitement liée à l'activité des ET. La voie des piRNA primaires (présente dans les cellules germinales et somatiques) est initiée par la transcription de clusters de piRNA unidirectionnels. Les séquences d'ET qui constituent ces clusters sont orientées préférentiellement dans le sens inverse de leur transcription. Les piRNA qu'ils produisent vont être associés à la protéine Piwi pour former des complexes Piwi-piRNA qui ont une localisation nucléaire. Ces complexes vont ensuite cibler spécifiquement, grâce à la complémentarité de séquence des piRNA, les ARN messagers des ET correspondant au cours de leur transcription.

Ces complexes Piwi-piRNA vont ensuite permettre le recrutement de protéines secondaires qui vont modifier les marques d'histones de la copie de l'ET correspondante, pour

la rendre inactive (voir Figure 8 page 43). Ces copies d'éléments désactivées pourraient être la source de la formation de clusters doubles de piRNA. Les clusters doubles sont transcrits dans les deux sens de lecture et sont la source de la synthèse de piRNA secondaires. Les piRNA secondaires orientés dans le sens inverse de la transcription de l'ET vont être associés à la protéine Aubergine (Aub) pour former des complexes Aub-piRNA. Alors que les piRNA secondaires orientés dans le sens de la transcription de la séquence de l'ET vont être associés à la protéine Argonaute 3 (Ago3) pour former des complexes Ago3-piRNA. Les complexes Aub-piRNA vont ensuite dégrader les transcrits des ET et des clusters doubles, dont ils sont complémentaires, pour former des petits ARN sens qui vont permettre la formation de nouveaux complexes Ago3-piRNA. Les complexes Ago3-piRNA vont à leur tour dégrader les transcrits anti-sens de clusters doubles pour former des petits ARN qui vont permettre la formation de complexes Aub-piRNA. Cette boucle d'amplification est appelée ping-pong [116] (voir Figure 8 page 43). La voie des piRNA secondaires semble donc étroitement liée avec l'activité des ET dont elle réprime la transposition, puisque les transcrits de ces ET peuvent contribuer à la formation de clusters doubles, et que les transcrits de ces éléments vont contribuer à augmenter le nombre de complexes Aub-piRNA et Ago3-piRNA.

Drosophila simulans, que nous avons présentée dans le chapitre 2, est une espèce de drosophile cosmopolite proche de l'espèce modèle *D. melanogaster* [204]. Contrairement à *D. melanogaster*, on peut observer des variations importantes de contenu en ET entre différentes populations de cette espèce [221, 222]. Ainsi, certains ET sont présents dans des populations de *D. simulans* mais semblent être absents dans d'autres populations de cette même espèce. Ces différences entre populations peuvent aussi correspondre à des variations importantes du nombre de copies d'un ET. Cette variabilité observée chez *D. simulans* pourrait être expliquée par une population plus structurée que chez *D. melanogaster* [223]. Avec le modèle de décomposition en dèmes des populations d'une espèce [176], que nous avons présenté dans l'introduction, cette structure populationnelle pourrait permettre à des ET de coloniser certaines populations de *D. simulans* tout en étant absents dans d'autres. Ces différences populationnelles pourraient aussi être associées à des différences dans les mécanismes de la régulation des ET. Comme nous l'avons vu, la régulation par la voie de l'ARN interférence est une réponse spécifique à la séquence du transcrit d'un ET. Il devrait donc aussi exister des différences dans la quantité de petits ARN associés à un élément donné suivant les populations. Les données de séquençage de petits ARN peuvent être une source d'information importante pour comprendre les mécanismes de la régulation des ET, puisqu'elles permettent de mesurer directement le nombre de petits ARN associé à une séquence d'ET.

L'étude de la régulation des ET par la voie des piRNA entre différentes populations de *D. simulans* a été un des objectifs de ma thèse. Afin de mieux comprendre une possible différence au niveau de la régulation des ET par la voie des piRNA entre différentes populations de *D. simulans*, j'ai effectué l'analyse bioinformatique des données de séquençage de nouvelle génération (NGS) de petits ARN provenant de différentes populations naturelles de *D. simulans*. Comme nous l'avons vu, de nombreux gènes sont associés à la régulation par les petits ARN, de plus les ET eux-mêmes sont transcrits, c'est pourquoi, pour compléter notre vision des différences de mécanismes de régulation des ET entre les populations de *D. simulans*, nous avons aussi étudié des données de séquençage d'ARN messager.

Dans ce chapitre, je présente les méthodes que j'ai utilisées pour analyser ces données et les résultats que j'ai obtenus. L'analyse approfondie de ces résultats sera terminée après la fin de ma thèse. Ce travail sera par ailleurs complété par une étude, effectuée par Marie Fablet, de l'expression des gènes codant pour les protéines impliquées dans la voie des piRNA.

1 Les données

Des données de séquençage de petits ARN ont été obtenues à partir de l'extraction d'ARN d'ovaires de cinq populations de *D. simulans* sélectionnées pour leur polymorphisme en terme de présence et nombre d'insertions d'ET. Ces populations sont : Makindu (Kenya), Mayotte (Madagascar), Chicharo (Portugal), Camberra (Autralie), et O.195 (USA). O.195 est la souche majoritaire dont le génome a été séquencé en 2007 [14]. Nous avons étudié des données de tissu ovarien car la régulation de la transposition des ET est particulièrement importante dans les cellules germinales [224, 225]. En effet, chaque nouvelle insertion dans les cellules germinales va être dupliquée au cours des divisions cellulaires et être présente dans toutes les cellules des individus de la descendance. Il est donc primordial que la transposition des ET soit régulée dans les gonades.

Les ARN issus de l'extraction des ovaires ont été filtrés par taille sur gel d'agarose pour ne conserver que des séquences de 23 à 27 nucléotides, ce qui correspond à la taille attendue des piRNA. Le produit de cette filtration a ensuite été polyadénylé à l'aide de polymérases polyA. Cette étape a permis d'obtenir des fragments d'ARN assez longs pour remplir les lectures de 160 à 180 paires de base (pb) du séquenceur Illumina HiSeq2000 utilisé.

Après séquençage, il a donc fallu enlever les queues polyA des lectures obtenues pour ces cinq populations. Cette étape n'est pas triviale car les taux d'erreurs produites par

ce type de technologie de séquençage augmentent avec la taille des séquences et avec le nombre de fois où un nucléotide est répété. Pour enlever les queues polyA de ces données, j'ai développé une nouvelle méthode implémentée dans le programme UrQt qui est présentée au chapitre 5. Ce programme permet aussi d'augmenter la qualité d'un jeu de données NGS en raccourcissant ses lectures pour enlever les nucléotides de mauvaise qualité qu'elles contiennent.

Comme chacune des cinq populations a été séquencée sans réplicat, j'ai de plus utilisé d'autres jeux de données de petits ARN disponibles pour certaines de ces populations. J'ai utilisé deux jeux de données publiées pour les populations de Makindu et Chicharo [226] ainsi qu'un autre pour la population de Mayotte (non publié). Ces petits ARN ont été isolés par une méthode différente permettant de retenir uniquement les petits ARN associés à des protéines avant de les sélectionner par leur taille [227]. Nous avons donc travaillé avec deux réplicats pour les populations de Mayotte, Makindu et Chicharo, et un seul réplicat pour les populations de Camberra et O.195.

Pour pouvoir étudier l'expression des gènes impliqués dans les voies de régulation des ET ainsi que l'expression des ET, nous avons aussi séquencé les ARN messagers de différentes populations de *D. simulans*. Ces données ont été obtenues fin Août 2014 pour les populations O.195, Chicharo, Makindu, Mayotte et une population du Zimbabwe pour les mêmes tissus que les petits ARN. Ces données ont été séquencées en paired-end et avec deux réplicats pour chacune des populations.

2 Analyse des données de petits ARN

2.1 Identification des piRNA

Les différentes données de petits ARN que j'ai utilisées pour cette étude ne sont pas restreintes à des séquences de piRNA. Ainsi ces séquences peuvent aussi correspondre à d'autres types de petits ARN comme les "small interfering RNA" ou des produits de la dégradation de transcrits par exemple. Ce n'est pas le cas pour les trois jeux de données pour lesquels des petits ARN associés à des protéines ont été extraits, nous nous attendons aussi à observer un plus fort enrichissement en petits ARN correspondant à des piRNA et siRNA, puisque ces deux types de séquences sont associés à des protéines de la famille Argonaute.

Pour l'analyse des données d'expression de piRNA dans les espèces modèles (comme *D. melanogaster*), une des premières étapes consiste à aligner ces séquences le long du génome de référence de l'espèce. Cette étape permet d'identifier les piRNA s'alignant

à des positions multiples et ceux qui s'alignent à une unique position du génome. Ceci permet de différencier les piRNA primaires, qui s'alignent sur les clusters de piRNA unidirectionnels (alignement unique), des piRNA secondaires, qui s'alignent sur les clusters doubles ainsi que sur les copies d'ET du génome (alignement multiple)[116]. Cette classification est néanmoins conservative puisque une partie des piRNA primaires s'aligne aussi sur les copies d'ET (alignement multiple).

Comme nous l'avons vu dans le chapitre 2, il existe deux versions du génome de *D. simulans*. La première version, produite en 2007, correspond à l'assemblage hybride de séquences provenant de cinq souches différentes de *D. simulans* [14] ; et la deuxième version, produite en 2012, correspond au séquençage de données NGS de la souche majoritairement utilisée en 2007 [205]. Les séquences du génome de 2007 représentent donc un patchwork de la diversité d'ET présents dans cinq souches, alors que la version de 2012, de meilleure qualité, ne contient que les ET présents dans la souche O.195. Cependant, les clusters de piRNA sont des régions extrêmement compliquées à assembler et à séquencer, de part les répétitions des séquences qui les composent et de leur hétérochromatinisation. Chez *D. melanogaster*, qui est une espèce modèle, il a fallu plusieurs années d'effort pour obtenir la séquence du cluster unidirectionnel *flamenco* [228]. On ne s'attend donc pas à pouvoir travailler avec les séquences de ces clusters, même dans la version du génome de *D. simulans* de 2012. De plus, contrairement à l'espèce modèle *D. melanogaster*, l'annotation de ces deux génomes est très limitée et nous n'avons pas accès à une liste de toutes les copies d'ET.

Pour étudier les mécanismes de régulation des ET par les voies d'ARN interférence, j'ai tout d'abord construit une liste de séquences de copies d'ET qui pourraient être présents chez *D. simulans*, pour pouvoir identifier les piRNA et siRNA parmi tous les petits ARN obtenus. J'ai commencé la construction de cette liste avec toutes les séquences d'ET identifiables dans les deux versions du génome de *D. simulans*, avec le programme RepeatMasker [184]. Les fichiers de sorties ont ensuite été traités avec le programme OneCodeToFindThemAll, pour construire la liste de toutes les séquences de ces copies [229].

Au vue de la variabilité de composition en ET observée chez *D. simulans* et des différences de contenu en ET entre les deux versions de son génome [221, 222, 230], il serait naturel de penser que ces deux génomes ne possèdent peut-être pas toutes les séquences des copies d'ET présentes chez *D. simulans*. J'ai donc ajouté à cette liste de séquences, les copies identifiées dans le génome de *D. melanogaster*. Cette liste représente 8040 copies d'ET. Mon intention initiale était d'utiliser l'ensemble des données de séquences d'ET de drosophiles présentes dans la base de données RepBase [8], afin d'obtenir le plus

de copies possible sur lesquelles puissent s'aligner nos données de séquençage. Malheureusement, beaucoup de ces séquences présentent des similarités importantes entre elles et il est difficile de déterminer quelles séquences correspondent vraiment à des séquences d'ET différents, ce qui est problématique pour assigner un petit ARN à un élément donné.

2.2 Association entre petits ARN et éléments transposables

Les piRNA et les siRNA ont des séquences très courtes de 21 à 27 pb. Cette caractéristique a l'avantage de nous permettre de les séquencer entièrement dans une unique lecture NGS, mais cela veut aussi dire que l'information disponible (nombre de bases) pour aligner ces séquences sur un génome est limitée. Par conséquent, un mauvais assortiment d'un faible nombre de bases pour un petit ARN peut être suffisant pour que son alignement ne soit plus possible. Afin d'augmenter le nombre de petits ARN alignables sur nos séquences d'ET, j'ai nettoyé l'ensemble des séquences de nos jeux de données avec le programme UrQt en fixant le seuil de qualité à un phred de 20. C'est aussi pour pouvoir travailler avec le plus grand nombre possible de lectures de ces petits ARN que nous avons travaillé avec les copies des ET, et non pas avec la séquence consensus.

Cette étape d'alignement effectuée avec Bowtie nous a permis d'aligner plus de 90 millions de petits ARN sur nos séquences d'ET [231]. Ces petits ARN représentent une proportion très variable des jeux de données, suivant les méthodes d'extraction utilisées. Ils représentent environ 60% des séquences des petits ARN extraits par protéine et 15% de ceux extraits uniquement par leur taille (voir Table 4.1).

Comme nous l'avons vu dans l'introduction de cette partie, la régulation par la voie des piRNA secondaires forme une boucle d'amplification appelée ping-pong. Les complexes Aub-piRNA et Ago3-piRNA ont une activité RNAse-H qui découpe l'ARN complémentaire à leurs séquences, après 10 pb de complémentarité (voir Figure 8 page 43). Le signal de cette complémentarité de 10 pb entre ces deux types de piRNA est appelé signal de ping-pong. Nous avons pu observer ce signal dans nos différents jeux de données avec plus de 60% des petits ARN présentant un recouvrement de 10 pb pour les jeux de données extraits par protéine et plus de 30% pour ceux extraits par taille (voir Figure 4.1). Le reste de ces petits ARN s'alignant sur les copies d'ET peuvent donc correspondre à des piRNA primaires sans partenaire de ping-pong, ou bien à des siRNA spécifiques de ces ET.

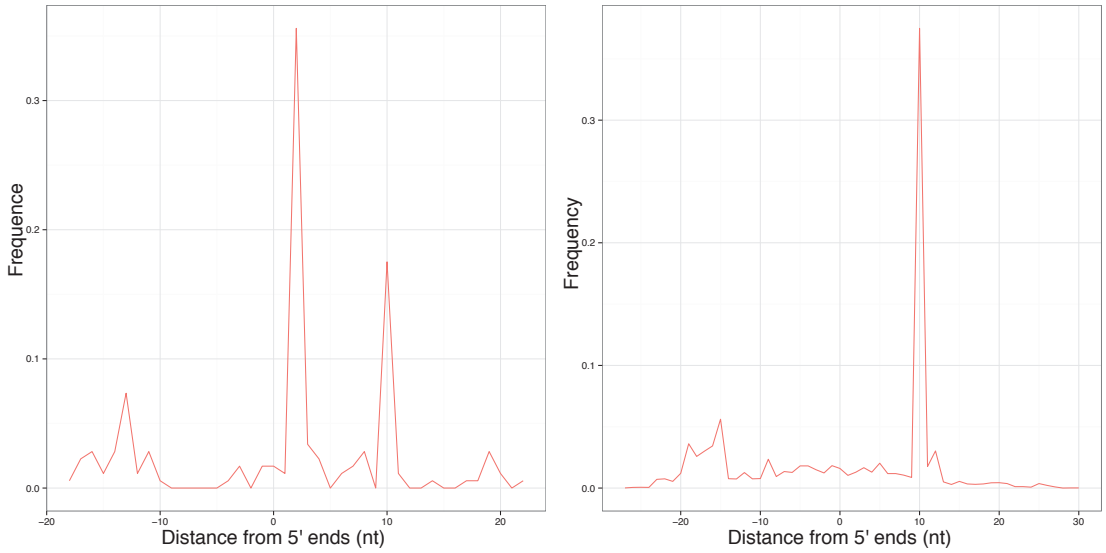


FIGURE 4.1 : Graphiques du signal de ping-pong. Le graphique de gauche correspond à la population Chicharo dont les petits ARN ont été extraits par protéine, alors que le graphique de droite correspond à la population O.195 avec les petits ARN extraits par taille. La courbe rouge correspond à la fréquence de la taille des recouvrements entre petits ARN sens et anti-sens.

Tableau 4.1 : Données de petit ARN et d'alignement sur les séquences des copies d'ET

population	protocol	nombre de lectures	lectures alignées	pourcentage aligné
Cambera	taille	21191864	3643033	17,19%
Chicharo	taille	21613466	3122463	14,45%
O.195	taille	17202387	4032617	23,44%
Makindu	taille	18769963	3119183	16,62%
Mayotte	taille	21458467	2929817	13,65%
Chicharo	protéine	45737349	28332805	61,95%
Makindu	protéine	42389730	23243252	54,83%
Mayotte	protéine	43768108	25646972	58,60%
total		232131334	94070142	-

2.3 Analyse d'expression différentielle des piRNA

Pour effectuer une analyse différentielle de la quantité de piRNA associée à chaque ET entre les différentes populations, il a fallu regrouper les comptages du nombre d'alignements par copie, par élément correspondant. Pour cette étape, j'ai développé le pipeline countTE qui permet à partir d'un fichier fastq de données NGS, d'une liste de séquences de copies d'ET et d'un fichier d'équivalence entre le nom de ces copies et de l'élément correspondant, appelé fichier rosette, de générer une liste de comptage de petits ARN pour chaque ET. Les différentes étapes de ce pipeline sont présentées dans la Figure 4.2. Le fichier rosette assigne pour chaque nom de séquences (copies d'ET), le nom de l'ET correspondant. Ce pipeline utilise le programme UrQt pour augmenter la qualité du jeu de données, le programme Bowtie ou Bowtie2 [231, 232] pour aligner les séquences des petits ARN sur celles des copies d'ET, et les outils des suites bedtools et samtools pour traiter ces données d'alignement afin d'obtenir le nombre de lecture alignées sur chaque copies d'ET [233, 234]. Ce pipeline peut aussi corriger le fichier rosette utilisé à partir des résultats d'un BLAT de toutes les copies d'ET contre elles-mêmes ou de ceux provenant du programme cd-hit-est [235, 236]. En effet, la construction d'un fichier rosette peut être le sujet d'erreurs d'annotation automatique assignant une copie à un mauvais ET ou d'erreurs humaines pendant sa création. En utilisant l'information de similarité entre copies et la règle des 80-80-80 (*i.e.* au moins 80% d'identité sur 80% de la séquence et 80

pb) [12], countTE peut donc vérifier si les copies les plus similaires à une copie donnée correspondent au même ET. Si ce n'est pas le cas, le nom de l'élément de cette copie est remplacé par le nom de l'élément le plus représenté parmi la liste des copies similaires.

Le pipeline countTE permet ensuite l'analyse de l'expression différentielle des petits ARN associés aux différents ET. Cette étape est effectuée via un script R et utilise la méthode d'analyse du package DESeq2 [237]. Il est important de souligner que la méthode de normalisation des comptages de cette méthode, ainsi que celle implémentée dans le package EdgeR [238], sont les seules méthodes de normalisation pour l'analyse d'expression différentielle à partir de données NGS actuellement valides [239]. En effet, la plupart des études sur les piRNA utilisent des méthodes de normalisation de type normalisation à 1 million de lectures ou de type RPKM (pour l'anglais "Reads Per Kilobase per Million reads") qui peuvent induire des biais importants. Bowtie semble avoir de meilleures performances que Bowtie2 pour aligner des lectures très courtes, c'est pourquoi j'ai utilisé ce programme pour traiter les données de petits ARN.

L'application de countTE pour l'analyse de nos données de petits ARN n'a conduit à la détection d'aucune différence significative dans le nombre de petits ARN entre les différentes paires de populations séquencées. Cette absence de résultat significatif semble principalement liée à l'absence de vrais réplicats techniques pour ces données de séquençage. Il est donc difficile d'isoler la part de variabilité biologique de la part de variabilité technique dans ces échantillons. Les données disponibles que j'ai utilisées pour avoir un réplicat biologique ont été extraites à partir d'un protocole différent. La différence entre l'extraction de données de petits ARN, par association avec des protéines ou par la taille uniquement, semble représenter plus de 90% de la variabilité dans ces données. La Figure 4.3 représente l'analyse en composantes principales des données, pour laquelle 90% de la variabilité est portée par l'axe 1 (en abscisse) qui semble discriminer les deux types de protocoles utilisés, alors que l'axe 2 (en ordonnée) et les axes suivants représentent respectivement 5% et moins de 1% de la variabilité.

J'ai aussi essayé d'utiliser ces données de piRNA pour la détection de clusters de piRNA le long des deux versions du génome de *D. simulans* disponibles. Malheureusement, comme nous l'avons dit précédemment, la qualité de ces génomes n'est pas suffisante pour avoir accès à ce type de régions difficiles à séquencer et à assembler. Les différentes méthodes utilisées, segmentSeq et piClust [240, 241], n'ont conduit qu'à l'identification de petits clusters doubles de piRNA d'une taille comparable à celle d'un élément transposable. Il est donc difficile de différencier des copies d'ET dont la transcription est régulée par le mécanisme du ping-pong, de clusters doubles de piRNA à partir de ces résultats, même si la découverte récente de l'activité des complexes formés

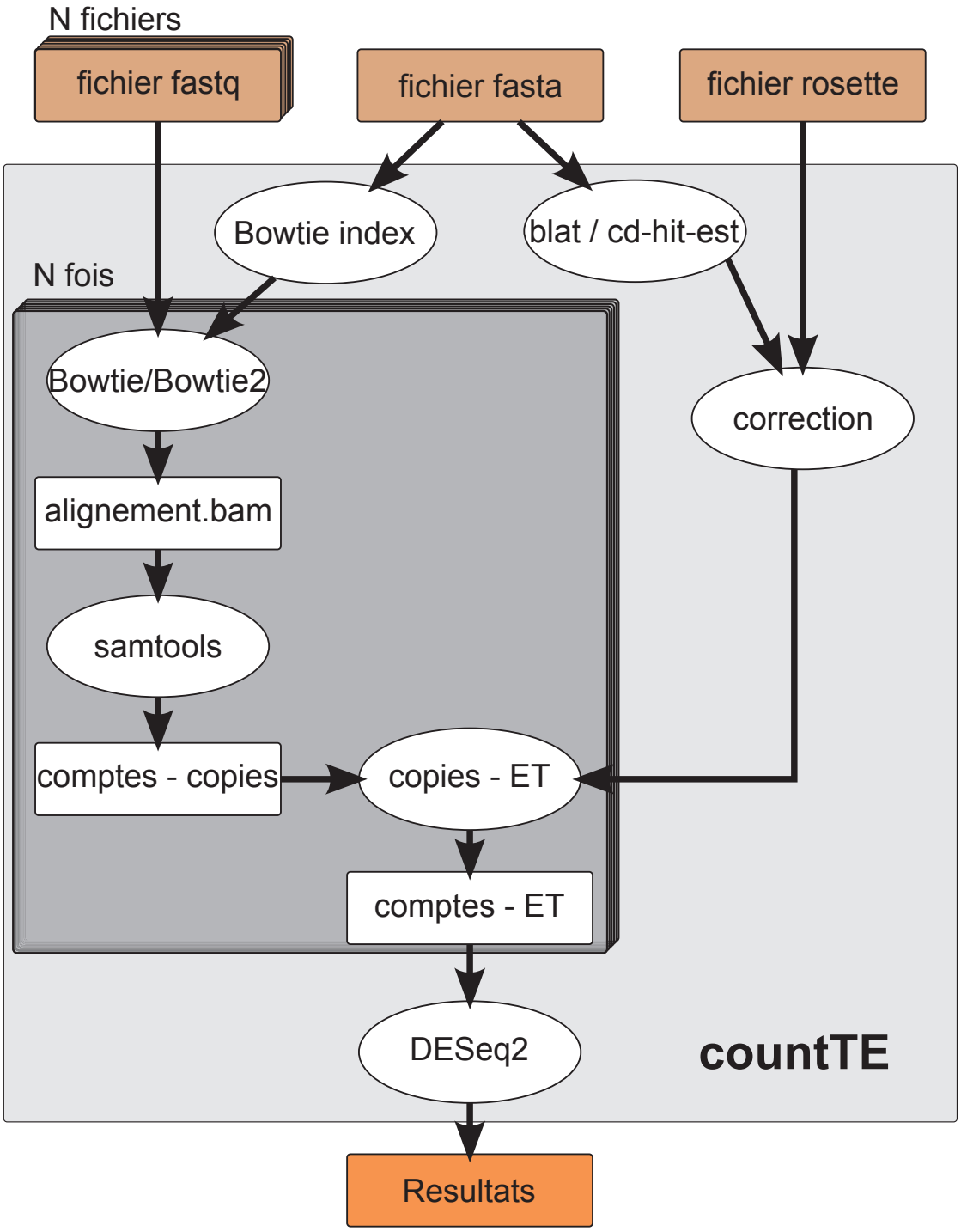


FIGURE 4.2 : Organigramme du fonctionnement de countTE. Les données sont représentées par des rectangles, alors que les étapes d'analyse sont représentées par des ovales. Les fichiers d'entrée apparaissent en marron et ceux de sortie en orange.

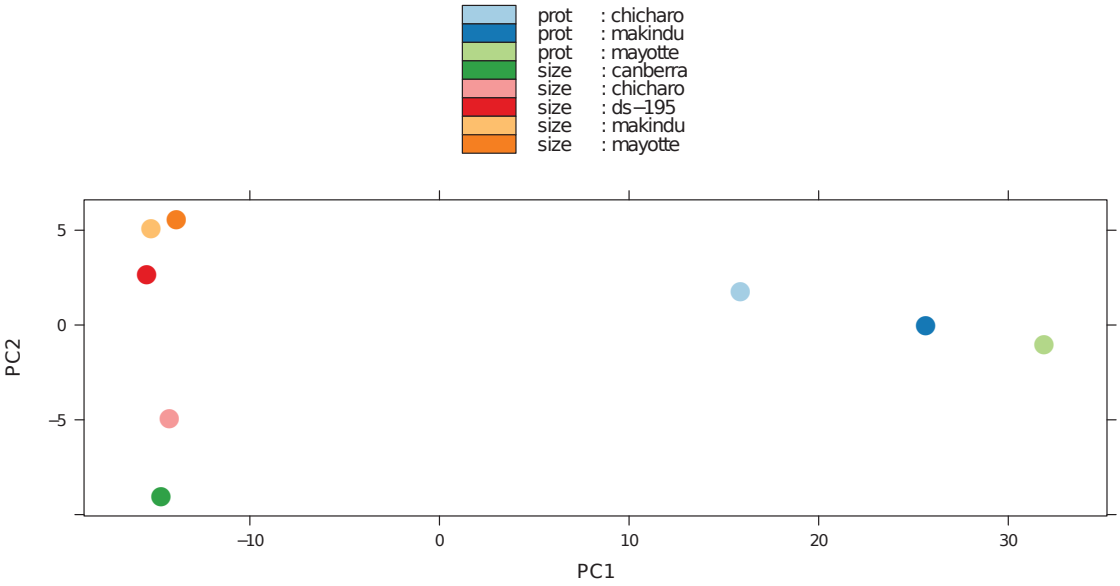


FIGURE 4.3 : Analyse en composantes principales des données de petits ARN. Chaque point représente une population avec “prot” pour le protocole d’extraction de protéine et “size” pour le protocole d’extraction par taille.

par les protéines Rhino, Cutoff et Deadlock peut venir tempérer la distinction qui existe entre ces deux types de séquences [130, 131].

3 Analyse des données d’ARN messager

Les données d’ARN totaux extraites des ovaires de drosophiles de cinq populations de *D. simulans* ont aussi été analysées avec le pipeline countTE. La même procédure que pour l’analyse des petits ARN a été utilisée, en utilisant Bowie2 à la place de Bowtie pour prendre en compte le fait que ces données sont de type paired-end.

Au total, 13 millions de séquences ont pu être alignées sur nos copies d’ET, ce qui représente environ 2% de notre jeu de données (voir Table 4.2). Ce faible pourcentage du jeu de données est à mettre en relation avec le fait que ce sont les ARN messagers totaux qui ont été extraits et que l’on ne s’attend pas à ce que l’expression associée aux séquences des copies d’ET représente la majorité des ARN messagers d’une cellule [242]. Cette analyse a permis de détecter une expression différentielle pour 6 ET entre les différentes populations. Ces éléments, ainsi que ces populations, sont présentés dans le tableau 4.3. La variabilité de certains de ces éléments, comme *Tirant* et *412*, avait

déjà été mise en évidence par des approches moléculaires [243, 244].

Tableau 4.2 : Données d'ARN messager et d'alignement sur les séquences des copies d'ET

population	réplicat	nombre de lectures	lectures alignées	pourcentage aligné
O.195	R1	83871434	1620864	1,93%
Chicharo	R1	74158274	1477004	1,99%
Makindu	R1	65259274	1420561	2,18%
Mayotte	R1	71648994	1322096	1,85%
Zimbabwe	R1	74453842	1454038	1,95%
O.195	R2	76490730	1403773	1,84%
Chicharo	R2	56784342	1040822	1,83%
Makindu	R2	60989658	1195946	1,96%
Mayotte	R2	65357738	1153773	1,77%
Zimbabwe	R2	101855344	1734526	1,70%
total		730869630	13823403	-

Ces résultats sont encore préliminaires, et il reste beaucoup d'information à extraire de ces données. Par exemple, il pourrait exister une association entre l'expression des ARN des ET et celle des petits ARN associés à ces ET. Nos données pourraient aussi permettre d'étudier la corrélation entre ces deux types de données d'expression et le nombre de copies des éléments caractérisés dans les populations naturelles de *D. simulans*.

Tableau 4.3 : Éléments transposables différentiellement exprimés entre les populations de *D. simulans*

couple dde population	ET	*p*-value ajustée
Chicharo - Makindu	412	$1{,}63\times10^{-12}$
Chicharo - Mayotte	412	$2{,}96\times10^{-6}$
O.195 - Makindu	412	$2{,}29\times10^{-9}$
O.195 - Mayotte	412	$6{,}08\times10^{-4}$
Makindu - Zimbabwe	412	$1{,}62\times10^{-4}$
Chicharo - Zimbabwe	ACCORD	$2{,}36\times10^{-3}$
O.195 - Zimbabwe	ACCORD	$1{,}23\times10^{-3}$
Makindu - Zimbabwe	ACCORD	$1{,}62\times10^{-4}$
Chicharo - Makindu	DM412	$3{,}90\times10^{-2}$
Chicharo - Makindu	DM412B	$9{,}12\times10^{-5}$
O.195 - Makindu	DM412B	$1{,}81\times10^{-3}$
O.195 - Zimbabwe	DOC6	$1{,}08\times10^{-2}$
Makindu - Mayotte	TIRANT	$2{,}00\times10^{-2}$

5

Procédure de contrôle qualité pour les analyses de données issues de séquençage de nouvelle génération

Comme nous l'avons vu dans le chapitre 1, les technologies de séquençage de nouvelle génération (NGS) ont permis de faire un véritable bond en avant dans la quantité des données exploitables en bioinformatique [245]. Ces technologies permettent d'obtenir différents types d'information comme des lectures de données génomiques (séquençage) [182], transcriptomiques (RNA-Seq), de méthylation de l'ADN (MeDIP-Seq) ou d'associations entre protéines et ADN (ChIP-Seq). Cependant, ces différentes technologies NGS ne sont pas exactes et font des erreurs dans les lectures qu'elles produisent [246]. Il est donc nécessaire de corriger ces erreurs quand cela est possible, et/ou de retirer les données erronées quand des corrections ne sont pas envisageables. Cette correction ou ce nettoyage des données fait partie de ce que l'on appelle l'étape de contrôle de la qualité des données de séquençage.

Les stratégies permettant de corriger les erreurs dans des données NGS, comme celles utilisées par le programme Quake [247], se basent sur le fait que ces erreurs ne sont pas systématiques dans les lectures produites. Par conséquent, une erreur de séquençage pour un nucléotide donné à une position du génome ne devrait pas être présente dans la majorité des lectures correspondant à cette portion du génome. Ce type d'approche utilise la profondeur de séquençage (nombre de lectures correspondant à une même séquence génomique) pour corriger les erreurs de séquençage. Pour cela, les lectures similaires

sont alignées entre elles et les nucléotides dont la fréquence dans ces lectures est trop faible (erreurs de lecture) sont remplacés par ceux dont la fréquence est élevée. Cependant, ce type d'approche présente plusieurs limitations. Elle nécessite un séquençage avec une profondeur de séquençage importante et n'est applicable que pour des données de séquençage génomique. Une méthode plus directe et plus largement utilisable pour nettoyer des données NGS consiste à identifier et retirer du jeu de données les lectures ou les portions de lecture de mauvaise qualité. C'est ce que l'on appelle le "trimming" des données NGS.

En plus de produire de nombreuses lectures de séquences d'ADN ou d'ARN, toutes les technologies NGS produisent des données sur la qualité de ces lectures. Ainsi, pour chaque nucléotide à chaque position de chaque lecture, on peut avoir accès à la probabilité que le nucléotide lu pour cette position soit correct. Cette probabilité est encodée sous la forme d'un score phred et correspond à la confiance que l'on peut accorder à un nucléotide. Le but des approches de "trimming" est d'éliminer les nucléotides dont le score phred est trop faible. Pour les technologies NGS les plus utilisées (Illumina et 454), ce score phred diminue avec la longueur de la lecture et la présence d'homopolymères. Ainsi, plus une lecture est longue, plus la probabilité pour chaque nouveau nucléotide séquencé d'être faux est élevée. C'est pourquoi les approches de "trimming" vont raccourcir les lectures en 3' pour essayer d'enlever ces nucléotides de mauvaise qualité. Dans les cas où il existe une succession de nucléotides identiques, la probabilité qu'un segment d'homopolymères contienne des erreurs de séquençage est aussi plus élevée.

Ces deux caractéristiques des données NGS se sont révélées problématiques pour le traitement des données de séquençage de piRNA qui ont été présentées dans le chapitre précédent. En effet, ces petits ARN ont une taille inférieure à 30 nucléotides alors que leur séquençage a été effectué par un séquenceur produisant des lectures de 160 à 180 nucléotides. Pour compléter les nucléotides manquant, ces piRNA ont donc été polyadénylisés à leur extrémité 3'. Pour pouvoir exploiter ces données, il a fallu enlever ces queues polyA pour avoir accès à la séquence des piRNA. Comme nous l'avons vu, ce problème est double puisque ces queues polyA sont des homopolymères de A et qu'elles constituent l'extrémité 3' des lectures produites. Pour enlever ces segments de A, j'ai donc développé une nouvelle méthode de segmentation non supervisée afin de déterminer le meilleur point de coupure entre un segment d'ADN indéterminé et une queue polyA. Cette méthode prend en compte la probabilité pour chaque nucléotide d'être correct afin de définir une queue polyA constituée de vrais A et de faux nucléotides T, C, et G qui pourraient en réalité correspondre à un A. J'ai ensuite étendu cette méthode au problème plus général du contrôle de qualité des données NGS. Dans ce cas, le problème

de segmentation correspond à trouver le meilleur point de coupure entre un segment de nucléotides de bonne qualité et un segment de nucléotides de mauvaise qualité.

Les méthodes de "trimming" existantes reposent sur deux types d'algorithmes : les algorithmes utilisant des fenêtres glissantes et ceux utilisant des sommes roulantes. Ces deux types d'algorithmes sont rapides mais produisent des points de coupures en considérant des variations locales de la qualité de la séquence et nécessitent plusieurs paramètres pour définir la qualité désirée après "trimming" [248]. Le "trimming" effectué n'est donc pas optimal en terme de la contrepartie effectuée entre le nombre de nucléotides retirés et la qualité de données obtenue. De plus l'utilisation de ces programmes nécessite des ajustements manuels des paramètres en fonction de la qualité des données traitées pour obtenir des résultats satisfaisants. Notre modèle de segmentation non supervisé est donc particulièrement adapté puisqu'il permet de trouver la meilleure segmentation globalement entre un segment de bonne qualité et un segment de mauvaise qualité à partir d'un seuil phred, au-dessus duquel une séquence est considérée comme étant de bonne qualité.

Ce travail a conduit à la rédaction d'un article :

- "UrQt : an efficient software for the Unsupervised Quality trimming of NGS data" soumis à BMC Bioinformatics.

et a fait l'objet d'une présentation orale à la conférence BGE à Lyon en 2014 intitulée :

- UrQt : unsupervised quality trimming of NGS data.

Le modèle probabilistique utilisé pour le "trimming" et pour enlever les queues polyA de données NGS est présenté dans cet article. Ce modèle a été implémenté dans le programme UrQt que j'ai écrit en C++ et qui utilise le calcul parallèle pour tirer profit des architectures multi-cœurs modernes. Dans cet article j'ai effectué une étude comparative entre UrQt et différents programmes de "trimming" de données NGS. Ces programmes ont été sélectionnés pour leur performance et leur représentativité des deux types d'algorithmes de "trimming" cités précédemment [248].

Les conclusions de ces tests comparatifs sont que UrQt permet d'obtenir des données de meilleure qualité que les autres programmes de "trimming". L'utilisation de UrQt est simple avec un unique paramètre pour définir un seuil phred au-delà duquel on considère qu'une séquence est de bonne qualité. De plus, les résultats fournis par UrQt sont indépendants de la qualité des données traitées, ce qui ne semble pas être le cas des autres procédures de "trimming". Ces deux caractéristiques –un unique paramètre pour gouverner la procédure et des résultats indépendants du type et de la qualité des

données traitées– peuvent donc permettre une meilleure automatisation de l'étape de contrôle de qualité des données NGS pour laquelle l'utilisateur peut définir un seuil de qualité voulue et simplement utiliser UrQt. UrQt pourrait donc facilement être intégré dans de nombreux pipelines bioinformatiques dans lesquels différents programmes sont exécutés les uns à la suite des autres pour automatiser certaines analyses.

UrQt : an efficient software for the Unsupervised Quality trimming of NGS data

Laurent Modolo et Emmanuelle Lerat

Soumis à BMC Bioinformatics

Modolo and Lerat *BMC Bioinformatics* (2015) 16:137
DOI 10.1186/s12859-015-0546-8

BMC Bioinformatics

SOFTWARE Open Access

UrQt: an efficient software for the Unsupervised Quality trimming of NGS data

Laurent Modolo and Emmanuelle Lerat*

Abstract

Background: Quality control is a necessary step of any Next Generation Sequencing analysis. Although customary, this step still requires manual interventions to empirically choose tuning parameters according to various quality statistics. Moreover, current quality control procedures that provide a "good quality" data set, are not optimal and discard many informative nucleotides. To address these drawbacks, we present a new quality control method, implemented in UrQt software, for Unsupervised Quality trimming of Next Generation Sequencing reads.

Results: Our trimming procedure relies on a well-defined probabilistic framework to detect the best segmentation between two segments of unreliable nucleotides, framing a segment of informative nucleotides. Our software only requires one user-friendly parameter to define the minimal quality threshold (phred score) to consider a nucleotide to be informative, which is independent of both the experiment and the quality of the data. This procedure is implemented in C++ in an efficient and parallelized software with a low memory footprint. We tested the performances of UrQt compared to the best-known trimming programs, on seven RNA and DNA sequencing experiments and demonstrated its optimality in the resulting tradeoff between the number of trimmed nucleotides and the quality objective.

Conclusions: By finding the best segmentation to delimit a segment of good quality nucleotides, UrQt greatly increases the number of reads and of nucleotides that can be retained for a given quality objective. UrQt source files, binary executables for different operating systems and documentation are freely available (under the GPLv3) at the following address: https://lbbe.univ-lyon1.fr/-UrQt-.html.

Keywords: Quality control, Trimming, Next-generation sequencing, Unsupervised segmentation, Parallel computing

Background

Next Generation Sequencing (NGS) technologies produce calling error probabilities for each sequenced nucleotide [1]. These probabilities, encoded as phred scores [2], are often high at the heads and tails of the reads, indicating low-quality nucleotides [3]. The presence of these unreliable nucleotides can result in missing or wrong alignments that can either increase the number of false negatives and false positives in subsequent analyses or can produce false *k*-mers in *de novo* assembly, increasing both the complexity of an assembly and the chance of producing misassemblies [4]. To remove these unreliable nucleotides and only work with informative nucleotides, most NGS data analyses start with a quality control (QC) step before any downstream analysis.

There are three types of approaches to address low-quality nucleotides. Classical QC strategies begin by removing an arbitrary number of nucleotides at the head and tail of each read, with tools such as the fastx_trimmer from the FASTX-Toolkit [5], after visualization of the per nucleotide sequence quality with tools such as FastQC [6]. Then, only reads of high quality are retained by other filters; for example, all reads with a given percentage of their length below a given phred score are excluded, using tools such as the fastq_quality_filter from FASTX-Toolkit. More recent approaches modify incorrectly called nucleotides by superimposing reads to each other and removing low frequency polymorphisms. This kind of approach often works using motifs of *k* nucleotides or *k*-mer to modify low frequency motifs based on the most frequent ones. However, this type of approach requires

*Correspondence: emmanuelle.lerat@univ-lyon1.fr
Université de Lyon; Université Lyon 1; CNRS; UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 bd du 11 novembre 1918, 69622 Villeurbanne cedex, France

potentially high sequencing coverage (15x in the case of Quake [7] and 100x in the case of ALLPATHS-LG [8]) and cannot be applied to non-uniform sequencing experiments, such as RNA sequencing (RNA-Seq). Other approaches trim unreliable nucleotides at the head and tail of each read. With these approaches, one wants to find the best trade-off between removing unreliable nucleotides and keeping the longest reliable or informative subsequence for the entire read. Current trimming approaches rely on two types of algorithms: the running sum algorithm and the window-based algorithm (for a review see [4]). However, these algorithms only return good local cutting points for each read when it is necessary to find a good global cutting point to get the best trade-off between removing unreliable nucleotides and losing too much information. Moreover, most of these QC strategies rely heavily on manually chosen parameters that are difficult to interpret and cannot be easily automatized.

In the present work, we have developed the program UrQt to trim unreliable nucleotides at the heads and tails of NGS reads based on their phred scores. We define an informative segment as a segment whose nucleotides are on average informative and an informative nucleotide as a nucleotide with a quality score above a specified quality threshold. Our approach takes advantage of the expected shape of the calling error probability along each read (abruptly decreasing for the first nucleotides and slowly increasing with the size of the reads) to find the best partition between two segments of unreliable nucleotides to be trimmed –the head and the tail of the reads– and a central informative segment. UrQt implements an unsupervised segmentation algorithm to find the best trimming cut-points in each read by maximum likelihood. We use a probabilistic model to handle more naturally the trimming problem than other procedures using window-based or running sum algorithms [4]. Moreover, UrQt requires no data-dependent parameters and takes advantage of modern multicore achitectures, which makes it particularly interesting to be routinely applied for NGS reads in fastq or fastq.gz format [9] and attractive for the development of future analytical pipelines.

Implementation

In this section, we present the probabilistic model that we use to find the best position to trim a read to increase its quality without removing more nucleotides than necessary. We also present an extension of this model for homopolymer trimming.

A read is defined as a vector $(n_1, \ldots, n_m)$ of m nucleotides associated with a vector of phred scores $(q_1, \ldots, q_m)$. We want to find the best cut-point $k_1 \in [1, m]$ in a read of length m between an informative segment for nucleotide $n_i, i \in [1, k_1]$ and a segment of unreliable quality for nucleotide $n_i, i \in [k_1 + 1, m]$ (Figure 1). Then, having found k_1, we want to find the best cut-point $k_2 \in [1, k_1]$ between a segment of unreliable quality for nucleotide $n_i, i \in [1, k_2 - 1]$ and an informative segment for nucleotide $n_i, i \in [k_2, k_1]$. Given the shape of the calling error probability distribution, there is less signal to find k_1 (the probability slowly increases at the extremity of the read) than k_2 (abruptly decreases). Therefore, we want to have the highest number of nucleotides to support the choice of k_1 when k_2 can be found with a subsequence of the read (Figure 1).

With q the quality value of a nucleotide, the probability for this nucleotide to be correct is defined by:

$$p_a(q) = 1 - 10^{\frac{-q}{10}} \quad \text{(A)}$$

which gives, for example, a probability $p_a(q) = 0.99$ for a phred $q = 20$ [2]. However, in QC, the word "informative" is typically defined as a phred score above a certain threshold and not the probability of calling the correct nucleotide. From a probabilistic point of view, we need to discriminate informative nucleotides (with $p_a(q) \geq p_a(t)$ and t a given threshold) from other nucleotides, rather than discriminate fairly accurate nucleotides (with $p_a(q) \geq 0.5$) from the others. Therefore, we propose to define the probability of having an informative nucleotide

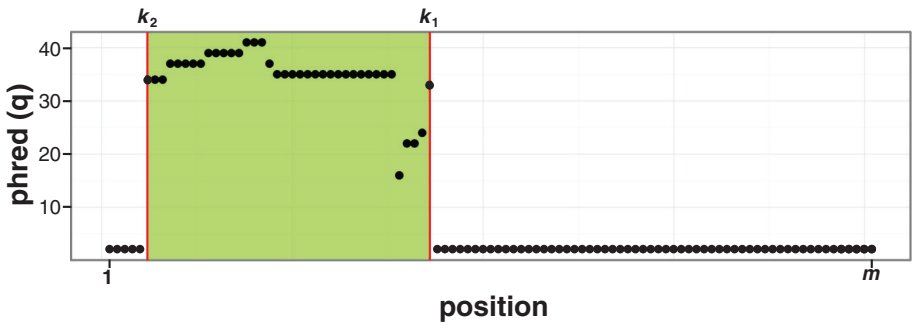


Figure 1 Quality trimming. Position of the cut-points k_1 and k_2 in a read. After trimming, the retained part corresponds to the section with a green background, which indicates an informative segment of nucleotides between k_1 and k_2.

as $p_b(q,t) = 1 - 2^{\frac{-q}{t}}$ with t the minimal phred score acceptable to be informative. This definition shifts the probability function such that for $q = t$, we have $p_b(q,t) = 0.5$. Therefore, at the threshold t, nucleotides with $p_b(q,t) \geq 0.5$ are informative and the others are not. With $t = 3.0103$, we go back to the classical phred function (Figure 2) in which $p_b(q,t) = p_a(q)$.

With the function $p_b(q,t)$, low phred scores are associated with a low probability to be correct ($p_b(0,t) = 0$), but for $t \leq 20$ a high phred score does not correspond to a high probability to be correct (for example, $p_b(40,20) = 0.75$). Therefore, from a probabilistic point of view, unreliable nucleotides will have more weight than informative ones. To associate a high phred score with a high probability of having an informative nucleotide, we constrain this probability to reach 1 for a phred score of 45 by using the following spline function (Figure 2):

$$p(q,t) = \begin{cases} 1 - 2^{\frac{-q}{t}} & \text{if } q \leq \max(20,t), \\ B(q^\star, p_1, p_2, 1, 1) & \text{otherwise} \end{cases} \quad \text{(B)}$$

with $B(q^\star, p_1, p_2, p_3, p_4)$ the cubic Bezier curve starting at p_1 toward p_2 and arriving at p_4 coming from the direction of p_3 for $q^\star \in [0,1]$. We have $p_1 = 1 - 2^{-\max(20,t)/t}$, $p_2 = g(1/3 \times (45 - \max(20,t)))$ with $g(q)$ the tangent to the function $1 - 2^{\frac{-q}{t}}$ in $\max(20,t)$. We scale the Bezier curve to the interval $[t,45]$ with $q^\star = (q-t)/(45-t)$. The constraint $\max(20,t)$ ensures that $\frac{d}{dq^\star} B(q^\star, p_1, p_2, p_3, p_4) < 0$ for $q^\star \in [0,1]$ (see Figure 2).

With the maximum likelihood framework, finding the position of the cut-point between a segment of informative nucleotides ($q > t$) and a segment of unreliable nucleotides ($q < t$) consists in estimating k_1 by:

$$\widehat{k_1} = \arg\max_k \prod_{i=1}^{k} \frac{1}{k} f_0(n_i,t) \prod_{i=k+1}^{m} \frac{1}{m-k-1} f_1(n_i,t) \quad \text{(C)}$$

with $f_0(n_i,t)$ the probability that the nucleotide n_i comes from the segment of informative nucleotides and $f_1(n_i,t)$ the probability that the nucleotide n_i comes from the segment of unreliable nucleotides for a given t. Such that:

$$f_0(n_i,t) = p(q_i,t) \prod_{N \in \Omega} \Pr(N)^{\mathbf{1}(n_i = N)} \quad \text{(D)}$$

$$f_1(n_i,t) = (1 - p(q_i,t)) \frac{1}{4} \quad \text{(E)}$$

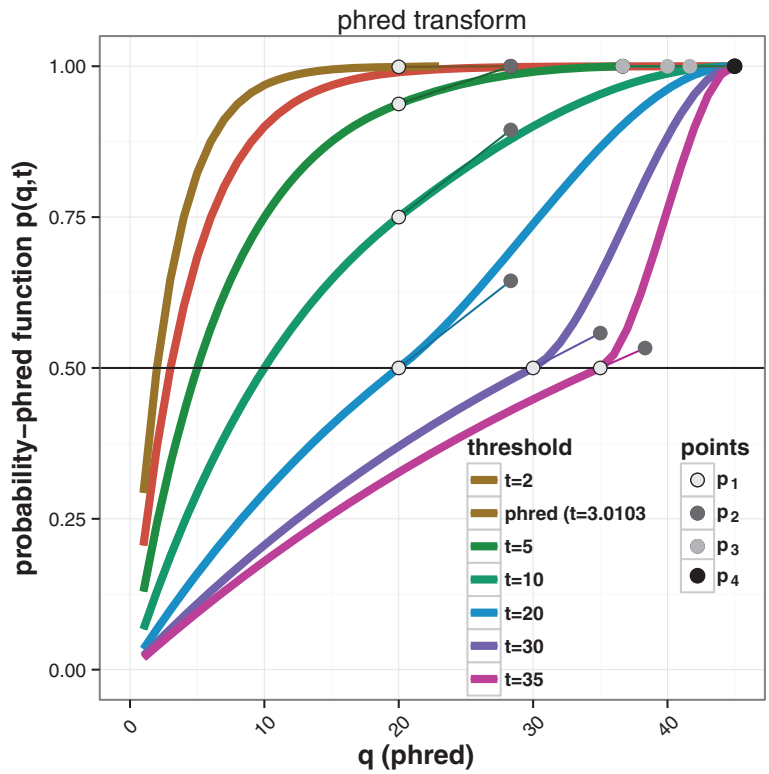


Figure 2 Probability-phred functions. $p(q,t)$ according to the choice of t. The white, dark grey, light grey and black dots represent respectively the position of p_1, p_2, p_3 and p_4 for the corresponding probability-phred functions. Before p_1 we have the $1 - 2^{\frac{-q}{t}}$ part of the function (B) and after p_1 the $B(q^\star, p_1, p_2, p_3, p_4)$ part of the function (B).

with $\mathbf{1}(n_i = N)$ an indicator variable such that $\mathbf{1}(n_i = N) = 1$ if the nucleotide n_i is equal to N and 0 otherwise, $\Pr(N) = \sum_{i=1}^{k} \mathbf{1}(n_i = N)/k$ the probability to observe the nucleotide N between 1 and k, and Ω the standard IUB/IUPAC dictionary [10]. $\Pr(N)_{N\in\Omega}$ and k_1 are estimated with the complete data framework of the EM algorithm [11]. After finding $\widehat{k_1}$, we apply the same procedure on the interval $[1, \widehat{k_1}]$ to estimate the best cut-point k_2 between a segment of unreliable nucleotides ahead of a segment of informative nucleotides. This double binary segmentation ensures to provide the best two cut-points for a given read [12].

For $p(q,t) = p_a(q)$, we can interpret the segment of informative nucleotides as a segment for which on average we are confident that a given nucleotide is the correct one, whereas the segment of unreliable nucleotides is composed of uninformative nucleotides in which on average any of the four nucleotides can be present at a given position. The cut-point k_1 maximizes the probability that the nucleotides $n_i, i \in [1, k_1]$ are informative and that nucleotides $n_i, i \in [k_1, m]$ are not.

With our model, trimming nucleotides of unreliable quality is somewhat similar to removing homopolymers from the extremities of the reads. The task of removing homopolymers, such as poly*A* tails in RNA-Seq experiments, is not trivial, because the quality of a given nucleotide decreases both at the end of the read and with the size of the homopolymer. Therefore, because the number of incorrectly called nucleotides increases, we are less likely to observe *A*s at the end of the poly*A* tail. UrQt implements a procedure for the unsupervised trimming of poly*N* with a straightforward modification of equation (E) such that:

$$f_1(n_i, t) = p_a(q_i, t)^{\mathbf{1}(n_i = A)} \left((1 - p_a(q_i, t)) \frac{1}{4} \right)^{\mathbf{1}(n_i \neq A)} \quad \text{(F)}$$

in which we can replace *A* by any letter of the standard IUB/IUPAC dictionary. With this definition of f_1, we consider the calling error probability of the nucleotide at position i if $n_i = A$ or if $n_i \neq A$, the probability that the nucleotide could have been an *A*.

Results and discussion

To assess the performance of our approach, we compared the performance of UrQt to other publicly available programs on different NGS data sets (see Table 1). The quality of the data generated during an NGS experiment can vary greatly depending on the type of data (DNA or RNA) and the sequencing pipeline. To analyze these two types of data on the same genome, we chose paired-end RNA and paired-end DNA sequencing experiments from the species *Drosophila melanogaster*. For this species, the DNA sample quality quickly drops at the end of the reads (see Additional file 1), and the RNA sample presents a large variability of quality among its reads. We also included in our analysis four other data sets from four different species which are the same ones as used in the comparative study of Del Fabbro et al. [4]. One single-end RNA sample from the species *Homo sapiens* of poor overall quality and one single-end RNA sample of good overall quality from the species *Arabidopsis thaliana*. For the DNA sample, we used one paired-end sample from the species *Prunus persica* of excellent overall quality and one paired-end DNA sample from the species *Saccharomyces cerevisiae* of average quality. Finally, we also included one paired-end RNA sample from the species *Homo sapiens* of overall good quality. The seven data sets (Table 1) were downloaded from the NCBI website. Rather than using the complete data set, we uniformly sampled 500,000 reads from each experiment using the software fastq_sampler.py (available at `https://github.com/l-modolo/fastq_sampler`), to speed-up the computation and work with comparable reads number for each sample.

For testing purposes, we choose the better trimming programs, according to their performances in the study of Del Fabbro et al. [4] and representing both running sum algorithms (Cutadapt [13], which implement the algorithm proposed for BWA [14]) and sliding-windows

Table 1 NGS data sets used for testing

Accession number	Species	Sample type	Paired-end	Read size (bp)	Reference genome
SRR002073	*Homo sapiens*	RNA	no	33	hg19
SRR521463	*Homo sapiens*	RNA	yes	75	hg19
SRR420813	*Arabidopsis thaliana*	RNA	no	83	TAIR10
SRX150254	*Prunus persica*	DNA	yes	100	1.22
SRR452441	*Saccharomyces cerevisiae*	DNA	yes	100	EF4
SRR988074	*Drosophila melanogaster*	DNA	yes	101	5.41
SRR919326	*Drosophila melanogaster*	RNA	yes	101	5.41

algorithms (Trimmomatic [15] and Sickle [16]). The different programs were compared on two points: the overall quality of the resulting trimmed data set and the number of reads mapped on the corresponding reference genome with Bowtie2 [17] for different quality thresholds. For the analyses presented in this work, we used the latest available versions of Cutadapt (version 1.4.1), Trimmomatic (version 0.32) and Sickle [16] (version 1.290). The value of the quality threshold *t* for the three programs, corresponded respectively to the parameter *–t* for UrQt, *–q* for Cutadapt and Sickle and *SLIDINGWINDOW:4:t* for Trimmomatic. All the other parameters were set to default values, except for the minimum read length that was set to 1 bp. All quality figures were generated with FastQC [6] and the quality statistics were computed using R [18] and the FASTX-Toolkit [5].

Consistency of the trimming procedures

It is expected that the quality in the trimmed data set will increase with the quality threshold up to a certain saturation point. We computed the median quality (phred) in the trimmed data for different quality thresholds (Figure 3, and Additional file 2 for the seven data sets). We observed from this comparison that except for UrQt, all other programs failed to produce a stable relationship between the chosen quality threshold and the resulting median quality score across different samples. For example, we observed a logarithmic-like relationship between the quality threshold and the median for data sets of overall poor quality, such as the *H. sapiens* data of overall poor quality, and an exponential-like relationship for data sets of overall good quality, such as the *A. thaliana* and the *S. cerevisiae* data (Figure 3). These different types of relationships indicate that an increase of the threshold does not have the same effect from one data set to another, and that this effect also depends on the value of the threshold. However, with UrQt, we observe a stable relationship between the threshold and the median quality that is representative of more consistent cutting-points. With a stable relationship between the threshold and the quality of the trimmed data

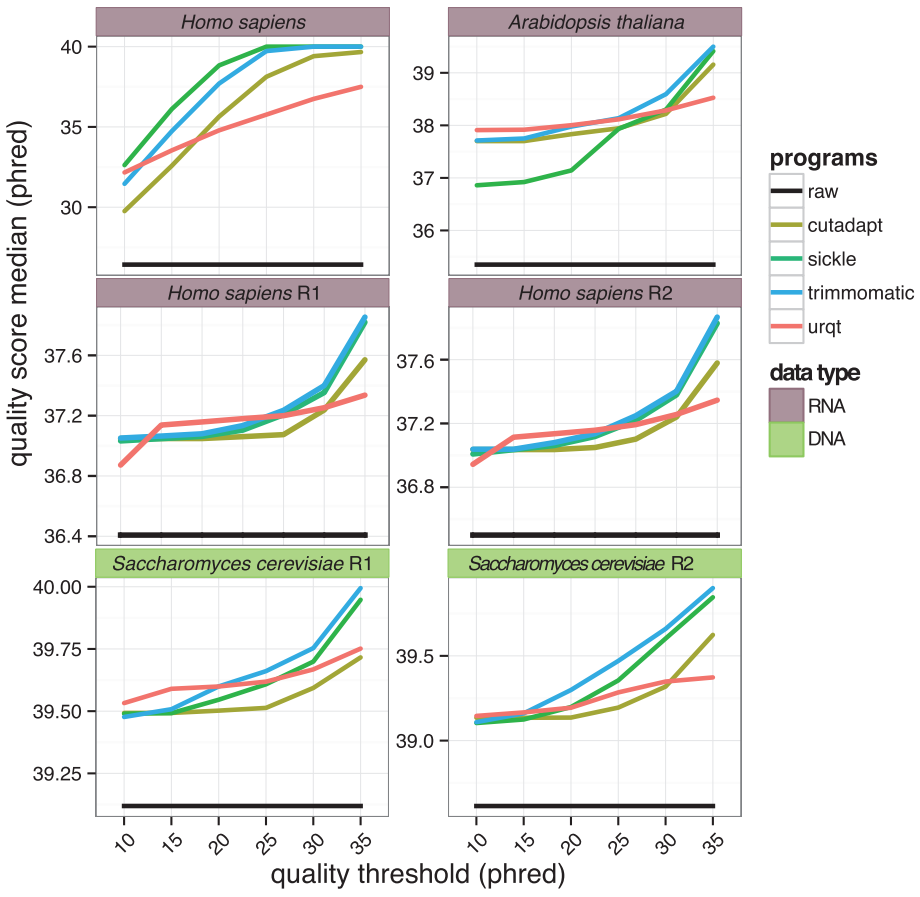


Figure 3 Quality of the trimmed data for each software. Performances of different trimming algorithms in terms of the median quality (phred) of the resulting trimmed data set for different quality thresholds. The choices of *t* correspond to the parameter *–t* for UrQt, *–q* for Cutadapt and Sickle and *SLIDINGWINDOW:4:t* for Trimmomatic. The black line corresponds to raw (untrimmed) data, and R1 and R2 correspond to the two ends of paired-end data.

set, it is thus possible to set the quality threshold beforehand according to a targeted quality and independently of the data.

Optimality of the trimming procedures

Although increasing the quality of a data set by trimming nucleotides of poor quality is easy, the remaining difficulty lies in minimizing the information (nucleotides) lost in the process. A simple metric to evaluate this trade-off is the number of trimmed reads that can be mapped on the corresponding reference genome. With better quality information after trimming, we expect an increase of the number of mapped reads, whereas by removing too many nucleotides, we expect less information and thus a decrease in the number of mapped reads. For the mapping procedure, we used Bowtie2 [17] (version 2.2.2) (with default parameters and the *–very-sensitive* option) and the genome indexes available from the igenome project (see Table 1 and Additional file 3 for the version). For the paired-end data, each end was mapped independently. We examined the number of mapped reads on the corresponding reference genomes (Table 1) for different quality thresholds (Figure 4 and Additional file 4 for the seven data sets). The same mapping procedure was also performed using BWA [14] (version 0.7.10) (with default parameters) (Additional file 3). We observed that UrQt was the only software that consistently increased the number of mapped reads for all data sets. The other programs provided the desired effect only for data sets of overall poor quality, such as for the single-end *H. sapiens* data (SRR002073), and produced worse results than those obtained by mapping the raw data for data sets of better quality (Figure 4). For the single-end *H. sapiens* data, we observed that UrQt better respected the chosen threshold, thus producing worst results than the other programs for the low quality threshold. For example, with this dataset and a threshold of 5, we expect a large number of reads with an average quality slightly above 5 which are difficult to map. This respect of the threshold can also be seen for the paired-end *H. sapiens* data (SRR521463) or the *D.*

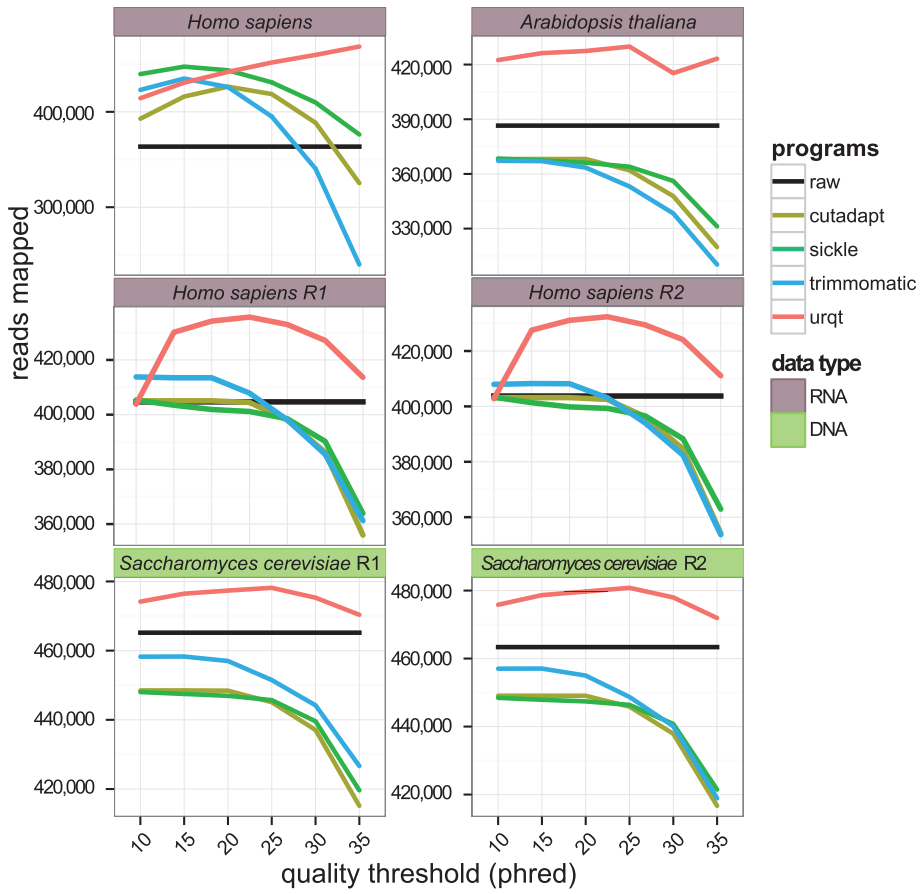


Figure 4 Remaining information in the trimmed data for each software. Mapping performances for different quality threshold. The choice of *t* corresponds to the parameter *–t* for UrQt, *–q* for Cutadapt and Sickle and *SLIDINGWINDOW:4:t* for Trimmomatic. The black line corresponds to raw (untrimmed) data, and R1 and R2 correspond to the two ends of paired-end data.

melanogaster RNA data (SRR919326) and a low threshold of 5 where UrQt is the only program that produces results comparable to the raw data (see Additional file 4).

For data sets of excellent quality, such as *P. persica* (see Additional file 4), all the trimming programs except for UrQt deteriorated the mapping performances compared with the ones obtained by mapping the raw data. This result provides additional evidence of better trimming cut-points identified by UrQt compared with the ones found by other procedures that remove too many nucleotides for data sets of excellent quality.

When considering the output of a mapping software, we can discriminate between reads, which map to a unique position and reads, which map to multiple positions. The number of reads mapping at multiple positions depends on three factors: the number of reads associated with repetition, the sensitivity of the mapping procedure (we can expect more reads mapping at multiple positions when allowing for more missmatches and gaps), and the information contained in the reads. Thus with trimming procedures, the information loss of over-trimming could lead to an increase of the number of reads mapping at multiple positions. This over-trimming effect can be seen with Cutadapt, Trimmomatic and Sickle for high threshold values (superior to 20) (see Additional file 3 for the results with Bowtie2 and BWA). However, with UrQt, the number of reads mapping to unique position increase with the choice of the threshold which is also consistent with better cut-point. These results hold for every dataset with the exception of the *H. sapiens* RNA sample of poor overall quality (SRR002073) for which removing a large number of uninformative nucleotides also correspond to removing a large number of reads.

Overall, the results obtained with UrQt correspond to the expected results for a trimming procedure and a given quality threshold in opposition to the other programs in our test panel (see Additional file 3 and 5). The output of UrQt depends on the choice of *t* that defines an informative sequence for which we expect nucleotides to have a phred score above this threshold. Contrary to current methods in which the choice of the threshold is set according to the quality of the data, the UrQt *–t* parameter only depends on the goal of the analysis (SNP calling, *de novo*-assembly, mapping, etc.).

Conclusions

UrQt is a new tool for the key QC step of any NGS data analysis to trim low-quality nucleotides and poly*A* tails from reads in fastq or fastq.gz format with an efficient C++ implementation. By finding the best segmentation to delimit a segment of informative nucleotides, UrQt greatly increases the number of reads and of nucleotides that can be retained for a given quality objective. Using this software should provide a significant gain for many NGS applications. Moreover, the consistency of our trimming procedure with the quality of the trimmed data set for a given quality threshold, will allow for better automation of the trimming step in a pipeline. We also provide a galaxy wrapper for UrQt to facilitate its integration in existing pipelines developed on this platform [19-21]. Finally, with our simple probabilistic model for the trimming of NGS data, we hope that users will have a better grasp on the quality threshold *–t* to obtain the largest trimmed data set with the required quality.

Availability and requirements

Project name: UrQt
Project home page: `https://lbbe.univ-lyon1.fr/-UrQt-.html`
Operating system(s): Platform independent
Programming language: C++
Other requirements: zlib and c++0x compiler
License: GNU GPLv3
Any restrictions to use by non-academics: GNU GPLv3

Additional files

Additional file 1: Quality analysis of the seven NGS samples. Quality analysis of the seven NGS samples (Table 1) with the FastQC software.

Additional file 2: Quality of the trimmed data for each programs. Performances of different trimming algorithms in terms of the median quality (phred) of the resulting trimmed data set for different quality thresholds. The choice of *t* corresponds to the parameter *–t* for UrQt, *–q* for Cutadapt and Sickle and *SLIDINGWINDOW:4:t* for Trimmomatic. The black line corresponds to raw (untrimmed) data, and R1 and R2 correspond to the two ends of paired-end data. This figure complements the Figure 3 with the seven data sets (Table 1).

Additional file 3: Mapping performances for the four tested programs. Mapping performances for different quality threshold with the four tested programs and the seven NGS sample (Table 1). Mapping results with Bowtie2 [17] and BWA [14].

Additional file 4: Remaining information in the trimmed data for each programs. Mapping performances for different quality threshold. The choice of *t* correspond to the parameter *–t* for UrQt, *–q* for Cutadapt and Sickle, and *SLIDINGWINDOW:4:t* for Trimmomatic. The black line corresponds to raw (untrimmed) data, and R1 and R2 correspond to the two ends of paired-end data. This figure complements the Figure 4 with the seven data sets (Table 1).

Additional file 5: Quality analysis of the seven NGS samples for the four tested programs. Quality analysis of the output of the four programs for the seven NGS samples and different quality thresholds (Table 1) with the FastQC [6] software.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Conceived and designed the experiments: LM and EL. Performed the experiments: LM. Analyzed and interpreted the data: LM; Drafted the manuscript: LM and EL. All authors read and approved the final version of the manuscript.

Acknowledgements
We thank V. Lacroix for his advice and discussions, and H. Lopez-Maestre and V. Romero-Soriano for their feedback in testing the software. We also thank the anonymous reviewers for their careful reading of our manuscript and their

many insightful comments and suggestions. The English of the manuscript has been edited by the American Journal Experts company.

Funding
This work was performed using the computing facilities of the CC LBBE/PRABI.

Received: 11 September 2014 Accepted: 20 March 2015
Published online: 29 April 2015

References

1. Ewing B, Hillier L, Wendl MC, Green P. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. Genome Res. 1998;8(3):175–85.
2. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998;8(3):186–94.
3. Datta SS, Kim S, Chakraborty S, Gill RS. Statistical analyses of next generation sequence data: a partial overview. J Proteomics Bioinf. 2010;3(6):183–90.
4. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLOS ONE. 2013;8(12):85024.
5. Lab H. FASTX Toolkit. 2011. http://hannonlab.cshl.edu/fastx_toolkit/.
6. Andrews S. Fastqc a quality control tool for high throughput sequence data. 2012. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
7. Kelley DR, Schatz MC, Salzberg SL. Quake: quality-aware detection and correction of sequencing errors. Genome Biol. 2010;11(11):116.
8. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A. 2011;108(4):1513–8.
9. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010;38(6):1767–71.
10. Cornish-Bowden A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res. 1985;13(9):3021–30.
11. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc. 1977;39(1):1–38.
12. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5(4):557–72.
13. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10.
14. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25(14):1754–60.
15. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
16. Joshi NA. Sickle: A Sliding-window, Adaptive, Quality-based Trimming Tool for FastQ Files. 2011. https://github.com/najoshi/sickle.
17. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
18. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. http://www.R-project.org.
19. Goecks J, Nekrutenko A, Taylor J, Team TG. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11(8):86.
20. Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, Mangan M, et al. Galaxy: A web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol. 2010. doi:10.1002/0471142727.mb1910s89.
21. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 2005;15(10):1451–5.

Conclusion

1 Conclusions générales

Comme nous l'avons vu le long de cette thèse, la présence des éléments transposables (ET) dans les génomes eucaryotes, ainsi que leur dynamique, peuvent avoir des conséquences importantes sur l'évolution des espèces. Les ET représentent une fraction extrêmement dynamique du génome de leurs hôtes et peuvent causer des changements importants au niveau de l'information génétique et épigénétique. Ces éléments sont caractérisés par leur capacité à transposer en différentes positions des génomes et par la multiplicité de leurs copies qui résultent de ces transpositions. La dynamique des ET est complexe et étroitement liée aux mécanismes de la régulation de leur transposition chez leurs hôtes. Le modèle de cycle de vie et de mort pour décrire cette dynamique est une théorie élégante pour décrire cette dynamique. Ce modèle propose que, après un événement de transfert horizontal (TH) introduisant un nouvel ET dans une espèce, cet élément devrait rapidement se multiplier et envahir les génomes des individus de cette espèce, avant que des mécanismes de régulation contrôlant son activité soient mis en place. Après un certain laps de temps, l'accumulation de mutations aléatoires devrait conduire à la perte des copies de cet élément par attrition.

Les TH et les voies de régulation des ET sont donc au cœur de la dynamique de ces éléments dans les génomes. Au cours de cette thèse, j'ai développé différents outils bioinformatiques pour étudier ces deux aspects importants de la dynamique des ET.

1.1 Détection de transferts horizontaux d'éléments transposables par approche comparative de génomes

Il existe de plus en plus d'exemples de TH d'ET entre des espèces eucaryotes, mais ces études utilisent des approches séquences spécifiques et sont retreintes à un faible nombre d'éléments. J'ai développé une méthode originale pour la détection de TH dans des génomes complets par analyse comparative de génomes, qui a permis de détecter de nouveaux cas de TH d'ET entre un organisme modèle *Drosophila melanogaster* et son espèce sœur *D. simulans*. Ces résultats illustrent que de nombreux TH d'ET auraient eu lieu entre ces deux espèces, ainsi qu'entre *D. melanogaster* et d'autres espèces de drosophiles plus ou moins proches. De plus, il semblerait que ces transferts soient souvent effectués en direction du génome de *D. melanogaster*, ce qui pourrait expliquer le fait que la plupart des copies d'ET sont jeunes et actives dans ce génome.

Ces résultats soulignent l'intérêt d'utiliser des approches pour génomes complets comme la nôtre, pour avoir une vision plus globale de la dynamique des ET chez les eucaryotes. Si cette approche nécessite peut-être de trop grosses ressources de calcul pour pouvoir être appliquée systématiquement entre toutes les paires possibles de génomes séquencés disponibles, elle pourrait cependant être appliquée à la liste des séquences de tous les ET identifiables dans ces génomes, pour pouvoir comparer systématiquement l'intégralité du contenu en ET entre deux génomes. L'utilisation d'autres programmes de recherche de séquences que BLASTN pourrait aussi être un moyen de diminuer le temps de calcul de cette étape. Ce type d'analyse permettrait d'avoir une vision plus globale des TH d'ET chez les eucaryotes. De plus, le test d'identité de séquences utilisé pour nos analyses pourrait être facilement modifié pour prendre en compte des mutations multiples pour un nucléotide donné, qui sont attendues après des temps de divergence plus grands que celui observé entre *D. simulans* et *D. melanogaster*. Cette modification pourrait alors permettre la détection de transferts plus anciens entre des espèces plus divergentes.

Ce travail sur la détection des TH a nécessité le développement d'une nouvelle procédure de correction pour les tests multiples unilatéraux. Cette méthode permet une plus grande puissance statistique, ce qui a permis la détection de nombreux cas de TH d'ET entre deux espèces proches. Cette procédure n'est cependant pas limitée à la problématique de la détection de TH et pourrait fournir une plus grande puissance de détection pour de nombreuses autres applications en biologie comme l'analyse différentielle de données d'expression par exemple.

1.2 Analyse de la régulation par la voie des "Piwi-interacting RNA"

En ce qui concerne l'étude de la régulation des ET par la voie des "Piwi-interacting RNA" présentée dans le chapitre 6, je n'ai pas pu détecter de différences significatives de quantité de piRNA associés aux différents ET entre les populations de *D. simulans*. Ce manque de significativité est sûrement dû à l'absence de vrais réplicats permettant d'isoler la variabilité technique de la variabilité biologique pour ces données. Néanmoins, ces données pourraient être exploitées pour étudier la position des clusters de piRNA dans le génome de *D. simulans*, ainsi que la distribution de ces petits ARN le long des séquences d'ET. Ce type d'analyse pourrait permettre d'identifier des portions de séquences d'ET exaptés par *D. simulans*, comme cela a été décrit pour les séquences de l'élément *LINE-1* chez l'Homme [249]. Dans cet exemple, une portion de l'ORF2 de cet élément est libre de piRNA, ce qui pourrait correspondre à une exaptation de cette partie codante de l'élément. Par ailleurs, même si la variabilité en contenu d'ET de *D. melanogaster* est inférieure à celle observée chez *D. simulans*, une étude récente sur le séquençage de piRNA dans 16 lignées différentes de *D. melanogaster* a produit des résultats similaires aux nôtres [250]. Les auteurs expliquent l'absence de différences entre la quantité de piRNA associés à un élément à fort nombre de copies et celle de piRNA associés au même élément en faible nombre de copies, comme la résultante de deux modèles. Dans un premier modèle, on s'attend à ce que la quantité de piRNA augmente en réponse à l'augmentation du nombre de copies d'un ET, ce qui devrait conduire à une corrélation positive entre le nombre de copies de l'élément et la quantité de piRNA associés. Par opposition, dans un second modèle, une plus grande production de piRNA devrait conduire à une répression de l'activité de l'ET, ce qui devrait amener à une corrélation négative entre le nombre de copies d'un ET et la quantité de piRNA associés. La combinaison de ces deux modèles pourrait décrire un modèle plus complexe expliquant cette absence de corrélation entre l'activité d'un ET et la quantité de piRNA associés.

Cette deuxième partie du travail de ma thèse a aussi conduit au développement d'un pipeline, countTE, pour l'analyse différentielle de données d'expression associées à des piRNA. Ce pipeline permet de calculer le nombre de petits ARN ou ARN messagers associés à une séquence donnée à partir de données de séquençage et d'une liste de séquences de copies d'ET. A la différence des autres approches utilisées par d'autres analyses de ce type dans la littérature, toutes les copies d'un ET sont utilisées pour calculer le nombre de petits ARN associés à cet élément et non pas une seule référence, ce qui permet de pouvoir analyser les petits ARN qui s'alignent à plusieurs endroits du

génome [116]. De plus, notre pipeline utilise la méthode de normalisation des comptages proposée par [237] qui est beaucoup moins biaisée que les méthodes de normalisation à 1 million de lectures ou de type RPKM (pour l'anglais "Reads Per Kilobase per Million reads"), généralement utilisées dans les études sur les piRNA [239]. Ce pipeline pourrait donc facilement être utilisé dans le futur pour d'autres analyses de ce type, concernant les piRNA ou l'expression des ET.

Enfin, ce travail sur la régulation des ET entre différentes populations de *D. simulans* a aussi conduit au développement d'un nouveau programme de "trimming" de données de séquençage de nouvelle génération (NGS) : UrQt. À l'inverse d'autres programmes de ce type, UrQt est simple d'utilisation avec un seul paramètre de qualité à choisir par l'utilisateur, et ses performances sont indépendantes de la qualité ou du type de jeu de données traité. Ces différents avantages pourraient permettre d'obtenir des jeux de données de meilleure qualité pour de nombreuses applications des technologies NGS en biologie.

2 Perspectives

Comme nous l'avons vu tout au long de cette thèse, les approches bioinformatiques sont un outil puissant pour l'étude de la dynamique des ET dans les génomes. Cependant, la multiplicité et la dynamique des séquences d'ET est souvent négligée par ces approches. Ceci pourrait être dû à l'influence des méthodes développées pour étudier les gènes qui par opposition aux ET, sont statiques et ont généralement peu de copie(s). La multiplicité des copies d'ET est une caratéristique importante de leur étude et devrait donc être systématiquement basée sur l'étude de leurs copies dans les génomes.

2.1 Annotation et classification des éléments transposables

Un des exemples les plus frappants de ce problème est illustré par le fait que la plupart des bases de données de séquences d'ET ne stockent pas la séquence de toutes les copies des éléments détectés dans une espèce donnée, mais une séquence consensus [8, 251]. Ce type de représentation permet sûrement une meilleure compatibilité avec les formats standards développés pour des séquences de gènes, comme EMBL ou GenBank, mais elle conduit aussi à une perte importante d'information sur les ET. Si ce type de données est peut-être suffisant pour masquer les séquences d'ET, il est souvent insuffisant pour pouvoir étudier leur dynamique. Par exemple, dans le cas des mécanismes de régulation par la voie des piRNA, l'information perdue dans la construction de la séquence consen-

sus pourrait être suffisante pour perdre la complémentarité de séquence avec un petit ARN de 21 paires de bases (pb). Avec l'augmentation massive de la quantité de données génomiques associée à l'utilisation des technologies NGS, il devient donc indispensable de construire des bases de données de séquences de copies d'ET pour pouvoir étudier plus facilement leur dynamique. Ces technologies permettent de générer de plus en plus de données populationnelles et ces nouvelles bases de données pourraient aussi contenir des informations sur les sites d'insertion des différentes copies d'un ET dans une population [252]. Ce type de données pourrait faciliter la datation et l'estimation du degré d'activité des différents ET présents dans différents génomes. La prise en compte de l'environnement génétique de chacune de ces insertions pourrait aussi permettre la classification de ces copies selon d'autres critères comme, par exemple, celui de leur intérêt adaptatif, avec des métriques similaires à celles proposées par Jurka et al. [173] qui compare la distribution du ratio du nombre de copies dans des régions conservées [253] au nombre de copies total dans le génome. Enfin, la construction de telles bases de données pourrait aussi permettre une meilleure caractérisation des différences structurales ainsi que des relations qui existent entre différents ET. Cette source d'information pourrait conduire à l'amélioration de la classification automatique des ET en utilisant des méthodes d'apprentissage automatique qui pourraient à leur tour faciliter la caractérisation et la classification de nouvelles copies d'ET [254].

2.2 Méthodes de détection spécifiques aux transferts horizontaux d'éléments transposables

Un autre exemple du problème lié à l'utilisation de méthodes développées pour les gènes est celui de la détection des transferts horizontaux d'ET. Jusqu'à présent la détection de TH par des méthodes bioinformatiques a reposé sur trois types de preuves, comme nous l'avons évoqué dans le chapitre 2 [158]. Un premier type de preuves peut être d'ordre phylogénétique, avec l'analyse de différences de topologie entre l'arbre phylogénétique des séquences étudiées et celui des espèces. Un second type de preuves repose sur une identité nucléotidique entre les séquences étudiées plus importante que celle attendue entre les deux espèces. Enfin le troisième type de preuves repose sur une distribution irrégulière de la séquence étudiée dans la phylogénie des espèces. Généralement, une détection de TH est jugée valide si elle est confirmée par au moins deux de ces approches pour augmenter la puissance de la détection. Les outils d'analyse utilisant ces trois types de preuves ont été développés pour la détection de TH de gènes chez les procaryotes. Il existe cependant une différence fondamentale entre la détection de TH de gènes chez les procaryotes et

celle d'ET chez les eucaryotes. En effet, il existe une grande différence au niveau de la taille des génomes et de la complexité du problème à traiter comme nous l'avons vu dans le chapitre 2, mais il existe avant tout une différence fondamentale entre les types d'éléments génétiques étudiés. Les séquences de gènes sont généralement uniques et statiques dans un génome procaryote, alors que les séquences d'ET sont multiples et mobiles.

Comme nous l'avons vu dans le chapitre 2 avec la présentation de l'"activity track", une explosion de transposition d'un ET dans un génome est un signal important que l'on peut s'attendre à détecter après un événement de TH. Certaines études sur les TH présentent parfois des arbres phylogénétiques des copies d'un élément donné pour illustrer la dynamique de cet ET. Une phylogénie en étoile de ces copies est attendue après une explosion de transposition [255]. Malheureusement, ce type d'analyse est pour l'instant uniquement qualitative [256]. De plus, il faut garder à l'esprit que l'histoire de certains ET est complexe et peut être ponctuée des différentes explosions de transpositions dont l'origine peut correspondre à des copies différentes de cet élément [257]. Ces histoires complexes pourraient par exemple expliquer la présence de différents variants pour un même ET [203].

Dans la présentation de l'"activity track" au chapitre 2, nous avons aussi évoqué le fait que la multiplicité des copies pour un élément donné puisse conduirent à l'observation de deux copies plus identiques que l'attendu par hasard entre deux génomes. Pour décrire ce phénomène, nous pouvons faire un parallèle entre la détection du TH d'une copie d'un élément entre deux espèces, quand cet élément est présent en plusieurs copies, et la problématique des tests multiples. Dans le cas des gènes, la détection d'un TH via son identité de séquence revient à effectuer un test statistique pour contrôler si cet élément est plus identique que l'attendu entre deux espèces. Dans le cas d'un ET, la détection d'une copie plus identique que l'attendu entre deux espèces correspond à un test positif parmi de nombreux tests, correspondant au nombre de copies de cet élément. Il est donc crucial de prendre en compte cet aspect des séquences d'ET pour étudier leurs TH.

Pour pouvoir prendre en compte ces différentes caractéristiques des ET dans la détection des TH, il faudrait utiliser des approches intégratives utilisant toute l'information disponible à partir des copies d'un élément entre deux espèces. En effet, les trois types d'approches séquence-spécifique sont étroitement liées entre elles. Par exemple deux séquences similaires vont nécessairement être proches dans un arbre phylogénétique reconstruit par des méthodes de parcimonie. Construire un arbre phylogénétique de séquences revient à hiérarchiser ces séquences entre elles en fonction de leur divergence, mais aussi en tenant compte de la position des sites divergents dans ces séquences. Les méthodes

de comparaison de topologies d'arbres phylogénétiques utilisent des métriques qui résument la géométrie de l'arbre et qui contiennent donc une part de l'information utilisée par les méthodes basées sur la divergence de séquences [258, 259, 260]. Il existe aussi un lien étroit entre les approches basées sur des différences de topologies des arbres phylogénétiques et celles basées sur une distribution inégale de la séquence étudiée dans les différentes feuilles de l'arbre. En effet, pour ce dernier type d'approches, même si elle est qualitative, l'information analysée correspond bien à une différence de topologie avec la topologie attendue dans le cas d'une transmission verticale de la séquence étudiée.

Un exemple d'approche intégrative développée pour étudier la dynamique d'un ET a été récemment publié [195]. La méthode AnTE repose sur un modèle probabiliste pour identifier les copies d'un élément correspondant à des copies "maîtres" ("master copy") et inférer l'histoire de la dynamique de cet ET dans un génome. Une copie "maître" est définie comme une copie dont l'activité de transposition va produire de nombreuses autres copies. Par opposition, l'activité de transposition des autres copies est considérée comme inexistante, ce qui pourrait correspondre à des insertions dans des régions hétérochromatiques du génome. Ce modèle utilise la différence de fréquence qui existe entre les mutations aléatoires qui vont s'accumuler dans la séquence de toutes les copies (et former des phylogénies en étoile après une explosion de transposition), et les mutations présentes dans la séquence des copies "maîtres" qui vont aussi être retrouvées dans toutes les copies qu'elles ont produites. En utilisant cette information, AnTE permet de calculer la probabilité qu'une copie donnée soit une copie "maître", et permet la reconstruction des copies "maîtres" ancestrales qui ne sont plus présentes dans le génome. Ce type d'approche pourrait être étendu au calcul de la probabilité qu'une ou plusieurs copies "maître" d'un ET correspondent à des copies présentes dans le génome d'une autre espèce ou population.

Bibliographie

[1] Ohno S (1972) So much "junk" DNA in our genome. Brookhaven symposia in biology 23 : 366–70.

[2] Dawkins R (2006) The Selfish Gene : 30th Anniversary Edition. OUP Oxford, 360 pp.

[3] Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284 : 601–3.

[4] Shapiro JA, Adhya SL (1969) The galactose operon of e. coli k-12. ii. a deletion analysis of operon structure and polarity. Genetics 62 : 249–264.

[5] Böhne A, Brunet F, Galiana-Arnoux D, Schultheis C, Volff JN (2008) Transposable elements as drivers of genomic and biological diversity in vertebrates. Chromosome Res 16 : 203–215.

[6] Biémont C, Vieira C (2006) Genetics : junk DNA as an evolutionary force. Nature 443 : 521–524.

[7] Biémont C (2010) A brief history of the status of transposable elements : from junk DNA to major players in evolution. Genetics 186 : 1085–93.

[8] Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research 110 : 462–7.

[9] Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Annual review of genetics 41 : 331–68.

[10] McDonald JF (1998) Transposable elements, gene silencing and macroevolution. Trends Ecol Evol 13 : 94–95.

[11] Finnegan DJ (1997) Transposable elements : how non-LTR retrotransposons do it. Current Biology 7 : R245–R248.

[12] Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, et al. (2007) A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8 : 973–982.

[13] Schnable PS, Ware D, Fulton RS, Al (2009) The B73 maize genome : complexity, diversity, and dynamics. Science 326 : 1112–1115.

[14] Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al. (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature 450 : 203–18.

[15] Pritham EJ (2009) Transposable elements and factors influencing their success in eukaryotes. J Hered 100 : 648–655.

[16] Kim J, Martignetti Ja, Shen MR, Brosius J, Deininger P (1994) Rodent BC1 RNA gene as a master gene for ID element amplification. Proc Natl Acad Sci U S A 91 : 3607–11.

[17] Song SU, Gerasimova T, Kurkulos M, Boeke JD, Corces VG (1994) An env-like protein encoded by a drosophila retroelement : evidence that gypsy is an infectious retrovirus. Genes Dev 8 : 2046–2057.

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

[19] Poulter RTM, Goodwin TJD (2005) Dirs-1 and the other tyrosine recombinase retrotransposons. Cytogenet Genome Res 110 : 575–588.

[20] Goodwin TJD, Butler MI, Poulter RTM (2003) Cryptons : a group of tyrosine-recombinase-encoding dna transposons from pathogenic fungi. Microbiology 149 : 3099–3109.

[21] Rebollo R, Horard B, Hubert B, Vieira C (2010) Jumping genes and epigenetics : Towards new species. Gene 454 : 1–7.

[22] Malik HS, Burke WD, Eickbush TH (1999) The age and evolution of non-LTR retrotransposable elements. Molecular biology and evolution 16 : 793–805.

[23] Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked Alu sequences. Nature genetics 35 : 41–48.

[24] Kramerov DA, Vassetzky NS (2005) Short retroposons in eukaryotic genomes. International Review of Cytology 247 : 165–221.

[25] Evgen'ev MB, Zelentsova H, Shostak N, Kozitsina M, Barskyi V, et al. (1997) Penelope, a new family of transposable elements and its possible role in hybrid dysgenesis in Drosophila virilis. Proc Natl Acad Sci U S A 94 : 196–201.

[26] Kojima KK, Jurka J (2011) Crypton transposons : identification of new diverse families and ancient domestication events. Mob DNA 2 : 12.

[27] Kapitonov VV, Jurka J (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. Nature Reviews Genetics 9 : 411–412.

[28] Jiang N, Feschotte C, Zhang X, Wessler SR (2004) Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). Curr Opin Plant Biol 7 : 115–119.

[29] Feschotte C, Wessler SR (2001) Treasures in the attic : rolling circle transposons discovered in eukaryotic genomes. Proc Natl Acad Sci U S A 98 : 8923–4.

[30] Kapitonov VV, Jurka J (2001) Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci U S A 98 : 8714–9.

[31] Kapitonov VV, Jurka J (2007) Helitrons on a roll : eukaryotic rolling-circle transposons. Trends in Genetics 23 : 521–529.

[32] Yang L, Bennetzen JL (2009) Distribution, diversity, evolution, and survival of Helitrons in the maize genome. Proc Natl Acad Sci U S A 106 : 19922–7.

[33] Li Y, Dooner HK (2009) Excision of Helitron transposons in maize. Genetics 182 : 399–402.

[34] Feschotte C, Pritham EJ (2005) Non-mammalian c-integrases are encoded by giant transposable elements. Trends in genetics 21 : 551–2.

[35] Pritham EJ, Putliwala T, Feschotte C (2007) Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene 390 : 3–17.

[36] Fischer MG, Suttle CA (2011) A virophage at the origin of large DNA transposons. Science 332 : 231–234.

[37] Maxwell P (2014) Consequences of ongoing retrotransposition in mammalian genomes. Advances in Genomics and Genetics : 129.

[38] Brandt J, Schrauth S, Veith AM, Froschauer A, Haneke T, et al. (2005) Transposable elements as a source of genetic innovation : expression and evolution of a family of retrotransposon-derived neogenes in mammals. Gene 345 : 101–111.

[39] Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nature reviews Genetics 9 : 397–405.

[40] Alzohairy AM, Gyulai G, Jansen RK, Bahieldin A (2013) Transposable elements domesticated and neofunctionalized by eukaryotic genomes. Plasmid 69 : 1–15.

[41] Kapitonov VV, Jurka J (2005) Rag1 core and v(d)j recombination signal sequences were derived from transib transposons. PLoS Biol 3 : e181.

[42] Ono R, Nakamura K, Inoue K, Naruse M, Usami T, et al. (2006) Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. Nature genetics 38 : 101–6.

[43] Sekita Y, Wagatsuma H, Nakamura K, Ono R, Kagami M, et al. (2008) Role of retrotransposon-derived imprinted gene, Rtl1, in the feto-maternal interface of mouse placenta. Nature genetics 40 : 243–8.

[44] Dupressoir A, Marceau G, Vernochet C, Bénit L, Kanellopoulos C, et al. (2005) Syncytin-a and syncytin-b, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in muridae. Proc Natl Acad Sci U S A 102 : 725–730.

[45] Pardue ML, DeBaryshe PG (2003) Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. Annu Rev Genet 37 : 485–511.

[46] Cowan RK, Hoen DR, Schoen DJ, Bureau TE (2005) MUSTANG is a novel family of domesticated transposase genes found in diverse angiosperms. Molecular biology and evolution 22 : 2084–9.

[47] Casacuberta E, González J (2013) The impact of transposable elements in environmental adaptation. Molecular ecology 22 : 1503–1517.

[48] Aminetzach YT, Macpherson JM, Petrov DA (2005) Pesticide resistance via transposition-mediated adaptive gene truncation in Drosophila. Science 309 : 764–767.

[49] Kim YB, Oh JH, McIver LJ, Rashkovetsky E, Michalak K, et al. (2014) Divergence of Drosophila melanogaster repeatomes in response to a sharp microclimate contrast in Evolution Canyon, Israel. Proceedings of the National Academy of Sciences 111 : 10630–10635.

[50] González J, Karasov TL, Messer PW, Petrov DA, Al (2010) Genome-wide patterns of adaptation to temperate environments associated with transposable elements in Drosophila. PLoS Genet 6 : e1000905.

[51] Oliver KR, Greene WK (2009) Transposable elements : powerful facilitators of evolution. BioEssays 31 : 703–14.

[52] O'Donnell KA, Burns KH (2010) Mobilizing diversity : transposable element insertions in genetic variation and disease. Mobile DNA 1 : 21.

[53] Babcock M, Pavlicek A, Spiteri E, Kashork CD, Ioshikhes I, et al. (2003) Shuffling of genes within low-copy repeats on 22q11 (lcr22) by alu-mediated recombination events during evolution. Genome Res 13 : 2519–2532.

[54] Feschotte C, Pritham EJ (2009) A cornucopia of Helitrons shapes the maize genome. Proc Natl Acad Sci U S A 106 : 19747–8.

[55] Lerman DN, Michalak P, Helin aB, Bettencourt BR, Feder ME (2003) Modification of Heat-Shock Gene Expression in Drosophila melanogaster Populations via Transposable Elements. Molecular Biology and Evolution 20 : 135–144.

[56] Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends in genetics 19 : 68–72.

[57] Rebollo R, Karimi MM, Bilenky M, Gagnier L, Miceli-Royer K, et al. (2011) Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. PLoS Genet 7 : e1002301.

[58] Hollister JD, Smith LM, Guo YL, Ott F, Weigel D, et al. (2011) Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata. Proc Natl Acad Sci U S A 108 : 2322–7.

[59] Gray YH (2000) It takes two transposons to tango : transposable-element-mediated chromosomal rearrangements. Trends in genetics 16 : 461–8.

[60] Castro JP, Carareto CMA (2004) Drosophila melanogaster P transposable elements : mechanisms of transposition and regulation. Genetica 121 : 107–118.

[61] Venner S, Feschotte C, Biémont C (2009) Dynamics of transposable elements : towards a community ecology of the genome. Trends in genetics 25 : 317–23.

[62] Koonin EV (2012) A half-century after the molecular clock : new dimensions of molecular evolution. EMBO reports 13 : 664–6.

[63] Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, et al. (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature 461 : 1130–1134.

[64] Hickey DA (1982) Selfish dna : a sexually-transmitted nuclear parasite. Genetics 101 : 519–531.

[65] Lynch M, Conery JS (2003) The origins of genome complexity. Science 302 : 1401–4.

[66] Ellison CE, Bachtrog D (2013) Dosage compensation via transposable element mediated rewiring of a regulatory network. Science 342 : 846–50.

[67] Willetts N, Skurray R (1980) The conjugation system of F-like plasmids. Annual review of genetics 14 : 41–76.

[68] Toussaint JVH, Ariane (2012) Baterial Molecular Networks. 542 pp. doi :10.1007/978-1-61779-361-5.

[69] Boutin TS, Le Rouzic A, Capy P (2012) How does selfing affect the dynamics of selfish transposable elements ? Mobile DNA 3 : 5.

[70] Charlesworth B, Sniegowski P, Stephan W (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371 : 215–20.

[71] Le Rouzic A, Capy P (2005) The first steps of transposable elements invasion : parasitic strategy vs. genetic drift. Genetics 169 : 1033–1043.

[72] Le Rouzic A, Capy P (2006) Population genetics models of competition between transposable element subfamilies. Genetics 174 : 785–93.

[73] Castillo DM, Mell JC, Box KS, Blumenstiel JP (2011) Molecular evolution under increasing transposable element burden in Drosophila : a speed limit on the evolutionary arms race. BMC evolutionary biology 11 : 258.

[74] Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nature reviews Genetics 8 : 272–85.

[75] Goldberg AD, Allis CD, Bernstein E (2007) Epigenetics : a landscape takes shape. Cell 128 : 635–8.

[76] Strahl BD, Allis CD (2000) The language of covalent histone modifications. Nature 403 : 41–45.

[77] Fan L, Roberts Va (2006) Complex of linker histone H5 with the nucleosome and its implications for chromatin packing. Proc Natl Acad Sci U S A 103 : 8384–9.

[78] Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, et al. (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. Nature genetics 40 : 897–903.

[79] Bekker-Jensen S, Mailand N (2010) Assembly and function of DNA double-strand break repair foci in mammalian cells. DNA Repair 9 : 1219–1228.

[80] Greer EL, Shi Y (2012) Histone methylation : a dynamic mark in health, disease and inheritance. Nature reviews Genetics 13 : 343–57.

[81] Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. (2007) High-Resolution Profiling of Histone Methylations in the Human Genome. Cell 129 : 823–837.

[82] Campos EI, Reinberg D (2009) Histones : annotating chromatin. Annual review of genetics 43 : 559–599.

[83] Grewal SIS, Elgin SCR (2007) Transcription and RNA interference in the formation of heterochromatin. Nature 447 : 399–406.

[84] Martens JH, O'Sullivan RJ, Braunschweig U, Opravil S, Radolf M, et al. (2005) The profile of repeat-associated histone lysine methylation states in the mouse epigenome. The EMBO journal 24 : 800–12.

[85] Peng JC, Karpen GH (2007) H3K9 methylation and RNA interference regulate nucleolar organization and repeated DNA stability. Nature cell biology 9 : 25–35.

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

[87] Vanyushin BF, Ashapkin VV (2011) DNA methylation in higher plants : past, present and future. Biochimica et biophysica acta 1809 : 360–8.

[88] Klose RJ, Bird AP (2006) Genomic DNA methylation : The mark and its mediators. Trends in Biochemical Sciences 31 : 89–97.

[89] Reddington JP, Pennings S, Meehan RR (2013) Non-canonical functions of the DNA methylome in gene regulation. The Biochemical journal 451 : 13–23.

[90] Suzuki MM, Bird A (2008) DNA methylation landscapes : provocative insights from epigenomics. Nature reviews Genetics 9 : 465–76.

[91] Boffelli D, Takayama S, Martin DIK (2014) Now you see it : Genome methylation makes a comeback in Drosophila. BioEssays .

[92] Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462 : 315–22.

[93] Laurent L, Wong E, Li G, Huynh T, Tsirigos A, et al. (2010) Dynamic changes in the human methylome during differentiation. Genome research 20 : 320–31.

[94] Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature reviews Genetics 11 : 204–20.

[95] Maksakova IA, Mager DL, Reiss D (2008) Keeping active endogenous retroviral-like elements in check : the epigenetic perspective. Cell Mol Life Sci 65 : 3329–3347.

[96] Howard G, Eiges R, Gaudet F, Jaenisch R, Eden A (2008) Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. Oncogene 27 : 404–408.

[97] Carmell MA, Xuan Z, Zhang MQ, Hannon GJ (2002) The Argonaute family : Tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. Genes and Development 16 : 2733–2742.

[98] Piriyapongsa J, Rutledge MT, Patel S, Borodovsky M, Jordan IK (2007) Evaluating the protein coding potential of exonized transposable element sequences. Biol Direct 2 : 31.

[99] Wang Q, Carmichael GG (2004) Effects of length and location on the cellular response to double-stranded RNA. Microbiology and molecular biology reviews 68 : 432–452.

[100] Castel SE, Martienssen RA (2013) RNA interference in the nucleus : roles for small RNAs in transcription, epigenetics and beyond. Nature reviews Genetics 14 : 100–12.

[101] Piatek MJ, Werner A (2014) Endogenous siRNAs : regulators of internal affairs. Biochemical Society transactions 42 : 1174–9.

[102] Cecere G, Grishok A (2014) A nuclear perspective on RNAi pathways in metazoans. Biochimica et Biophysica Acta - Gene Regulatory Mechanisms 1839 : 223–233.

[103] Chung WJ, Okamura K, Martin R, Lai EC (2008) Endogenous RNA interference provides a somatic defense against Drosophila transposons. Current biology 18 : 795–802.

[104] Czech B, Malone CD, Zhou R, Stark A, Schlingeheyde C, et al. (2008) An endogenous small interfering RNA pathway in Drosophila. Nature 453 : 798–802.

[105] Luteijn MJ, Ketting RF (2013) PIWI-interacting RNAs : from generation to transgenerational epigenetics. Nature reviews Genetics 14 : 523–34.

[106] Malone CD, Brennecke J, Dus M, Stark A, Mccombie WR, et al. (2009) Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. Cell 137 : 522–35.

[107] Malone CD, Hannon GJ (2009) Small RNAs as guardians of the genome. Cell 136 : 656–68.

[108] Zanni V, Eymery A, Coiffet M, Zytnicki M, Luyten I, et al. (2013) Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. Proc Natl Acad Sci U S A 110 : 19842–7.

[109] Karginov FV, Hannon GJ (2010) The CRISPR System : Small RNA-Guided Defense in Bacteria and Archaea. Molecular Cell 37 : 7–19.

[110] Kumar MS, Chen KC (2012) Evolution of animal Piwi-interacting RNAs and prokaryotic CRISPRs. Briefings in functional genomics 11 : 277–88.

[111] Yamanaka S, Siomi MC, Siomi H (2014) piRNA clusters and open chromatin structure. Mobile DNA 5 : 22.

[112] de Vanssay A, Bougé AL, Boivin A, Hermant C, Teysset L, et al. (2013) Profiles of piRNA abundances at emerging or established piRNA loci are determined by local DNA sequences. RNA biology 10 : 1233–9.

[113] Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, et al. (2006) A novel class of small RNAs bind to MILI protein in mouse testes. Nature 442 : 203–7.

[114] Girard A, Sachidanandam R, Hannon GJ, Carmell MA (2006) A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature 442 : 199–202.

[115] Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, et al. (2006) A distinct small RNA pathway silences selfish genetic elements in the germline. Science 313 : 320–4.

[116] Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, et al. (2007) Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 128 : 1089–1103.

[117] Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, et al. (2008) An epigenetic role for maternally inherited piRNAs in transposon silencing. Science 322 : 1387–1392.

[118] Dennis C, Zanni V, Brasset E, Eymery A, Zhang L, et al. (2013) "Dot COM", a nuclear transit center for the primary piRNA pathway in Drosophila. PloS one 8 : e72752.

[119] Olivieri D, Senti KA, Subramanian S, Sachidanandam R, Brennecke J (2012) The Cochaperone Shutdown Defines a Group of Biogenesis Factors Essential for All piRNA Populations in Drosophila. Molecular Cell 47 : 954–69.

[120] Sienski G, Dönertas D, Brennecke J (2012) Transcriptional Silencing of Transposons by Piwi and Maelstrom and Its Impact on Chromatin State and Gene Expression. Cell 151 : 964–980.

[121] Huang XA, Yin H, Sweeney S, Raha D, Snyder M, et al. (2013) A major epigenetic programming mechanism guided by piRNAs. Developmental cell 24 : 502–16.

[122] Le Thomas A, Rogers AK, Webster A, Marinov GK, Liao SE, et al. (2013) Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. Genes and Development 27 : 390–399.

[123] Rozhkov NV, Hammell M, Hannon GJ (2013) Multiple roles for Piwi in silencing Drosophila transposons. Genes & development 27 : 400–12.

[124] Aravin Aa, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, et al. (2008) A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Molecular cell 31 : 785–99.

[125] Senti KA, Brennecke J (2010) The piRNA pathway : a fly's perspective on the guardian of the genome. Trends in genetics 26 : 499–509.

[126] Pandey RR, Pillai RS (2014) Primary piRNA biogenesis : caught up in a Maelstrom. The EMBO journal 33 : 1979–80.

[127] Akkouche A, Grentzinger T, Fablet M, Armenise C, Burlet N, et al. (2013) Maternally deposited germline piRNAs silence the tirant retrotransposon in somatic cells. EMBO reports 14 : 458–64.

[128] Li C, Vagin VV, Lee S, Xu J, Ma S, et al. (2009) Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. Cell 137 : 509–21.

[129] Khurana JS, Wang J, Xu J, Koppetsch BS, Thomson TC, et al. (2011) Adaptation to P element transposon invasion in Drosophila melanogaster. Cell 147 : 1551–63.

[130] Shpiz S, Ryazansky S, Olovnikov I, Abramov Y, Kalmykova A (2014) Euchromatic transposon insertions trigger production of novel Pi- and endo-siRNAs at the target sites in the drosophila germline. PLoS genetics 10 : e1004138.

[131] Mohn F, Sienski G, Handler D, Brennecke J (2014) The rhino-deadlock-cutoff complex licenses noncanonical transcription of dual-strand piRNA clusters in Drosophila. Cell 157 : 1364–79.

[132] Rebollo R, Romanish MT, Mager DL (2012) Transposable elements : an abundant and natural source of regulatory sequences for host genes. Annual review of genetics 46 : 21–42.

[133] Grandbastien MA (2014) LTR retrotransposons, handy hitchhikers of plant regulation and stress response. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms .

[134] Zeh DW, Zeh JA, Ishida Y (2009) Transposable elements and an epigenetic basis for punctuated equilibria. BioEssays 31 : 715–26.

[135] Le Rouzic A, Boutin TS, Capy P (2007) Long-term evolution of transposable elements. Proc Natl Acad Sci U S A 104 : 19375–80.

[136] Schaack S, Pritham EJ, Wolf A, Lynch M (2010) DNA transposon dynamics in populations of Daphnia pulex with and without sex. Proc Biol Sci 277 : 2381–2387.

[137] Frost LS, Leplae R, Summers AO, Toussaint A (2005) Mobile genetic elements : the agents of open source evolution. Nature reviews Microbiology 3 : 722–732.

[138] Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. Molecular biology and evolution 19 : 2226–38.

[139] Lang AS, Zhaxybayeva O, Beatty JT (2012) Gene transfer agents : phage-like elements of genetic exchange. Nature Reviews Microbiology 10 : 472–82.

[140] Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, de la Cruz F (2010) Mobility of plasmids. Microbiology and Molecular Biology Reviews 74 : 434–52.

[141] Ivancevic AM, Walsh AM, Kortschak RD, Adelson DL (2013) Jumping the fine LINE between species : horizontal transfer of transposable elements in animals catalyses genome evolution. BioEssays 35 : 1071–1082.

[142] O'Brochta DA, Stosic CD, Pilitt K, Subramanian RA, Hice RH, et al. (2009) Transpositionally active episomal hAT elements. BMC molecular biology 10 : 108.

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

[144] Maruyama K, Hartl DL (1991) Evidence for interspecific transfer of the transposable element mariner between drosophila and zaprionus. J Mol Evol 33 : 514–524.

[145] Simmons GM (1992) Horizontal transfer of hobo transposable elements within the drosophila melanogaster species complex : evidence from dna sequencing. Mol Biol Evol 9 : 1050–1060.

[146] Wallau GL, Ortiz MF, Loreto ELS (2012) Horizontal transposon transfer in eukarya : detection, bias, and perspectives. Genome biology and evolution 4 : 689–99.

[147] Pace JK, Gilbert C, Clark MS, Feschotte C (2008) Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. Proc Natl Acad Sci U S A 105 : 17023–8.

[148] Novick P, Smith J, Ray D, Boissinot S (2010) Independent and parallel lateral transfer of DNA transposons in tetrapod genomes. Gene 449 : 85–94.

[149] Gilbert C, Waters P, Feschotte C, Schaack S (2013) Horizontal transfer of OC1 transposons in the Tasmanian devil. BMC genomics 14 : 134.

[150] El Baidouri M, Carpentier MC, Cooke R, Gao D, Lasserre E, et al. (2014) Widespread and frequent horizontal transfers of transposable elements in plants. Genome research 24 : 831–838.

[151] Lynch M (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci U S A 104 : 8597–8604.

[152] Han MV, Hahn MW (2012) Inferring the history of interchromosomal gene transposition in Drosophila using n-dimensional parsimony. Genetics 190 : 813–25.

[153] Dupuy C, Periquet G, Serbielle C, Bézier A, Louis F, et al. (2011) Transfer of a chromosomal Maverick to endogenous bracovirus in a parasitoid wasp. Genetica 139 : 489–96.

[154] Gilbert C, Chateigner A, Ernenwein L, Barbe V, Bézier A, et al. (2014) Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons. Nature communications 5 : 3348.

[155] Stroun M, Lyautey J, Lederrey C, Mulcahy HE, Anker P (2001) Alu repeat sequences are present in increased proportions compared to a unique gene in plasma/serum dna : evidence for a preferential release from viable cells ? Ann N Y Acad Sci 945 : 258–264.

[156] Gilbert C, Schaack S, Pace 2nd JK, Brindley PJ, Feschotte C, et al. (2010) A role for host-parasite interactions in the horizontal transfer of transposons across phyla. Nature 464 : 1347–50.

[157] Ghosh A, Meirmans PG, Haccou P (2012) Quantifying introgression risk with realistic population genetics. Proceedings of the Royal Society B : Biological Sciences 279 : 4747–54.

[158] Wallau GL, Kaminski VL, Loreto EL (2011) The role of vertical and horizontal transfer in the evolution of Paris-like elements in drosophilid species. Genetica 139 : 1487–1497.

[159] Belyayev A (2014) Bursts of transposable elements as an evolutionary driving force. Journal of Evolutionary Biology .

[160] Naito K, Cho E, Yang G, Campbell MA, Yano K, et al. (2006) Dramatic amplification of a rice transposable element during recent domestication. Proc Natl Acad Sci U S A 103 : 17620–5.

[161] García Guerreiro MP (2012) What makes transposable elements move in the Drosophila genome ? Heredity 108 : 461–8.

[162] Lerat E, Rizzon C, Biémont C (2003) Sequence divergence within transposable element families in the Drosophila melanogaster genome. Genome Res 13 : 1889–1896.

[163] Kofler R, Betancourt AJ, Schlötterer C (2012) Sequencing of pooled dna samples (pool-seq) uncovers complex dynamics of transposable element insertions in drosophila melanogaster. PLoS Genet 8 : e1002487.

[164] Oliver KR, McComb JA, Greene WK (2013) Transposable elements : powerful contributors to angiosperm evolution and diversity. Genome Biol Evol 5 : 1886–1901.

[165] Fontdevila A (2005) Hybrid genome evolution by transposition. Cytogenet Genome Res 110 : 49–55.

[166] Michalak P (2009) Epigenetic, transposon and small rna determinants of hybrid dysfunctions. Heredity (Edinb) 102 : 45–50.

[167] Hashida SN, Uchiyama T, Martin C, Kishima Y, Sano Y, et al. (2006) The temperature-dependent change in methylation of the antirrhinum transposon tam3 is controlled by the activity of its transposase. Plant Cell 18 : 104–118.

[168] Ebina H, Levin HL (2007) Stress management : how cells take control of their transposons. Mol Cell 27 : 180–181.

[169] Cho K, Lee YK, Greenhalgh DG (2008) Endogenous retroviruses in systemic response to stress signals. Shock 30 : 105–116.

[170] Loreto ELS, Carareto CMA, Capy P (2008) Revisiting horizontal transfer of transposable elements in Drosophila. Heredity 100 : 545–554.

[171] Rouzic AL, Capy P (2005) The first steps of transposable elements invasion : parasitic strategy vs. genetic drift. Genetics 169 : 1033–1043.

[172] Rozhkov NV, Schostak NG, Zelentsova ES, Yushenova Ia, Zatsepina OG, et al. (2012) Evolution and dynamics of small RNA response to a retroelement invasion in Drosophila. Molecular biology and evolution : 1–27.

[173] Jurka J, Bao W, Kojima KK, Kohany O, Yurka MG (2012) Distinct groups of repetitive families preserved in mammals correspond to different periods of regulatory innovations in vertebrates. Biology direct 7 : 36.

[174] Lu J, Clark AG (2010) Population dynamics of PIWI-interacting RNAs (piRNAs) and their targets in Drosophila. Genome research 20 : 212–27.

[175] Khan H, Smit A, Boissinot S (2006) Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Research 16 : 78–87.

[176] Jurka J, Bao W, Kojima KK (2011) Families of transposable elements, population structure and the origin of species. Biology direct 6 : 44.

[177] Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74 : 5463–7.

[178] Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, et al. (1998) New goals for the U.S. Human Genome Project : 1998-2003. Science 282 : 682–9.

[179] Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, et al. (2000) The genome sequence of Drosophila melanogaster. Science 287 : 2185–95.

[180] Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420 : 520–62.

[181] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437 : 376–380.

[182] Shendure J, Ji H (2008) Next-generation DNA sequencing. Nature biotechnology 26 : 1135–45.

[183] Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. Hum Mol Genet 19 : R227–R240.

[184] Smit A, Hubley, R, Green P (2010) RepeatMasker Open-3.0. URL `http://www.repeatmasker.org`.

[185] Bergman CM, Quesneville H (2007) Discovering and detecting transposable elements in genome sequences. Briefings in bioinformatics 8 : 382–92.

[186] Saha S, Bridges S, Magbanua ZV, Peterson DG (2008) Empirical comparison of ab initio repeat finding programs. Nucleic Acids Res 36 : 2284–2294.

[187] Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes : how to find your way through the dense forest of programs. Heredity 104 : 520–533.

[188] Janicki M, Rooke R, Yang G (2011) Bioinformatics and genomic analysis of transposable elements in eukaryotic genomes. Chromosome Res 19 : 787–808.

[189] Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95 : 315–327.

[190] Novák P, Neumann P, Macas J (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC bioinformatics 11 : 378.

[191] Fiston-Lavier AS, Carrigan M, Petrov DA, González J, Gonza J (2011) T-lex : a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. Nucleic Acids Res 39 : e36.

[192] Wheeler TJ, Eddy SR (2013) nhmmer : DNA homology search with profile HMMs. Bioinformatics 29 : 2487–9.

[193] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature methods 9 : 357–9.

[194] Schbath S, Martin V, Zytnicki M, Fayolle J, Loux V, et al. (2012) Mapping reads on a genomic sequence : an algorithmic overview and a practical comparative analysis. Journal of Computational Biology 19 : 796–813.

[195] Wacholder AC, Cox C, Meyer TJ, Ruggiero RP, Vemulapalli V, et al. (2014) Inference of transposable element ancestry. PLoS genetics 10 : e1004482.

[196] Koch P, Platzer M, Downie BR (2014) RepARK–de novo creation of repeat libraries from whole-genome NGS reads. Nucleic acids research 42 : e80.

[197] Zytnicki M, Akhunov E, Quesneville H (2014) Tedna : a transposable element de novo assembler. Bioinformatics 30 : 2656–8.

[198] Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nature Reviews Genetics 12 : 671–682.

[199] Lerat E, Brunet F, Bazin C, Capy P (1999) Is the evolution of transposable elements modular ? Genetica 107 : 15–25.

[200] Engels WR, Johnson-Schlitz DM, Eggleston WB, Sved J (1990) High-frequency p element loss in drosophila is homolog dependent. Cell 62 : 515–525.

[201] Saito K, Nishida KM, Mori T, Kawamura Y, Miyoshi K, et al. (2006) Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. Genes & development 20 : 2214–22.

[202] Aravin AA, Hannon GJ, Brennecke J (2007) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science 318 : 761–4.

[203] Lerat E, Burlet N, Biémont C, Vieira C (2011) Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. Gene 473 : 100–109.

[204] Tamura K, Subramanian S, Kumar S (2004) Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. Mol Biol Evol 21 : 36–44.

[205] Hu TT, Eisen MB, Thornton KR, Andolfatto P (2013) A second-generation assembly of the Drosophila simulans genome provides new insights into patterns of lineage-specific divergence. Genome research 23 : 89–98.

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

[207] de la Chaux N, Wagner A (2009) Evolutionary dynamics of the LTR retrotransposons roo and rooA inferred from twelve complete Drosophila genomes. BMC evolutionary biology 9 : 205.

[208] Carareto CM (2011) Tropical Africa as a cradle for horizontal transfers of transposable elements between species of the genera Drosophila and Zaprionus. Mobile genetic elements 1 : 179–186.

[209] Barbash DA (2010) Ninety years of Drosophila melanogaster hybrids. Genetics 186 : 1–8.

[210] Shi SY, Cai XH, Ding Df (2005) Identification and categorization of horizontally transferred genes in prokaryotic genomes. Acta biochimica et biophysica Sinica 37 : 561–6.

[211] Putonti C, Luo Y, Katili C, Chumakov S, Fox GE, et al. (2006) A computational tool for the genomic identification of regions of unusual compositional properties and its utilization in the detection of horizontally transferred sequences. Molecular biology and evolution 23 : 1863–8.

[212] Podell S, Gaasterland T (2007) DarkHorse : a method for genome-wide prediction of horizontal gene transfer. Genome biology 8 : R16.

[213] Wei X, Cowen L, Brodley C, Brady A, Sculley D, et al. (2008) A Distance-Based Method for Detecting Horizontal Gene Transfer in Whole Genomes. Proceedings of the 4th international conference on Bioinformatics research and applications 4983 : 26–37.

[214] Jaron KS, Moravec JC, Martínková N (2014) SigHunt : horizontal gene transfer finder optimized for eukaryotic genomes. Bioinformatics 30 : 1081–1086.

[215] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215 : 403–410.

[216] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate : a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 57 : 289–300.

[217] Efron B, Tibshirani R (2002) Empirical Bayes methods and false discovery rates for microarrays. Genetic epidemiology 23 : 70–86.

[218] Sun W, Cai TT (2007) Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control. Journal of the American Statistical Association 102 : 901–912.

[219] Pounds S, Cheng C (2006) Robust estimation of the false discovery rate. Bioinformatics 22 : 1979–87.

[220] Saito K, Siomi MC (2010) Small RNA-mediated quiescence of transposable elements in animals. Developmental cell 19 : 687–97.

[221] Vieira C, Lepetit D, Dumont S, Biémont C (1999) Wake up of transposable elements following Drosophila simulans worldwide colonization. Molecular biology and evolution 16 : 1251–5.

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

[223] Mousset S, Derome N (2004) Molecular polymorphism in Drosophila melanogaster and D. simulans : what have we learned from recent studies ? Genetica 120 : 79–86.

[224] Girard A, Hannon GJ (2008) Conserved themes in small-rna-mediated transposon control. Trends Cell Biol 18 : 136–148.

[225] Siomi MC, Sato K, Pezic D, Aravin AA (2011) Piwi-interacting small rnas : the vanguard of genome defence. Nat Rev Mol Cell Biol 12 : 246–258.

[226] Akkouche A, Grentzinger T, Fablet M, Armenise C, Burlet N, et al. (2013) Maternally deposited germline piRNAs silence the tirant retrotransposon in somatic cells. EMBO reports 14 : 458–64.

[227] Grentzinger T, Armenise C, Brun C, Mugat B, Serrano V, et al. (2012) piRNA-mediated transgenerational inheritance of an acquired trait. Genome research 22 : 1877–88.

[228] Coline G, Théron E, Brasset E, Vaury C (2014) History of the discovery of a master locus producing piRNAs : the flamenco/COM locus in Drosophila melanogaster. Frontiers in genetics 5 : 257.

[229] Bailly-Bechet M, Haudry A, Lerat E (2014) “One code to find them all” : a perl tool to conveniently parse RepeatMasker output files. Mobile DNA 5 : 13.

[230] Modolo L, Picard F, Lerat E (2014) A new genome-wide method to track horizontally transferred sequences : application to Drosophila. Genome biology and evolution 6 : 416–32.

[231] Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10 : R25.

[232] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature methods 9 : 357–9.

[233] Quinlan AR, Hall IM (2010) BEDTools : a flexible suite of utilities for comparing genomic features. Bioinformatics 26 : 841–2.

[234] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25 : 2078–9.

[235] Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12 : 656–664.

[236] Li W, Godzik A (2006) Cd-hit : a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22 : 1658–9.

[237] Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome biology 11 : R106.

[238] Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26 : 139–40.

[239] Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, et al. (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in bioinformatics 14 : 671–83.

[240] Hardcastle TJ, Kelly KA, Baulcombe DC (2012) Identifying small interfering RNA loci from high-throughput sequencing data. Bioinformatics 28 : 457–63.

[241] Jung I, Park JC, Kim S (2014) piClust : A density based piRNA clustering algorithm. Computational biology and chemistry .

[242] Deloger M, Cavalli FMG, Lerat E, Biémont C, Sagot MF, et al. (2009) Identification of expressed transposable element insertions in the sequenced genome of drosophila melanogaster. Gene 439 : 55–62.

[243] Akkouche A, Rebollo R, Burlet N, Esnault C, Martinez S, et al. (2012) tirant, a newly discovered active endogenous retrovirus in Drosophila simulans. Journal of virology 86 : 3675–81.

[244] Biémont C, Vieira C, Borie N, Lepetit D (1999) Transposable elements and genome evolution : the case of Drosophila simulans. Genetica 107 : 113–20.

[245] Metzker ML (2010) Sequencing technologies - the next generation. Nature reviews Genetics 11 : 31–46.

[246] Minoche AE, Dohm JC, Himmelbauer H (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome biology 12 : R112.

[247] Kelley DR, Schatz MC, Salzberg SL (2010) Quake : quality-aware detection and correction of sequencing errors. Genome Biol 11 : R116.

[248] Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM (2013) An extensive evaluation of read trimming effects on Illumina NGS data analysis. PloS one 8 : e85024.

[249] Lukic S, Chen K (2011) Human piRNAs are under selection in Africans and repress transposable elements. Mol Biol Evol 28 : 3061–3067.

[250] Song J, Liu J, Schankenberg S, Ha H, Xing J, et al. (2014) Variation in piRNA and transposable element content in strains of Drosophila melanogaster. Genome Biology and Evolution : evu217.

[251] Xu HE, Zhang HH, Xia T, Han MJ, Shen YH, et al. (2013) BmTEdb : a collective database of transposable elements in the silkworm genome. Database 2013 : bat055.

[252] Nakagome M, Solovieva E, Takahashi A, Yasue H, Hirochika H, et al. (2014) Transposon Insertion Finder (TIF) : a novel program for detection of de novo transpositions of transposable elements. BMC bioinformatics 15 : 71.

[253] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15 : 1034–1050.

[254] Loureiro T, Camacho R, Vieira J, Fonseca NA (2013) Boosting the Detection of Transposable Elements Using Machine Learning. In : 7th International Conference on Practical Applications of Computational Biology & Bioinformatics. Springer, pp. 85–91.

[255] Dias ES, Carareto CMA (2012) Ancestral polymorphism and recent invasion of transposable elements in Drosophila species. BMC evolutionary biology 12 : 119.

[256] Le Rouzic A, Payen T, Hua-Van A (2013) Reconstructing the evolutionary history of transposable elements. Genome biology and evolution 5 : 77–86.

[257] Flutre T, Permal E, Quesneville H (2011) In search of lost trajectories : Recovering the diversification of transposable elements. Mob Genet Elements 1 : 151–154.

[258] Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. Journal of molecular evolution 29 : 170–9.

[259] Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Molecular biology and evolution 16 : 1114–1116.

[260] Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. Systematic biology 51 : 492–508.

ANNEXES :

GBE

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*)

Clément Goubert[1,2,3], Laurent Modolo[1,2,3], Cristina Vieira[1,2,3], Claire Valiente Moro[2,3,4], Patrick Mavingui[2,3,4,5], and Matthieu Boulesteix[1,2,3,*]

[1]Laboratoire de Biométrie et Biologie Évolutive, UMR 5558, CNRS, INRIA, VetAgro Sup, Villeurbanne, France

[2]Université de Lyon 1, Villeurbanne, France

[3]Université de Lyon, Lyon, France

[4]Ecologie Microbienne, UMR 5557, CNRS, USC INRA 1364, VetAgro Sup, FR41 BioEnvironment and Health, Villeurbanne, France

[5]Université de La Réunion, UMR PIMIT, CNRS 9192, INSERM 1187, IRD 249

*Corresponding author: E-mail: matthieu.boulesteix@univ-lyon1.fr.

Associate editor: Josefa Gonzalez

Accepted: March 6, 2015

Abstract

Repetitive DNA, including transposable elements (TEs), is found throughout eukaryotic genomes. Annotating and assembling the "repeatome" during genome-wide analysis often poses a challenge. To address this problem, we present dnaPipeTE—a new bioinformatics pipeline that uses a sample of raw genomic reads. It produces precise estimates of repeated DNA content and TE consensus sequences, as well as the relative ages of TE families. We shows that dnaPipeTE performs well using very low coverage sequencing in different genomes, losing accuracy only with old TE families. We applied this pipeline to the genome of the Asian tiger mosquito *Aedes albopictus*, an invasive species of human health interest, for which the genome size is estimated to be over 1 Gbp. Using dnaPipeTE, we showed that this species harbors a large (50% of the genome) and potentially active repeatome with an overall TE class and order composition similar to that of *Aedes aegypti*, the yellow fever mosquito. However, intraorder dynamics show clear distinctions between the two species, with differences at the TE family level. Our pipeline's ability to manage the repeatome annotation problem will make it helpful for new or ongoing assembly projects, and our results will benefit future genomic studies of *A. albopictus*.

Key words: transposable elements, repeated DNA, TE analysis, *Aedes albopictus*, Trinity, bioinformatic pipeline.

Introduction

Repeated DNA, including transposable elements (TEs), is widespread within eukaryotic genomes. In such a "repeatome," the spread of TEs, which might bear coding sequences and can reach thousands of base pairs in length, contributes substantially to genomic size and evolution. Because of their ability to insert within genes or regulatory regions and to cause ectopic recombination due to their repetitive nature, TEs are assumed to be frequently deleterious to their hosts (Goodier and Kazazian 2008; Beck et al. 2011; Vela et al. 2014). However, an increasing number of studies have shown that TE insertions can sometimes be adaptive and can be co-opted by their host genomes (Rebollo et al. 2010; Casacuberta and González 2013). Thus, understanding genomic evolution demands a comprehensive knowledge of TE composition within the genome, as well as of their dynamics and interactions with host genome. To this end, genome annotations that include TE annotation and quantification are crucial.

In the current era of short-read sequencing, the assembly of genomes bearing a significant amount of repeated sequence is a complex task. Reads overlapping a repeated element might correspond to several positions in the genome

and thus can be misplaced and can produce chimeric assembly. Therefore, repeats produce a large number of short contigs that cannot be properly positioned or annotated within the assembly. Accordingly, the quality of the assembly for TEs is often poor and can result in underrepresented and/or incorrect annotation of their sequences (Modolo and Lerat 2014).

The Asian tiger mosquito *Aedes albopictus* (Diptera: Culicidae) presents a striking example of a genome that is difficult to assemble due to its repeatome. This species—a vector of Dengue and Chikungunya viruses that is often viewed as one of the most threatening invasive species in the world—still has not had its genome sequence released, even though several projects have been aimed at this task over the last few years (see Bonizzoni et al. 2013 for a review). *Aedes aegypti*, the closest species whose genome has been fully sequenced and annotated, possesses a similar genome size, and repeated DNA comprises more than 50% of its genome. Unlike *A. albopictus*, the whole genome of *A. aegypti* has been fully sequenced using Sanger technology, which produces longer reads than current Next-Generation Sequencing (NGS) methods and therefore allowed the construction of a large library of TEs and repeats (Nene et al. 2007). Moreover, intraspecies variation of the *A. albopictus* genome size—ranging from 0.62 to 1.66 pg—has been suggested (Rao and Rai 1987; Kumar and Rai 1990), supporting the hypothesis of a significant amount of TE activity, with more copies present in some populations than in others (McLain et al. 1987; Black et al. 1988). However, no study is currently aimed at finding and quantifying TEs in a comprehensive manner in this species.

Several bioinformatic solutions now enable the de novo assembly of TE sequences directly from NGS genomic data sets without the need for a reference genome. These methods assume that reads belonging to TEs or other repetitive DNAs are overrepresented among the sequenced reads. Current pipelines such as RepARK (Koch et al. 2014) and TEdna (Zytnicki et al. 2014) use whole NGS genomic data sets or only the unassembled reads left after a genome assembly. These two programs use overrepresented k-mers to assemble TE sequences: Velvet (Zerbino and Birney 2008) or CLC (CLCbio, http://www.clcbio.com/products/clc-assembly-cell/, last accessed April 13, 2015) are used in RepARK, and an implementation of a de Bruijn graph assembler is used in TEdna. Although these programs are dedicated to TE assembly, they do not allow repeat quantification or annotation. An alternative way to explore a genome's repetitive content is to use low coverage sequencing. In such data sets, only TEs and other repetitive DNA sequences are expected to have a sufficient representation in the pool of reads to be assembled. For example, in average, for a sample with 0.1× coverage, only sequences that are present at least 10 times within the genome can be assembled. Based on this principle, the RepeatExplorer (RE) pipeline (Novák et al. 2010) was designed to cluster and then assemble similar reads from a small uniform genomic sample in order to retrieve repeats. In a uniform genomic sample, the proportion of reads assigned to a given cluster directly corresponds to the proportion of reads assigned to the relevant TE family in the genome. In addition to computing a direct quantification of each repeat family, RE can annotate repeat families using RepeatMasker (RM) and protein domain search (Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. 1996–2010, http://www.repeatmasker.org, last accessed April 13, 2015). However, although the RE pipeline can process NGS data sets, most of the tools it uses are not designed for this type of data, especially during the assembly step performed by CAP3 (Huang 1999)—a Bacterial Artificial Chromosome (BAC)-clone sequence type assembler.

Here, we present a new pipeline, dnaPipeTE (De Novo Assembly and Annotation Pipeline for Transposable Elements), that combines previous methods by allowing fast and accurate assembly of repeat sequences from a small genomic sample with dedicated NGS tools and by performing quantification and annotation of TEs and repeats for comparative analysis. The cornerstone of dnaPipeTE is the use of Trinity (Grabherr et al. 2011)—originally designed for RNAseq data assembly—to assemble repeats from low-coverage genomic data sets, which produce complete repeat sequences and enable the recovery of alternative consensuses within one TE family. Our pipeline also performs an automatic annotation of repeats using RM and the Repbase database (Jurka et al. 2005) and produces different data and figures for the quantification of repeats. We also implemented a computation of the TE age distribution for the most recent copies, using the divergence between reads and contigs.

With this pipeline and annotations from known TEs, we aimed to 1) estimate the number of repeated DNAs in *A. albopictus*, 2) annotate and quantify the diversity of TEs in its genome, and 3) compare this repeatome with that of *A. aegypti*, to infer the dynamics of TEs since the divergence of these two species.

Materials and Methods

dnaPipeTE: A Pipeline to Assemble, Annotate, and Quantify Repetitive Sequences from Small Unassembled NGS Data Sets

dnaPipeTE is a fully automated pipeline designed to assemble and quantify repeats from genomic NGS reads. It is freely available for download at https://lbbe.univ-lyon1.fr/-dnaPipeTE-.html (under the GPLv3). Figure 1 shows the main steps in the dnaPipeTE pipeline. Our pipeline takes as input a FASTQ (Cock et al. 2010) file containing quality filtered short reads. dnaPipeTE then performs uniform samplings of the reads to produce low coverage data sets used during analysis. The samples must represent less than 1× coverage to avoid the assembly of nonrepeated genome content; using

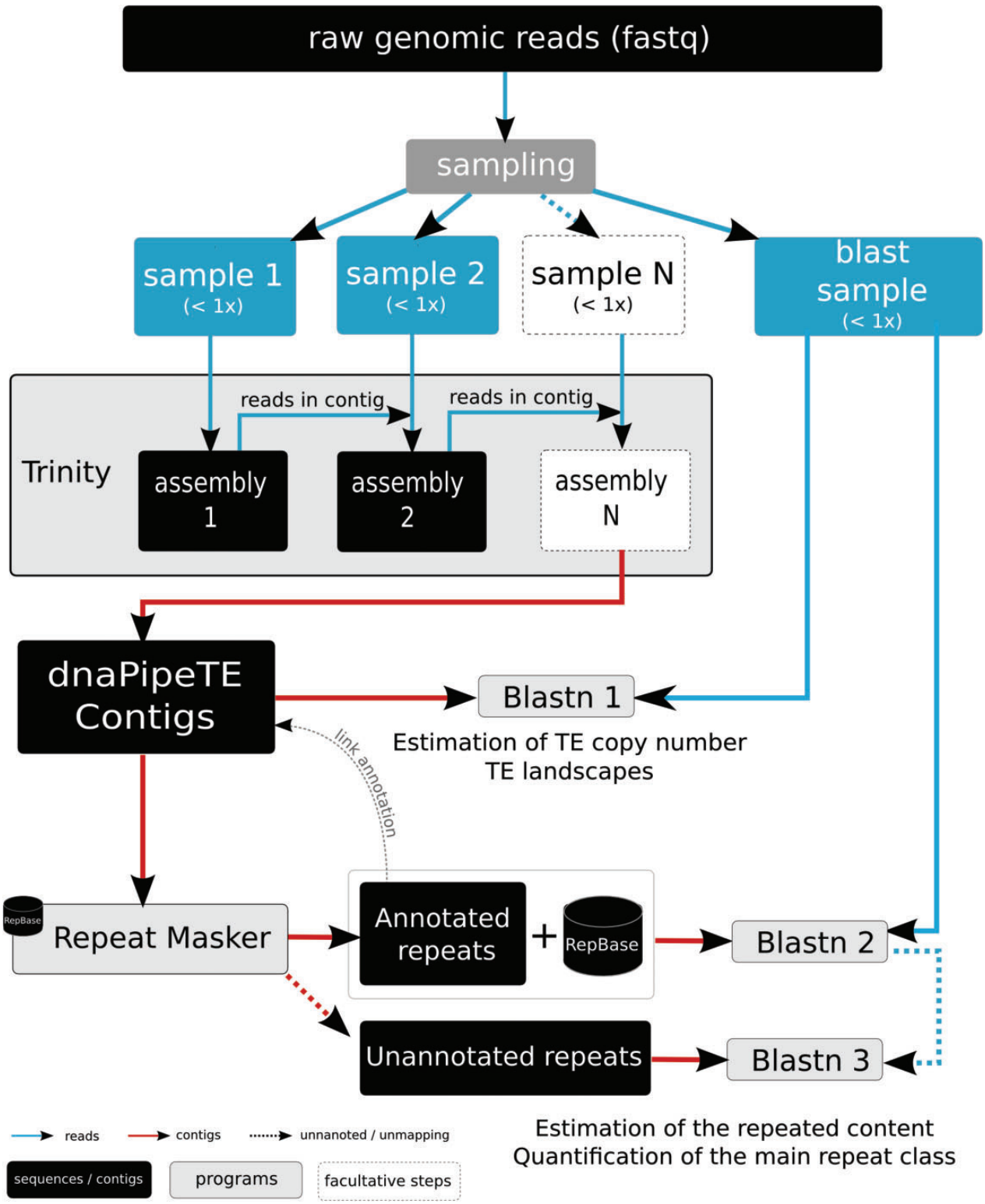


Fig. 1.—Overview of the dnaPipeTE pipeline. First, genomic reads in FASTQ format are sampled. Then, assembly of repeats is performed using two or more iterations of Trinity. For each iteration, the previously assembled reads are added to the next sample to improve the repeat assembly. In the next step, assembled contigs are annotated using RepeatMasker. Finally, reads from the "BLAST sample" are blasted against all the contigs to estimate the relative abundance of each assembled repeat and to compute the TE landscape. In a second BLAST, the same sample is successively blasted against the annotated contigs joined to the Repbase library, then with the unannotated contigs in order to retrieve copies that would not have been assembled and to obtain a more global repeat content estimation. See text for additional details.

a sample size of less than 0.25× of the genome is often sufficient to obtain a precise estimate of the repeated content (see supplementary fig. S1, Supplementary Material online, for examples with 0.1× and 0.25×). dnaPipeTE requires at least three samples of the original genomic data set: Two for the assembly step and an independent third used for the quantification steps. Our pipeline is currently designed to use only single-end reads because training analyses showed that using paired-end reads could produce chimeras during repeat assembly (data not shown). We developed dnaPipeTE using 100-bp reads, which are currently the most frequently generated NGS data sets, but our implementation would work with any read size.

Repeat Assembly with Trinity

After uniform sampling of the reads, dnaPipeTE builds contigs from the repeated sequences using Trinity. In an RNAseq experiment, a given gene can produce different transcripts, and the Trinity software is equipped to handle alternative transcripts with a hierarchical procedure: after identifying a "gene" (a subpart of the assembly graph), Trinity can produce different contigs that represent all the alternative transcripts of this gene. Similarly, TE copies from the same family, which may display an accumulation of mutations, deletions, insertions, or other structural changes, are treated by Trinity as alternative sequences of the same gene (TE family). Thus, with Trinity one can recover complete alternative consensus sequences from a given TE family. Retrieving good consensus increases the ability to perform an accurate estimation of TE abundance by improving read mapping to TEs. The rarest elements in the genome are predicted to generate few (or no) reads in the subset samples; thus, dnaPipeTE performs iterative runs of Trinity using new samples to decrease such risk. The first run uses a first sample; then, any reads mapping to k-mer contigs belonging to repeats ("inchworm" contigs; see Trinity manual) are added to a second independent sample, and Trinity is performed one more time. Each iteration enriches the number of reads associated with a repeat in the next sample and allows the recovery of more and larger contigs (some examples are given in supplementary Material, Supplementary Material online). In the case of *A. albopictus* sequences, our tuning experiments showed that two iterations performed on a data set with 0.1× coverage ensured the best assembly N50 and that supplementary iteration showed no significant improvement in the quality of the assembly (supplementary fig. S2, Supplementary Material online). In the latest versions of Trinity (≥r20140717), contigs are built from "clusters" that correspond to units of the de Bruijn graph made during the assembly. These clusters are divided into genes and finally "isoforms" that represent the alternative transcripts of a gene in RNAseq studies. Applied to low-coverage DNA data, one gene ideally represents one repeat family, in which isoforms are structural variant copies belonging to one family (copies with insertions or deletions for example) or to closely related families. An isoform present in Trinity.fasta output following all iterations of the Trinity program is referred to as a "dnaPipeTE contig." During the assembly step in dnaPipeTE, Trinity (version r20140717) was used with default parameters for single-end reads, with the exception of the minimum coverage to join k-mer contigs set to 1 to retain contigs from low copy repeats (Haas B, personal communication).

Contig Annotation with RepeatMasker

After the assembly step, dnaPipeTE contigs are annotated using RM, for which a built-in or custom repeat library can be specified. Following the 80-80-80 rule proposed by Wicker et al. (2007), contigs with 80% query coverage on 80% of subjects (databases) were stored as "full-length," and queries with 80% hits on fewer than 80% of subjects were stored as "partial" (fig. 2). Of the other contigs annotated by RM, only the order information (according to Wicker et al. 2007 classification)— Long Terminal Repeat (LTR), Long INterspersed Element (LINE), Short INterspersed Element (SINE), DNA, Miniature Inverted-repeat Transposable Elements (MITEs) (short TEs harboring terminal inverted repeats but without coding sequences), Ribosomal RNA, low complexity, and simple/tandem repeats—is retained. For our analysis, we used the Repbase libraries (version 2014-01-31 downloaded from http://www.girinst.org/, last accessed April 13, 2015) and the TEFam library (accessed at http://tefam.biochem.vt.edu/tefam/index.php, last accessed April 13, 2015). RM (version open-4.0.5) parameters were set to default values, slow-research mode with the NCBI BLAST program (RMBLASTN program, NCBI BLAST 2,2,23+), and only the best hit was kept following dnaPipeTE contig analysis, as determined by the highest Smith–Waterman score provided by RM.

Repeat Quantification

For quantifying the repeats, BLASTN software (Altschul et al. 1990) was found to perform better than classic short-read aligners such as Bowtie2 (Langmead and Salzberg 2012). Indeed, the divergence between a dnaPipeTE contig—that is, a consensus sequence for a repeat family—and its reads belonging to different copies can be higher than the divergence between a gene or a transcript and its reads, and requires a more sensitive approach. During the "BLAST 1" step (fig. 1), reads from the "BLAST" sample are matched against all the dnaPipeTE contigs to estimate the genome proportion of each assembled repeat. However, we cannot quantify the unassembled repeats during this step. Thus, to obtain an overall estimation of repeat content, the BLAST sample is first matched against a database composed of the annotated contigs of dnaPipeTE and the repeat library in order to recover reads associated with misassembled or missing repeats

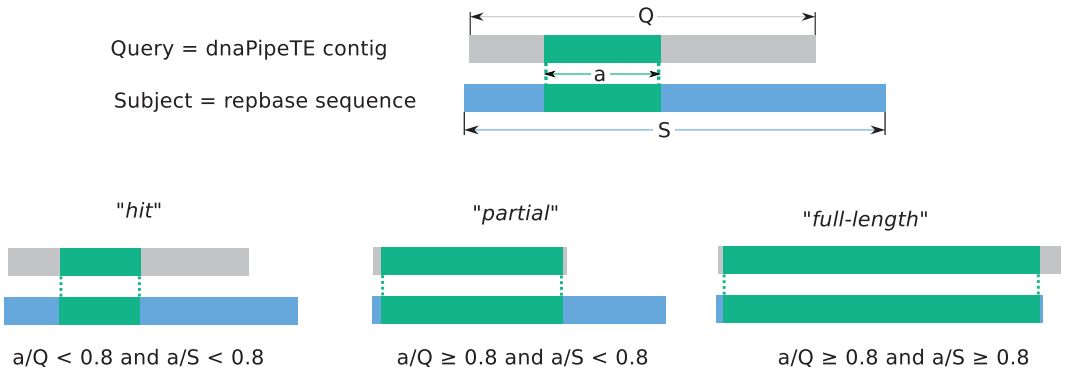


Fig. 2.—Classification procedure of RepeatMasker annotation for the dnaPipeTE contigs. According to the alignment overlap between the query (a/Q) and the subject (a/S), the dnaPipeTE contigs are annotated as one of the three categories. "Hit" is the weakest annotation, while partial and full-length indicate that the dnaPipeTE contig has annotated along more than 80% of its length.

("BLASTN 2," fig. 1). Then, the unmapped reads are matched against the unannotated contigs supplied by dnapipeTE ("BLASTN 3," fig. 1), and the remaining reads are assumed to belong to nonrepeated sequences. We use the BLAST sample for both estimations, and reads are mapped using discontinuous BLASTN (NCBI BLAST 2.2.29+), which keeps matches with 80% minimum identity and only the best hit per read. To speed-up computation, dnaPipeTE uses GNU Parallel (version 20 140 622) (Tange 2011) to parallelize BLASTN runs.

Finally, the divergence computed between one read and its contigs during the BLAST 1 step is used as a proxy of the divergence time between TE copies in a given family. This proxy is shown to be relevant compared with previous analyses of TE age distribution that used Kimura distances from a full-length TE copy and its consensus sequence in Repbase ("TE Landscapes," http://www.repeatmasker.org/ (last accessed April 13, 2015); several examples are given in supplementary fig. S3, Supplementary Material online).

Efficiency of dnaPipeTE

Prior to *A. albopictus* genome analysis, we tested the efficiency of dnaPipeTE on well-annotated genomes that varied in size and TE content. We used available Illumina reads from the species *Drosophila melanogaster* (Diptera: Drosophilidae), *Anopheles gambiae* (Diptera: Culicidae), *Caenorhabditis elegans* (Rhabditida: Rhabditidae), *Ciona intestinalis* (Enterogona: Cionidae), *Gasterosteus aculeatus* (Gasterosteiformes: Gasterosteidae), and *A. aegypti*—the closest fully sequenced species to *A. albopictus*. We also tested the behavior of dnaPipeTE on older repeatomes, such as that of the human genome (*Homo sapiens*), in which copies of one TE family are highly divergent. All data management information and references are given in supplementary table S1, Supplementary Material online.

Analysis of the *A. albopictus* Repeatome and Comparison with *A. aegypti*

Genomic Data

The two mosquito genomes were sequenced with Illumina NGS technology (Illumina HiSeq2000). The *A. albopictus* strain originated from La Reunion Island, Indian Ocean. Genomic DNA was prepared from four female individuals of generation F5 bred in an insectarium. Sequencing generated 440.2 million 100-bp paired-end reads (ProfilXpert platform, Lyon, France). A total sample of 4,243,902 single-end reads was also generated (R1's were used). *Aedes aegypti* female genomic reads (SRR871496; strain Liverpool; 213.4 million 100-bp paired-end reads; ~16.4× coverage, Virginia Tech) were downloaded from the short-read archive collection (http://www.ncbi.nlm.nih.gov/sra, last accessed April 13, 2015); only the first read of each pair was used for analysis.

Read Preprocessing

According to quality statistics, all reads were trimmed to 82 bp, keeping the nucleotides 10 through 91 in both *A. albopictus* and *A. aegypti* species. Then, sequences were filtered using FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/, last accessed April 13, 2015) with a minimum 20 average Phred score on 90% of the reads. Finally, reads from mitochondrial DNA were removed from the data with Bowtie 2 software (version 2.1.0) under default parameters to map reads to the whole mitochondrial genome sequence for each *Aedes* species available through the NCBI website (http://www.ncbi.nlm.nih.gov/, last accessed April 13, 2015).

Aedes albopictus and A. aegypti Sampling

In the literature, the genome size of *A. albopictus* is reported to be variable, ranging from 0.6 to 1.6 Gbp. Flow cytometry performed on the heads of *A. albopictus* females estimated the genome size of our sequenced strain to be 1.16 Gbp

(1.19 pg, unpublished data). The number of reads comprising the three independent samples used by dnaPipeTE was set to represent 0.1× of each genome. The subset sample of 4,243,902 reads (0.3×) was used to assemble TEs and repeats for *A. albopictus*, consisting of 2 samples of 0.1× genomic coverage for assembly and a third sample of 0.1× for the quantification step. This sample size was chosen after a preliminary analysis showed that 0.1× per Trinity run maximizes the assembly N50 for this genome (supplementary fig. S2, Supplementary Material online). We suggest that this will balance finding as many repeats as possible with limiting the assembly of nonrepeated DNA (noise). For *A. aegypti*, coverage was also set to 0.1×, using reads taken from the full sequencing experiment based on a genome size of 1.3 Gbp, according to the whole-genome assembly size and mean genome size estimations (Nene et al. 2007; Gregory, T.R. (2015); Animal Genome Size Database. http://www.genomesize.com, last accessed April 13, 2015).

TE Family Recovery and Quantification

To cluster dnaPipeTE contigs into TE families, we used the cd-hit-est program from the CD-HIT suite (version 4.6.1) (Li and Godzik 2006) with local alignment and the greedy algorithm. We set the clustering parameters to group pairs of sequences with at least 80% of the shortest sequence aligned, with a minimum of 80% identity in the longest sequence (parameters -aS 0,8 -c 0,8 -G 0 -g 1). This method results in better performance than grouping contigs per Trinity gene or by RM annotation. In the first case, contigs from one Trinity gene could be joined when they shared a conserved fragment (such as a protein domain), even if they did not actually belong to the same TE family. In the second case, RM annotations include only the closest sequences known, and one sequence could easily match to multiple TE families. This method allowed us to report the most abundant repeats (in relative genome proportion) and to estimate the number of TE copies for fully assembled repeats (dnaPipeTE contigs full-length, see above).

We then estimated the copy number of the fully assembled repeats (table 1) using the following formula:

$$(n/N) \times (G/L)$$

where n is the number of read-matching contigs from a TE family (contigs from one CD-HIT cluster), N is the total number of reads in the BLAST sample, G is the genome size in bp, and L is the length of the representative sequence of the TE family (reference sequence of the CD-HIT cluster) in bp.

TE Transcriptional Activity

To identify transcriptionally active TEs among the discovered repeats in *A. albopictus*, we mapped the *A. albopictus* transcriptome assembly (adult, embryo, and oocyte transcriptome merged reference assembly downloaded from http://www.albopictusexpression.org/, last accessed April 13, 2015) onto the dnaPipeTE contigs using BLAT. We filtered the results of the BLAT analysis such that only TE consensus sequences matching 80% of a transcriptome contig (minimum alignment 80 bp) with 80% minimum identity were retained.

Comparison between A. albopictus and A. aegypti

To avoid annotation bias due to the abundance of reference sequences from *A. aegypti* in Repbase, we performed a second analysis with dnaPipeTE on *A. albopictus* and *A. aegypti* using a TE library devoid of reference sequences from *A. aegypti*. Then, we used BLAT to match cd-hit-clustered dnaPipeTE contigs between species in order to identify shared TE families. We filtered the results of the BLAT analysis such that alignments with at least 80 bp and 75% identity and only one reference contig per species were retained. Finally, for each species we summed the total number of reads in the cluster for which the references belonged. Thus, we obtained pairs of counts for putatively shared TE families.

dnaPipeTE Comparison with RepeatExplorer

Compared with dnaPipeTE, RE requires only one sample for assembly and annotation. We thus ran it using the "BLAT" sample generated by dnaPipeTE for the *A. albopictus* data set, on which an estimation of repeated content and a quantification of the main repeat families is performed. Computations were performed online with the "clustering" tool of the RE Galaxy server (http://repeatexplorer.umbr.cas.cz/, last accessed April 13, 2015) with the following parameters: 44 bp (55% of the read length) minimum overlap for clustering, 0.01% cluster threshold for detailed analysis, 40 bp minimal overlap for read assembly and RepeatMasking against the "all" database. Computation time, contig number, N50, proportion of repeats in the sample, and percentage of annotation of the repeated content were calculated for comparison.

Results

Efficiency of dnaPipeTE

We report here the results obtained for *D. melanogaster* (fig. 3). Details and results from other species are presented in supplementary figures S1 and S3, Supplementary Material online. In *D. melanogaster*, as well as the other fully annotated genome tested, dnaPipeTE estimations for the different families of TEs are accurate when only a small subset sample of NGS sequencing reads was used as input (three samples of 0.25× coverage). The relative proportion of each TE order is respected in dnaPipeTE estimations. In *D. melanogaster*, however, the whole repeat content is underestimated (17.78% vs. 28.21%). For this species, our results indicate that dnaPipeTE seems to have underestimated the simple and tandem repeat content of the genome. For *A. aegypti* (supplementary fig. S1,

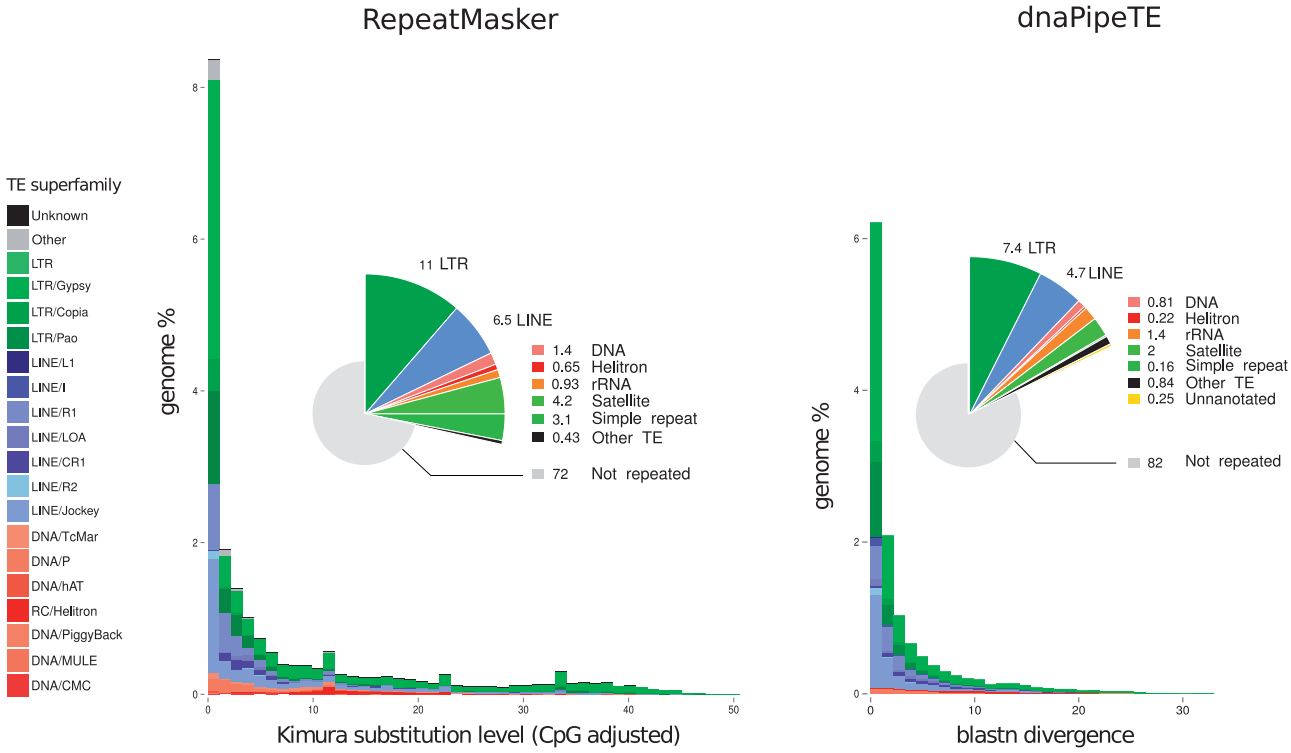


Fig. 3.—Relative genome proportions of the main repeat classes (pie charts) and TE landscapes (bar plots) from RepeatMasker on assembled genome (left) and dnaPipeTE (right, BLASTN with 0.25× genome coverage) for *Drosophila melanogaster* strain *w1118*. RepeatMasker analysis data were downloaded from http://repeatmasker.org and retranscribed according to the name used for annotation in dnaPipeTE.

Supplementary Material online), we estimate the TE content to be 45.6%, which is very close to the estimation of 47% made by Nene et al. from the assembled genome. Using genomes variable in size and TE content as benchmark, we also noticed that the more the genome is filled with repeated DNA, the less the number of Trinity iteration is needed, as well as the coverage provided as input.

Comparisons of TE age distributions obtained with dnaPipeTE (fig. 3 and supplementary fig. S3, Supplementary Material online) and those made from fully assembled genomes available on the RM website (TE landscapes) (http://repeatmasker.org/genomicDatasets/RMGenomicDatasets.html, last accessed April 13, 2015) were performed. These comparisons showed that dnaPipeTE provides a good estimate of the recent TE age distribution. As with other de novo TE assemblers, dnaPipeTE is limited in its ability to detect old TE families with degraded and divergent copies. For example, in *D. melanogaster* or *H. sapiens,* TEs with more than 30% divergence between reads and the consensus sequence are not identified (fig. 3 and supplementary fig. S4, Supplementary Material online). Our tuning tests show that dnaPipeTE performs well in the estimation of TE proportion and dynamics, with consensus-read divergence ranging from 0% to 15%, which is sufficient to compare closely related species and is close to the definition of a TE family as per the 80-80-80 rule (Wicker et al. 2007).

Aedes albopictus Repeatome Analysis

Repeat Assembly with dnaPipeTE

Assembly of the repeats produced 8,102 contigs with an N50 of 677 bp. Although no reference genome for *A. albopictus* exists at this point in time, dnaPipeTE was able to annotate 5,141 contigs including 949 "partial TEs" and 30 full-length elements. Among these, some full-length annotated dnaPipeTE contigs were found to represent different variants of the same family, including some internal deletions. Taking this into account, a total of 24 annotated families with full-length consensus sequences were quoted for *A. albopictus*.

Repeated DNA Content of A. albopictus

dnaPipeTE reported that the repeatome of *A. albopictus* comprises 49.73% of the genome. Annotation of this repeated

Table 1
The Most Abundant Identified Repeat Families in *Aedes albopictus*

Genome%	RM Annotation	RM Superfamily	dnaPipeTE Contig Size	Estimated Copy Number
1.26%	Lian-Aa1	LINE/LOA	4,080	3586
1.25%	RTE Ele4	LINE/RTE-BovB	3,447	4203
1.16%	JAM1	LINE/RTE-BovB	2,356	5728
1.10%	R1_Ele1	LINE/R1	5,797	2195
0.54%	RTE_Ele3	LINE/RTE-BovB	3,283	1911
0.41%	CACTA-3_AA	DNA/CMC-EnSpm	1,626	
0.37%	TF001239_mTA_Ele24Aedes	MITE	638	
0.33%	Chapaev3-2_AA	DNA/CMC-Chapaev-3	1,611	
0.29%	Loner_Ele2	LINE/I	6,335	526
0.28%	TF001239_mTA_Ele24_Aedes	MITE	469	
0.28%	Loner Ele1	LINE/I	6,329	513
0.23%	Lian-Aa1	LINE/LOA	934	
0.23%	FEILAI_AA	S1NE/tRNA	324	8215
0.22%	TF001248_mTA_E1e33_Aedes	MITE	2,407	1071
0.18%	MSAT-1_AAe Satellite		2,133	
0.17%	RTE Ele5	LINE/RTE-BovB	2,642	
0.17%	Lian-Aa1	LINE/LOA	1,865	1053
0.17%	LSU-rRNADme	rRNA	4,681	
0.16%	R1_Ele1	LINE/R1	3,362	
0.16%	JAM1B_AAe	LINE/RTE-BovB	793	
0.16%	LOA_Ele5	LINE/LOA	3,724	500
0.16%	TF001244_mTA_Ele29Aedes	MITE	578	
0.15%	MSAT-2_AAe	Satellite	1,301	
0.15%	TF001312_m8bp_Ele20_Aedes	MITE	1,532	
0.15%	TF000681_m4bp_Ele5_Aedes	MITE	674	2548
0.14%	CR1-50_AAe	LINE/CR1	678	
0.14%	Sola2-4_AAe	DNA/Sola	1,232	
0.14%	TF001310_m8bp_E1e19_Aedes	MITE	1,840	
0.14%	TF001280_otherMITEs_Ele7Aedes	MITE	252	
0.13%	JAM1B_AAe	LINE/RTE-BovB	424	
0.13%	MSAT-1_AAe	Satellite	663	
0.13%	MSAT-2_AAe	Satellite	575	
0.13%	Gecko	SINE/tRNA-I	249	5967
0.13%	TF001295_mTA_Ele38c_Aedes	MITE	1,377	
0.12%	MSAT-1AAe	Satellite	204	
0.12%	TF001257_m4bp_E1e16_Aedes	MITE	887	
0.12%	TF001280_otherMITEs_Ele7Aedes	MITE	1,379	
0.12%	TF001313_otherMITEs_Ele27Aedes	MITE	2,209	
0.12%	MSAT-1_AAe	Satellite	852	
0.12%	TF000746_mTA_Ele22_Aedes	MITE	557	2439
0.11%	LOA_Ele2B_AAe	LINE/LOA	2,484	
0.11%	Sola1-3_AA	DNA/Sola	349	
0.11%	otherMITEs_Ele11	DNA/hAT-hATm	421	
0.11%	TF001251_m3bp_Ele8a_Aedes	MITE	900	

Note.—An estimation of copy number was made only for TEs identified as full-length elements and was based on the size of the dnaPipeTE reference contig after TE family clustering. RM annotation, repeat family hit found by RepeatMasker; RM superfamily, repeat superfamily name in Repbase.

DNA showed that TEs occupy 33.58% of the genome. Tandem repeats (satellites and microsatellites) occupy 8% (fig. 4), while unannotated repeats represent 7.23%. The most abundant repeats were Class II (DNA) transposons and LINE (Class I non-LTR) retrotransposons, followed by LTR retrotransposons and SINEs. Details regarding the most abundant repeat families are reported in table 1. The most abundant TE family in terms of genome percentage is a "Lian-like" LINE element (similar to *Lian-a1* in *A. aegypti*), which occupies 1.267% of the genome with 3,586 estimated copies (table 1). The most highly represented families in terms of copy number among the full-length elements annotated by dnaPipeTE are two LINE elements from the "Loner" superfamily, with more than 6,000 estimated

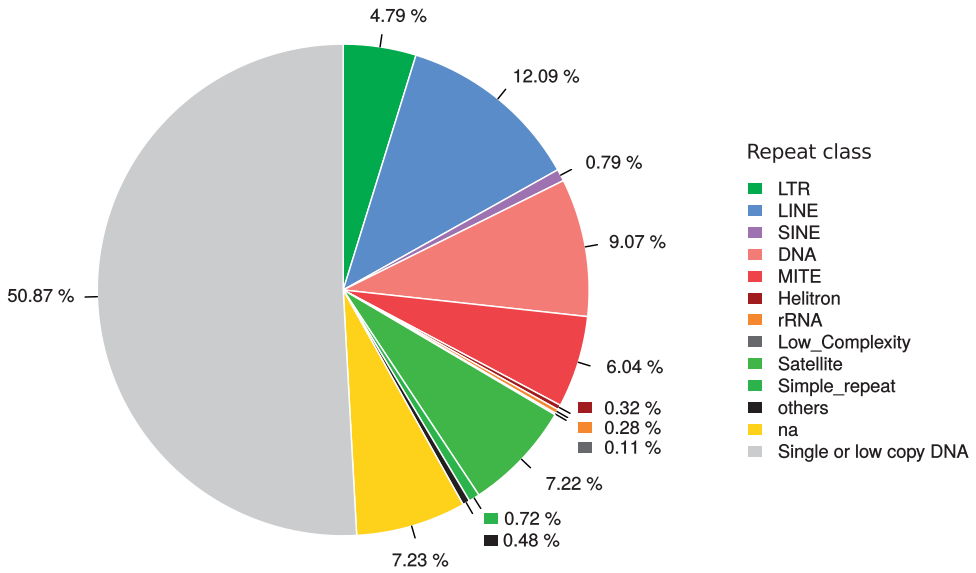


Fig. 4.—Relative genome proportions of the main repeat classes found in *Aedes albopictus* using dnaPipeTE, from a nucleotide BLAST of 1,414,634 reads (0.1×) against the repeat assemblies performed with a total of 2,829,268 reads (0.2×).

copies each. Thirteen other LINE families represent more than 0.10% of the genome each. Fourteen MITEs (non-autonomous Class II) also appear among the most repeated TE families.

In addition, we found using BLAT that 7,005 of the 8,102 dnaPipeTE contigs have significant hits with a sequence from the *A. albopictus* transcriptome assembly reported for adult, embryo, and oocyte (Poelchau 2011; http://www.albopictusexpression.org/ [last accessed April 13, 2015]; supplementary table S2, Supplementary Material online).

Comparison of TE Dynamics between *A. albopictus* and *A. aegypti*

Aedes albopictus TE age distribution was compared with that of the yellow fever mosquito, *A. aegypti* (the only available assembled genome for the *Aedes* genus). We showed that in both species, most of the reads are highly similar to their respective dnaPipeTE contigs (fig. 5). This indicates that most of the detected TE families are recent and possess a high degree of similarity between their copies. This similarity is particularly strong for the detected LTR retrotransposons and, to a lesser extent, for the LINEs that are the most represented TEs in these distributions. Class II DNA transposons are less represented than expected in these comparisons, as their detection suffered from the removal of *A. aegypti* reference sequences from the library for comparison (fig. 4 for the full analysis in *A. albopictus* vs. fig. 5 for the interspecies comparison). Between species, the most striking result is that the genomic proportion of LINE/Jockey reads in *A. aegypti* is high and is composed of mostly recent but also some older TEs, while this family is much less abundant in *A. albopictus*, with less divergence between reads and contigs. In addition, the distribution of the read divergence of LINE/R1 elements is strongly concentrated at the left of the graphic (representing recent TE copies) in *A. aegypti*, while in *A. albopictus* the proportion of reads in superfamilies of higher divergence decreases more slowly (representing older TE copies).

The weak positive correlation between *A. aegypti* and *A. albopictus* in the genomic abundance of the shared families (fig. 6, $r^2 = 0.186$, $P < 0.01$ on the $\log_{10}$ scale) is mostly due to the less abundant families (<0.1% of the genome). Some families display very high differences, such as the *Juan*-A (LINE/Jockey retrotransposon) family which represents almost 3% of the genome proportion in *A. aegypti* but only 0.08% in *A. albopictus*, or *Copia_Ele122* which displays a 5-fold change between the two species, while *R1-Ele*1 and *RTE*-3 are good examples of the mirror case. Globally, very few shared families have the same genomic proportion, with the exception of *CACTA*-3 (DNA transposon) and, less markedly, *Jam*-1 or *Lian-Aa*1 (LINEs), which contrast the general trend.

Comparison between dnaPipeTE and RepeatExplorer

Our pipeline dnaPipeTE operates on the same principles as RE to estimate, assemble, and annotate the repeatome of a species from a sample of reads. Therefore, it was expected that similar estimates of global repeated content in *A. albopictus* would be obtained by RE and dnaPipeTE (table 2). However, dnaPipeTE, in addition to being much faster, was also able to

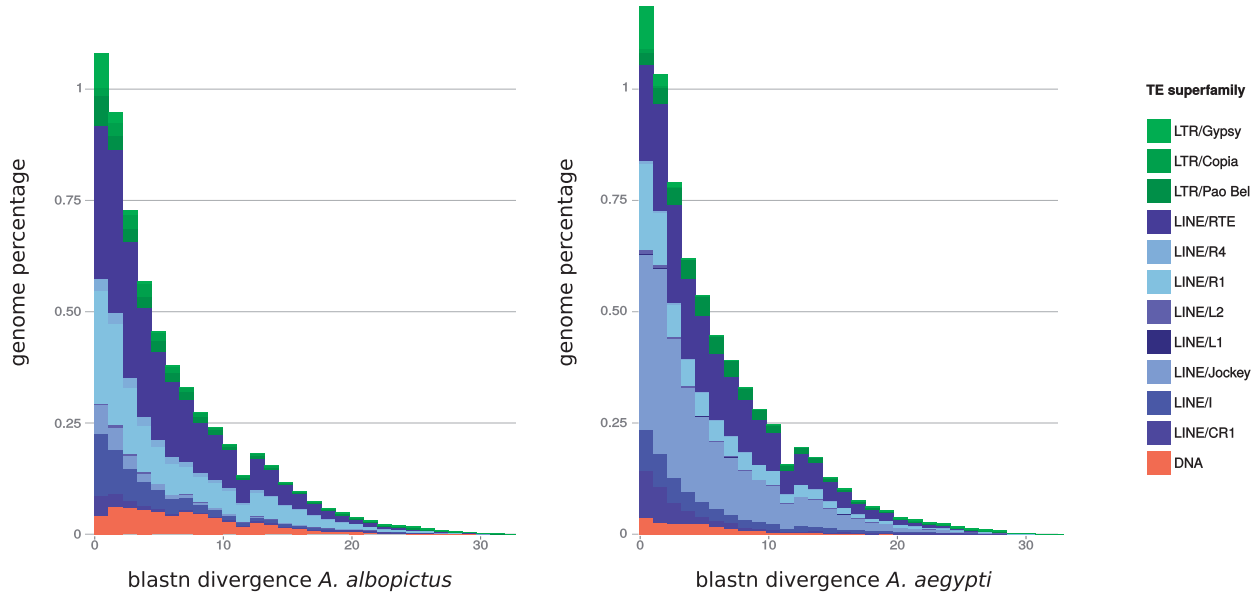


Fig. 5.—TE age distribution comparisons between *Aedes albopictus* (left) and *Aedes aegypti* (right). For each species, the nucleotide divergence from BLASTN is reported between a repeat read and the contig, where it matches the dnaPipeTE assembly.

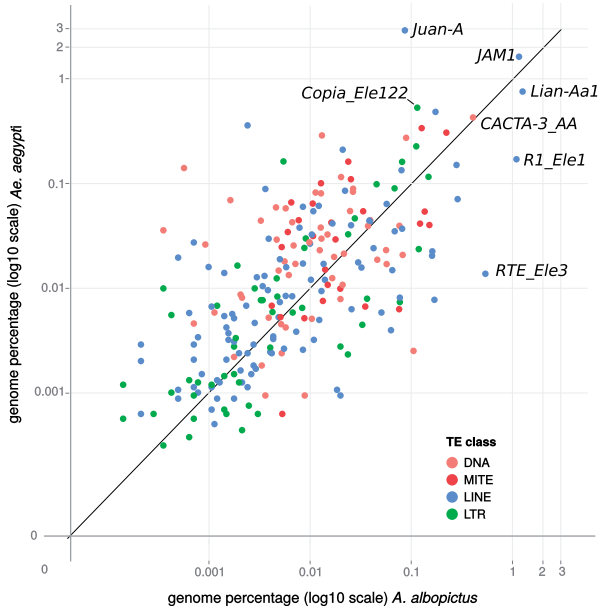


Fig. 6.—Comparison of the relative genome proportions of shared TE families between *Aedes albopictus* and *Aedes aegypti* in terms of genome percentage ($\log_{10}$ scale). Each dot represents a shared TE family, defined by a more similar BLAT hit between the TE family reference contig of each species. Names on the graphs correspond to the main TE annotation (from *A. aegypti*) discussed in the text.

annotate a larger fraction of TEs and to compute larger contigs. However, RE seems to more sensitively estimate the proportion of low complexity and tandem repeat sequences (data not shown).

Discussion

The *A. albopictus* repeatome

We report the first description of the *A. albopictus* repeatome using dnaPipeTE, a new bioinformatic pipeline for the de novo estimation, annotation, and assembly of repeatomes from raw genomic reads. We found that the total amount of repeated DNA reached 49.13% of the genome that includes at least 33.58% TEs. Taking into account that this method will underestimate low copy number TEs as well as older copies that were unable to be assembled due to mutation accumulation, our estimation should be viewed as a lower bound for the TE content of *A. albopictus*. As 7.23% of the genome is still unannotated repeats, it is possible that the TE content of *A. albopictus* ranks the largest among mosquitoes (fig. 7; Holt et al. 2002; Nene et al. 2007; Arensburger et al. 2011; Marinotti et al. 2013; Zhou et al. 2014). The large repeatome of *A. albopictus* contributes to half of its genome size, which is consistent with the observed relation between genome size and TE content (Biémont and Vieira 2004; Chénais et al. 2012). This relation exists between published genome sizes and TE content of other mosquitoes (fig. 7, $r^2 = 0.82$, $P < 0.01$).

TE families can be extremely different from each other and are classified into several subfamilies. In a given genome, some TE families are present in few copies, while others can reach hundreds of thousands of copies. In *A. albopictus*, the largest TE families in terms of genome proportion and copy numbers are LINE (non-LTR) retroelements, which harbor thousands of copies per family and represent 12.09% of the genome.

Table 2

Performance Comparison between dnaPipeTE and RepeatExplorer Using *Aedes Albopictus* and *Drosophila melanogaster* Samples

		Computing Time	Contig Number	Assembly N50 (bp)	Repeat Content Estimation	Repeat Annotation
A. albopictus	dnaPipeTE	3 h 07 min (8 CPUs/40 Go RAM)	8102	677	49.13%	85.3%
	RepeatExplorer	2 days 5 h 12 min (8 CPUs/16 Go RAM)	14615	198	51.0%	25.5%
D. melanogaster	dnaPipeTE	0 h 40 min (8 CPUs/15 Go RAM)	2054	2,590	18%	98.8%
	RepeatExplorer	6 h 05 min (8 CPUs/16 Go RAM)	1352	287	16.5%	86.1%

NOTE.—Repeat annotation percentage was computed by counting the number of genomic reads receiving an annotation for each method.

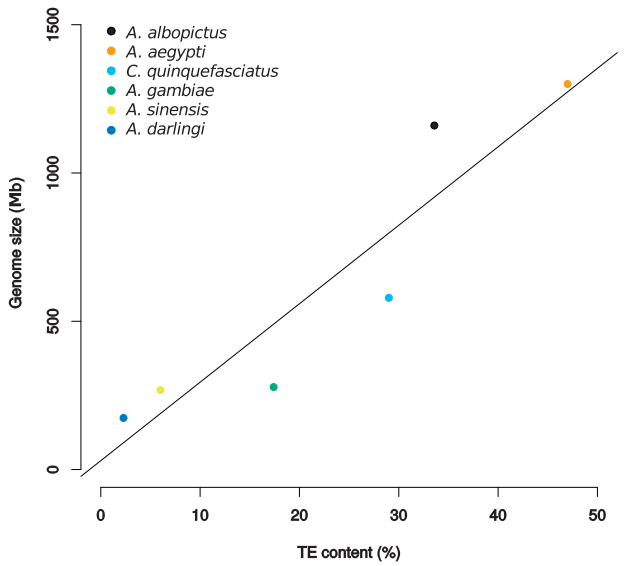


FIG. 7.—Linear regression of genome size over TE content in mosquitoes. Except for *Aedes albopictus*, data come from complete sequenced genomes cited in the text. ($r^2 = 0.827$, $P < 0.01$).

These LINEs represent several well-known superfamilies that have been described in mosquitoes, such as I (*Lian*, *R1, Loa*, and *Loner* families) and RTE (Tu et al. 1998; Biedler and Tu 2003; Boulesteix and Biémont 2005). LINEs are also found in high copy number in *A. aegypti*, where they represent 14% of the genome (Nene et al. 2007). At the class level, the most abundant class of TE is the Class II, with a majority of DNA transposons and MITEs. This feature is shared by the *A. aegypti* genome, in which Class II elements are also the most abundant repeats, comprising 20% of genome proportion, including 16% of MITEs.

TE Dynamics and Comparison with *Aedes aegypti*

Comparison of the two related *Aedes* species highlighted a convergence in TE landscapes at the superfamily level. Both species display a similar distribution of sequenced TE reads against their contig sequences for the three TEs studied (LTR, LINEs [Class I], and Class II). In these species, Class I elements (RNA-mediated transposition) showed a right-skewed distribution, meaning that copies of each TE family share a high identity. This is typical of recent or active TE families, in which the copy number increases faster than the accumulation of mutations within the copies (Lerat et al. 2011; Staton et al. 2012). This pattern can be seen in species such as *D. melanogaster* or *An. gambiae*, in which Class I elements showed recent amplifications (Biedler and Tu 2003; Kapitonov and Jurka 2003; see also the genome analysis available online at http://repeatmasker.org/genomicDatasets/RMGenomicDatasets.html, last accessed April 13, 2015).

In both mosquito species, DNA-based transposons (Class II) are poorly represented compared with their relative genome proportion. However, this result might be explained by the removal of *A. aegypti* TE references from the library to avoid any bias toward this species in the annotation, which might have removed elements specific to the *Aedes* genus. Another explanation is that DNA transposons could belong to families with very few copies and/or result from an old invasion of the genome. Thus, our methodology, which is weaker beyond 15% divergence and for elements with few copies, could have missed old Class II elements. Ultimately, this could mean either that members of Class II are the first TEs to have invaded *Aedes* genomes or that Class I TEs are undergoing a new expansion wave.

Despite these similarities in the TE age distributions, the LINE/Jockey superfamily is different between these two species. Indeed, these elements are rare (0.04% of the blasted reads) in *A. albopictus*, where only recent copies are found. However, in *A. aegypti*, they represent half of the LINEs, and the LINE/*Juan-A* is the most abundant TE, representing 3% of the genome (Nene et al. 2007). Conversely, *A. albopictus* harbors more LINE/I elements than *A. aegypti*, and their distribution indicates a higher number of divergent copies, which suggests that their amplification in the *A. albopictus* genome could have begun earlier than in *A. aegypti* following the divergence of these two species.

The distinction between *A. albopictus* and *A. aegypti* is even more striking when observing the abundance of the TE families they share. Indeed, the abundance of TEs copies is very

different from one genome to another. This indicates that while both species share similar trends in TE class dynamics, a TE expansion occurred independently in each species. This observation could be interpreted in the ecological framework of TE dynamics and evolution (Venner et al. 2009; Linquist et al. 2013). Indeed, "ecological" factors affecting the genome, such as GC content or genome size, have been shown to be linked to TE abundance and distribution in related species (Jurka et al. 2011). Thus, inheritance of a common genome and ecosystem from an ancestor could have constrained superfamily dynamics in both species, considering either the possible interaction between TEs (identical to interspecific competition) or between TEs and the genome architecture (Venner et al. 2009; Linquist et al. 2013). However, at the family level, the spread of one TE family instead of another is not subject to ecological constraint (Jurka et al. 2011). For instance, the general pattern of a recent invasion of LTRs and LINEs in the *Aedes* species studied here can still be observed, while the specific TE families amplified in each species differ. In addition, both *A. albopictus* and *A. aegypti* are examples of species with numerous subdivided populations in their native areas (Hawley 1988; Mousson et al. 2005; Brown et al. 2014) and a relatively limited natural dispersion capability (Reiter 1996; Bellini et al. 2010; Medley et al. 2015), which increases the probability of differential TE fixation in isolated subpopulations (Jurka et al. 2011). Therefore, the sequenced individuals are only representative of the subpopulations to which they belong, and it would be interesting to compare TE family diversity at the subpopulation level with regard to intraspecific genome size variation imparted by TEs in *A. albopictus* (McLain et al. 1987; Black and Rai 1988).

dnaPipeTE: A Novel Tool for TE Comparative Studies

Preliminary work on the *A. albopictus* repeatome led us to develop our own pipeline in order to address specific unmet needs. As the *A. albopictus* genome is especially large, we were interested in solutions using low coverage sequencing to find and quantify TEs and interspersed repeats. The most advanced software for this task previously available was RE (Novák et al. 2010), which allows the simultaneous location, quantification, and annotation of repeats from unassembled sequencing reads. However, we felt that some points could be improved by using NGS-specific tools. By using Trinity as a TE assembler on small genomic data sets, dnaPipeTE can recover larger TE contigs and can improve this step by performing multiple iterations with additional independent samples. dnaPipeTE can annotate and quantify TE families with its contigs and the number of mapped reads, while RE annotation is given only for sampled reads. Our method allowed the identification of more repeats in *A. albopictus* than RE, with a substantial decrease in computational time. As with other library-based tools, this automatic annotation should be considered with caution when working on species with very few reference libraries, where the similarities between hits might be weak and could lead to annotation errors. However, tests on model species showed that dnaPipeTE performed well in the estimation of the TE content and the proportions of the main TE families. Although it was not designed for de novo identification of new TE families, dnaPipeTE can produce full-length contigs of TEs that could be manually annotated at a later point. dnaPipeTE also provides a large amount of usable output (summary tables, graphs, sorted data sets). Finally, dnaPipeTE is the first method capable of generating a representation of TE age distribution without prior genome assembly. This analysis of course has some limitations. First, the BLAST method allows the detection of variation only from 0% to 15% divergence. Second, considering two divergent copies in a TE family, the accumulation of mutations will not be evenly distributed along the sequence; reads from a conserved protein domain will be more similar to the contig than nonfunctional regions due to selective constraints, biasing the TE age distribution toward recent divergence. In the future, the effects of these drawbacks will be reduced by the use of longer reads, which dnaPipeTE is already equipped to handle. In conclusion, this new bioinformatic pipeline, available for download at https://lbbe.univ-lyon1.fr/-dnaPipeTE-.html, allowed us to perform a fast and comprehensive analysis of TEs and repeat elements in a newly sequenced genome using NGS raw data with only 0.3× genome coverage. It allows the design of "low sequencing experiments" that reduce sequencing cost and facilitate an increase in the number of samples compared. The consistency and the robustness of dnaPipeTE also allow for comparative studies such as the one presented in this article.

Our study showed that the repeatome of *A. albopictus* is huge, encompassing 50% of the genome, and that it shares notable similarities with *A. aegypti* at the main TE order level. The intrafamily dynamics of TEs show high variation between species. Since the divergence of *A. albopictus* and *A. aegypti* 10 million years ago (Pashley and Rai 1983), TE families seemed to have evolved independently from ancestral TE ecology. These pictures of the two *Aedes* species' repeatomes could explain the large genome size variation due to repetitive DNA reported at the intraspecific level (McLain et al. 1987; Black and Rai 1988).

Supplementary Material

Supplementary figures S1–S4, tables S1 and S2, and Material are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank the members of the Transposable Elements, Evolution, and Population team for testing dnaPipeTE and for providing suggestions and constructive discussion about

this article. We also thank Petr Novák for his availability to provide help and information about the RepeatExplorer pipeline and Murray Patterson for English revisions. This work was performed using the computing facilities of the CC LBBE/PRABI. This work was supported by the Agence Nationale de la Recherche (ANR Genemobile), the Centre National de La Recherche Scientifique, the Institut Universitaire de France, and the Federation de Recherche 41 "Bio-Environnement et Santé." C.G. received a grant from the French Ministry of Superior Education. This study was also supported by The *A. albopictus* genome sequencing project which is partly funded by the Agence Nationale de la Recherche (ANR Immunsymbart).

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Arensburger P, Hice RH, Wright JA, Craig NL, Atkinson PW. 2011. The mosquito *Aedes aegypti* has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. BMC Genomics 12:606.

Beck CR, Garcia-Perez JL, Badge RM, Moran JV. 2011. LINE-1 elements in structural variation and disease. Annu Rev Genomics Hum Genet. 12: 187–215.

Bellini R, et al. 2010. Dispersal and survival of *Aedes albopictus* (Diptera: Culicidae) males in Italian urban areas and significance for sterile insect technique application. J Med Entomol. 47:1082–1091.

Biedler J, Tu Z. 2003. Non-LTR retrotransposons in the African malaria mosquito, *Anopheles gambiae*: unprecedented diversity and evidence of recent activity. Mol Biol Evol. 20:1811–1825.

Biémont C, Vieira C. 2004. [The influence of transposable elements on genome size]. J Soc Biol. 198:413–417.

Black WC, Ferrari JA, Sprengert D. 1988. Breeding structure of a colonising species: *Aedes albopictus* (Skuse) in the United States. Heredity (Edinb) 60(Pt 2):173–181.

Black WC, Rai KS. 1988. Genome evolution in mosquitoes: intraspecific and interspecific variation in repetitive DNA amounts and organization. Genet Res. 51:185–196.

Bonizzoni M, Gasperi G, Chen X, James AA. 2013. The invasive mosquito species *Aedes albopictus*: current knowledge and future perspectives. Trends Parasitol. 29:460–468.

Boulesteix M, Biémont C. 2005. Transposable elements in mosquitoes. Cytogenet Genome Res. 110:500–509.

Brown JE, et al. 2014. Human impacts have shaped historical and recent evolution in *Aedes aegypti*, the dengue and yellow fever mosquito. Evolution 68:514–525.

Casacuberta E, González J. 2013. The impact of transposable elements in environmental adaptation. Mol Ecol. 22:1503–1517.

Chénais B, Caruso A, Hiard S, Casse N. 2012. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. Gene 509: 7–15.

Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 38:1767–1771.

Goodier JL, Kazazian HH. 2008. Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell 135:23–35.

Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29: 644–652.

Hawley WA. 1988. The biology of *Aedes albopictus*. J Am Mosq Control Assoc Suppl. 1:1–39.

Holt RA, et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298:129–149.

Huang X. 1999. CAP3: a DNA sequence assembly program. Genome Res. 9:868–877.

Jurka J, Bao W, Kojima KK. 2011. Families of transposable elements, population structure and the origin of species. Biol Direct. 6:44.

Jurka J, et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110:462–467.

Kapitonov VV, Jurka J. 2003. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. Proc Natl Acad Sci U S A. 100:6569–6574.

Koch P, Platzer M, Downie BR. 2014. RepARK—de novo creation of repeat libraries from whole-genome NGS reads. Nucleic Acids Res. 42:e80.

Kumar A, Rai KS. 1990. Intraspecific variation in nuclear DNA content among world populations of a mosquito, *Aedes albopictus* (Skuse). Theor Appl Genet. 79:748–752.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9:357–359.

Lerat E, Burlet N, Biémont C, Vieira C. 2011. Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. Gene 473:100–109.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659.

Linquist S, et al. 2013. Distinguishing ecological from evolutionary approaches to transposable elements. Biol Rev Camb Philos Soc. 88: 573–584.

Marinotti O, et al. 2013. The genome of *Anopheles darlingi*, the main neotropical malaria vector. Nucleic Acids Res. 41:7387–7400.

McLain DK, Rai KS, Fraser MJ. 1987. Intraspecific and interspecific variation in the sequence and abundance of highly repeated DNA among mosquitoes of the *Aedes albopictus* subgroup. Heredity (Edinb) 58: 373–381.

Medley KA, Jenkins DG, Hoffman EA. 2015. Human-aided and natural dispersal drive gene flow across the range of an invasive mosquito. Mol Ecol. 24:284–295.

Modolo L, Lerat E. 2014. Identification and analysis of transposable elements in genomic sequences. In: Poptsova MS, editor. Genome analysis: current procedures and application. Norfolk (UK): Caister Academic Press. p. 165–181.

Mousson L, et al. 2005. Phylogeography of *Aedes (Stegomyia) aegypti* (L.) and *Aedes (Stegomyia) albopictus* (Skuse) (Diptera: Culicidae) based on mitochondrial DNA variations. Genet Res. 86:1–11.

Nene V, et al. 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science 316:1718–1723.

Novák P, Neumann P, Macas J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11:378.

Pashley DP, Rai KS. 1983. Comparison of allozyme and morphological relationships in some *Aedes* (*Stegomyia*) mosquitoes (Diptera: Culicidae). Ann Entomol Soc Am. 76:388–394.

Rao PN, Rai KS. 1987. Inter and intraspecific variation in nuclear DNA content in *Aedes* mosquitoes. Heredity (Edinb) 59:253–258.

Rebollo R, Horard B, Hubert B, Vieira C. 2010. Jumping genes and epigenetics: towards new species. Gene 454:1–7.

Reiter P. 1996. [Oviposition and dispersion of *Aedes aegypti* in an urban environment]. Bull Soc Pathol Exot. 89:120–122.

Staton SE, et al. 2012. The sunflower (*Helianthus annuus* L.) genome reflects a recent history of biased accumulation of transposable elements. Plant J. 72:142–153.

Tange O. 2011. GNU parallel: the command-line power tool. ;login USENIX Mag. 3:42–47.

Tu Z, Isoe J, Guzova JA. 1998. Structural, genomic, and phylogenetic analysis of *Lian*, a novel family of non-LTR retrotransposons in the yellow fever mosquito, *Aedes aegypti*. Mol Biol Evol. 15:837–853.
Vela D, Fontdevila A, Vieira C, García Guerreiro MP. 2014. A genome-wide survey of genetic instability by transposition in *Drosophila* hybrids. PLoS One 9:e88992.
Venner S, Feschotte C, Biémont C. 2009. Dynamics of transposable elements: towards a community ecology of the genome. Trends Genet. 25:317–323.
Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973–982.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829.
Zhou D, et al. 2014. Genome sequence of *Anopheles sinensis* provides insight into genetics basis of mosquito competence for malaria parasites. BMC Genomics 15:42.
Zytnicki M, Akhunov E, Quesneville H. 2014. Tedna: a transposable element de novo assembler. Bioinformatics 30:2656–2658.

Laurent Modolo

Analyse bioinformatique des événements de transferts horizontaux entre espèces de drosophiles et lien avec la régulation des éléments transposables

ANNEXES

Table des matières

1 A New Genome-Wide Method to Track Horizontally Transferred Sequences : Application to Drosophila

Additional Files

Supplementary Figure S1: Detailed data-flow diagram of the different steps in the genome-wide detection of horizontally transferred sequences between two genomes. It consists of three steps for the detection of all pairs of sequences with a nucleotidic identity higher than expected between two genomes, and two filtering steps to remove spurious detections from the results.

Supplementary Figure S2: Density distributions of the *activity tracks* for all the TEs detected between *D. melanogaster* and the 2012 version of the genome of *D. simulans*. The reds bars represent the *activity tracks* in *D. melanogaster* while the blue bars represent the *activity tracks* in the other species. The elements with an *activity tracks* consitent with a recent arrival in the genome of *D. melanogater* by HT are represented with a green background title

Supplementary Figure S3: Density distributions of the *activity tracks* for all the TEs detected between *D. melanogaster* and the 2007 version of the genome of *D. simulans*. The reds bars represent the *activity tracks* in *D. melanogaster* while the blue bars represent the *activity tracks* in the other species. The elements with an *activity tracks* consitent with a recent arrival in the genome of *D. melanogater* by HT are represented with a green background title.

Supplementary Figure S4: Density distributions of the *activity tracks* for all the TEs detected between *D. melanogaster* and *D. sechellia*. The red bars represent the *activity tracks* in *D. melanogaster* while the blue bars represent the *activity tracks* in the other species.The elements with an *activity tracks* consitent with a recent arrival in the genome of *D. melanogater* by HT are represented with a green background title

Supplementary Figure S5: Density distributions of the *activity tracks* for all the TEs detected between *D. melanogaster* and *D. yakuba*. The red bars represent the *activity tracks* in *D. melanogaster* while the blue bars represent the *activity tracks* in the other species.The elements with an *activity tracks* consitent with a recent arrival in the genome of *D. melanogater* by HT are represented with a green background title

Supplementary Figure S6: Density distributions of the *activity tracks* for all the TEs detected between *D. melanogaster* and *D. pseudoobscura*. The red bars represent the *activity tracks* in *D. melanogaster* while the blue bars represent the *activity tracks* in the other species.

Supplementary Figure S7: Density distributions of the *activity tracks* for all the TEs detected between *D. melanogaster* and *D. virilis*. The red bars represent the *activity tracks* in *D. melanogaster* while the blue bars represent the *activity tracks* in the other species.

Supplementary Table S1: Annotation of the intergenic DNA fragments detected with the *D. melangoaster-D. simulans* analysis, before and after the filtering with the results between *D. melanogaster* and other *Drosophila* species of the phylogeny.

Supplementary Table S2: List of TEs families detected and validated based on their *activity tracks* between *D. melanogaster* and the corresponding species.

Figure S1

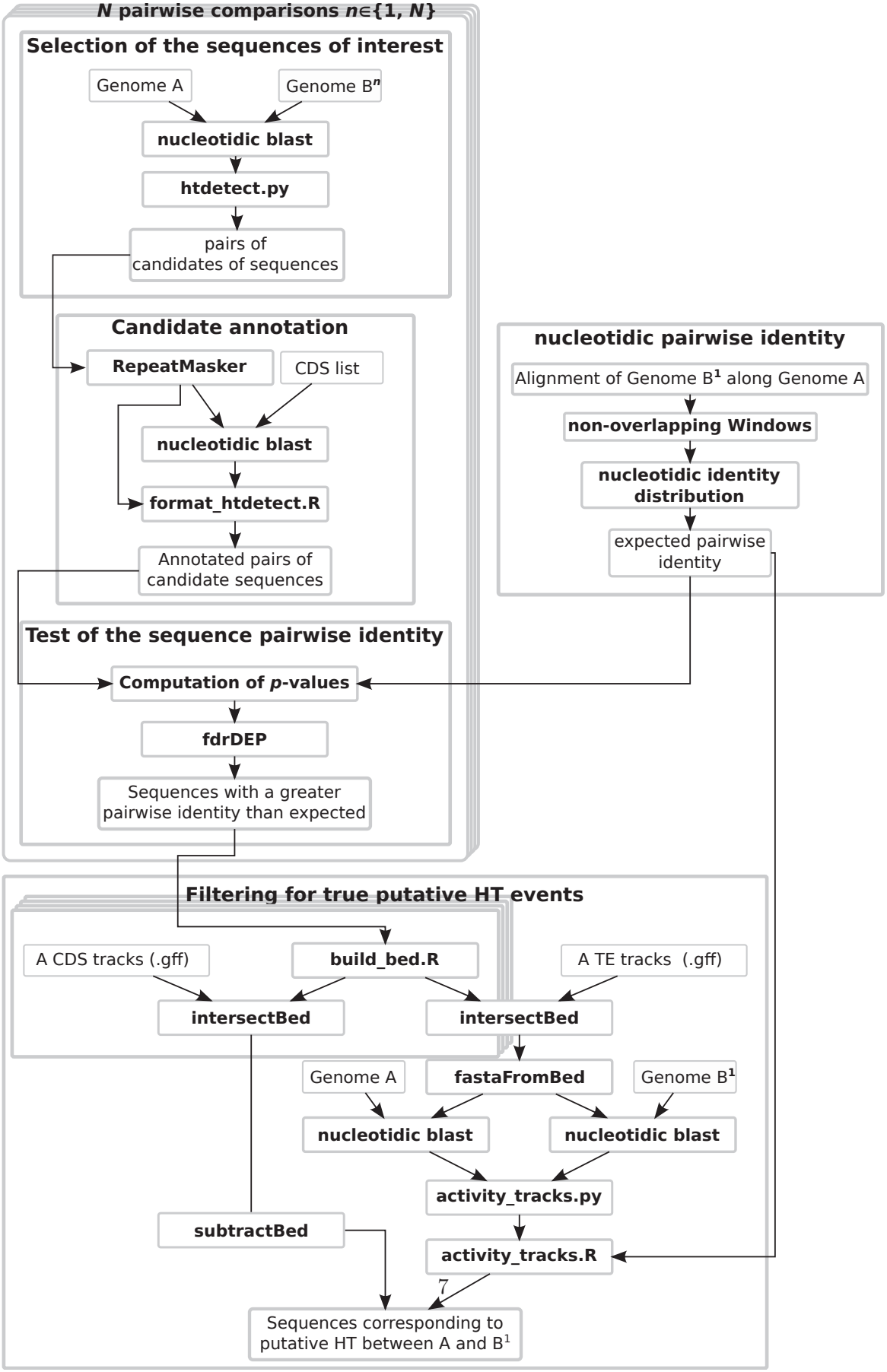
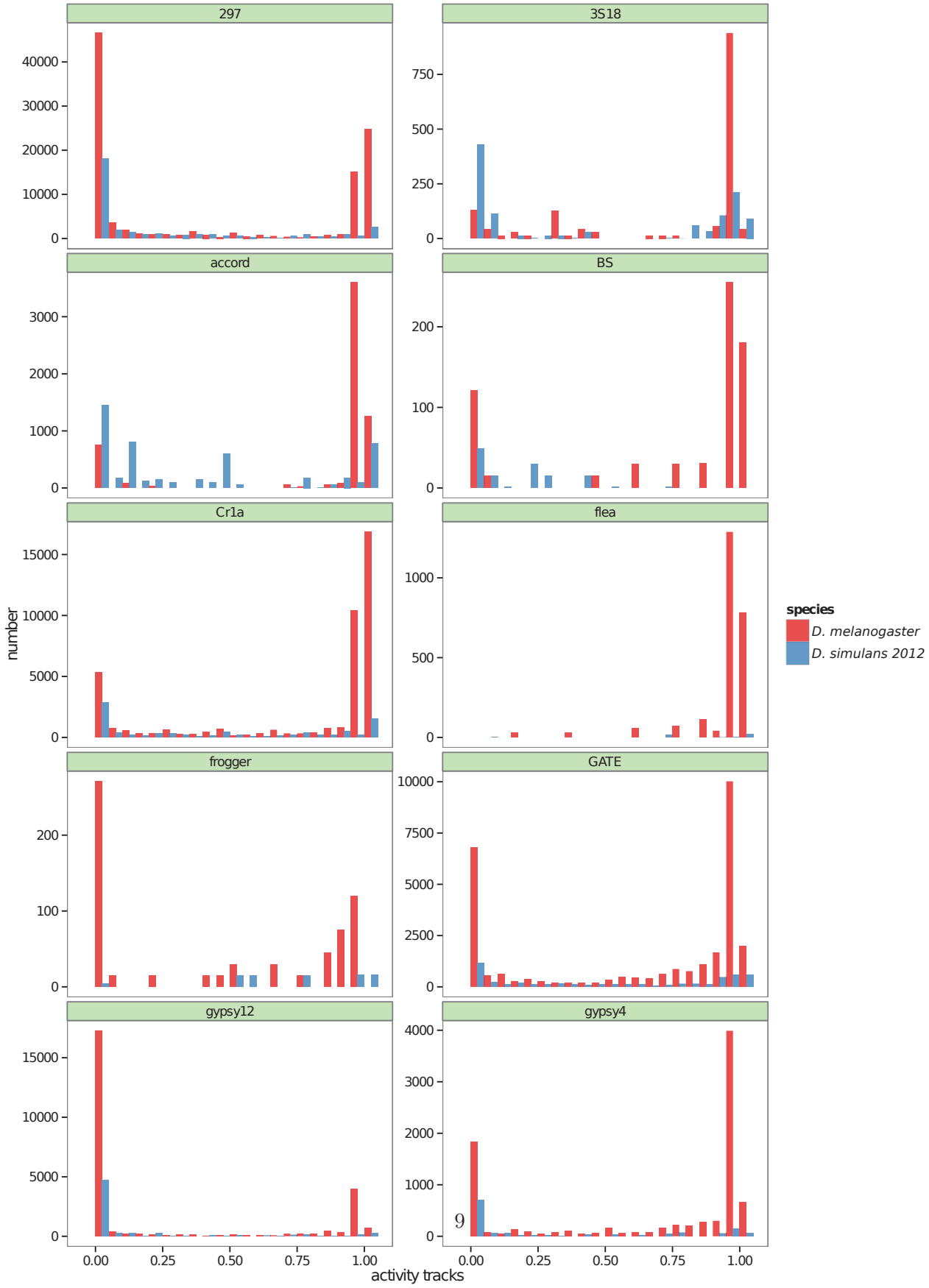
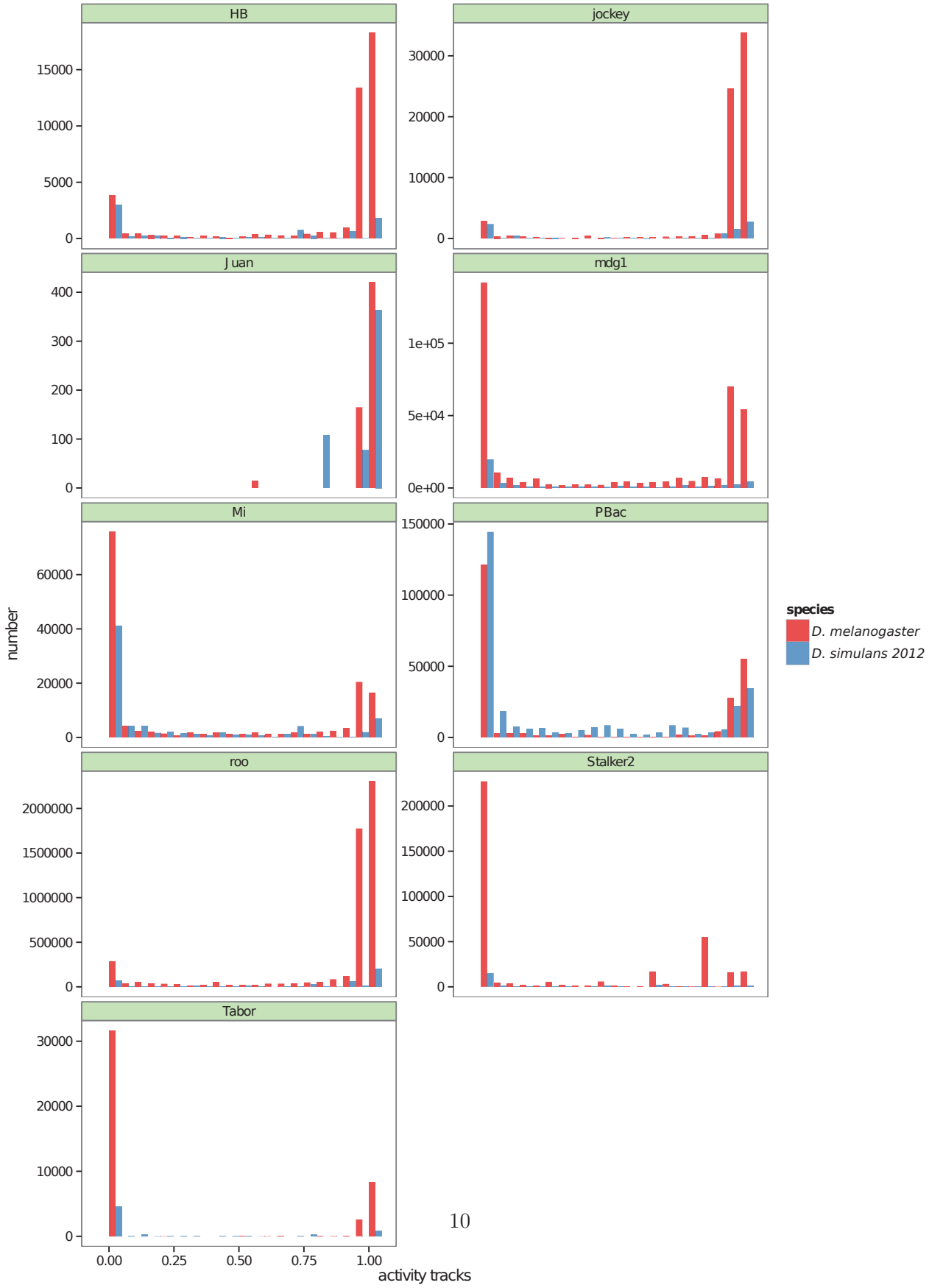

N pairwise comparisons n∈{1, N}
Selection of the sequences of interest
Genome A
Genome Bn
nucleotidic blast
htdetect.py
pairs of candidates of sequences
Candidate annotation
RepeatMasker
CDS list
nucleotidic blast
format_htdetect.R
Annotated pairs of candidate sequences
nucleotidic pairwise identity
Alignment of Genome B1 along Genome A
non-overlapping Windows
nucleotidic identity distribution
expected pairwise identity
Test of the sequence pairwise identity
Computation of p-values
fdrDEP
Sequences with a greater pairwise identity than expected
Filtering for true putative HT events
A CDS tracks (.gff)
build_bed.R
A TE tracks (.gff)
intersectBed
intersectBed
Genome A
fastaFromBed
Genome B1
nucleotidic blast
nucleotidic blast
activity_tracks.py
subtractBed
activity_tracks.R
Sequences corresponding to putative HT between A and B1

Figure S2


297
3S18
accord
BS
Cr1a
flea
frogger
GATE
gypsy12
gypsy4
number
activity tracks
species
D. melanogaster
D. simulans 2012


HB
jockey
Juan
mdg1
Mi
PBac
roo
Stalker2
Tabor
number
activity tracks
0.00
0.25
0.50
0.75
1.00
species
D. melanogaster
D. simulans 2012

10

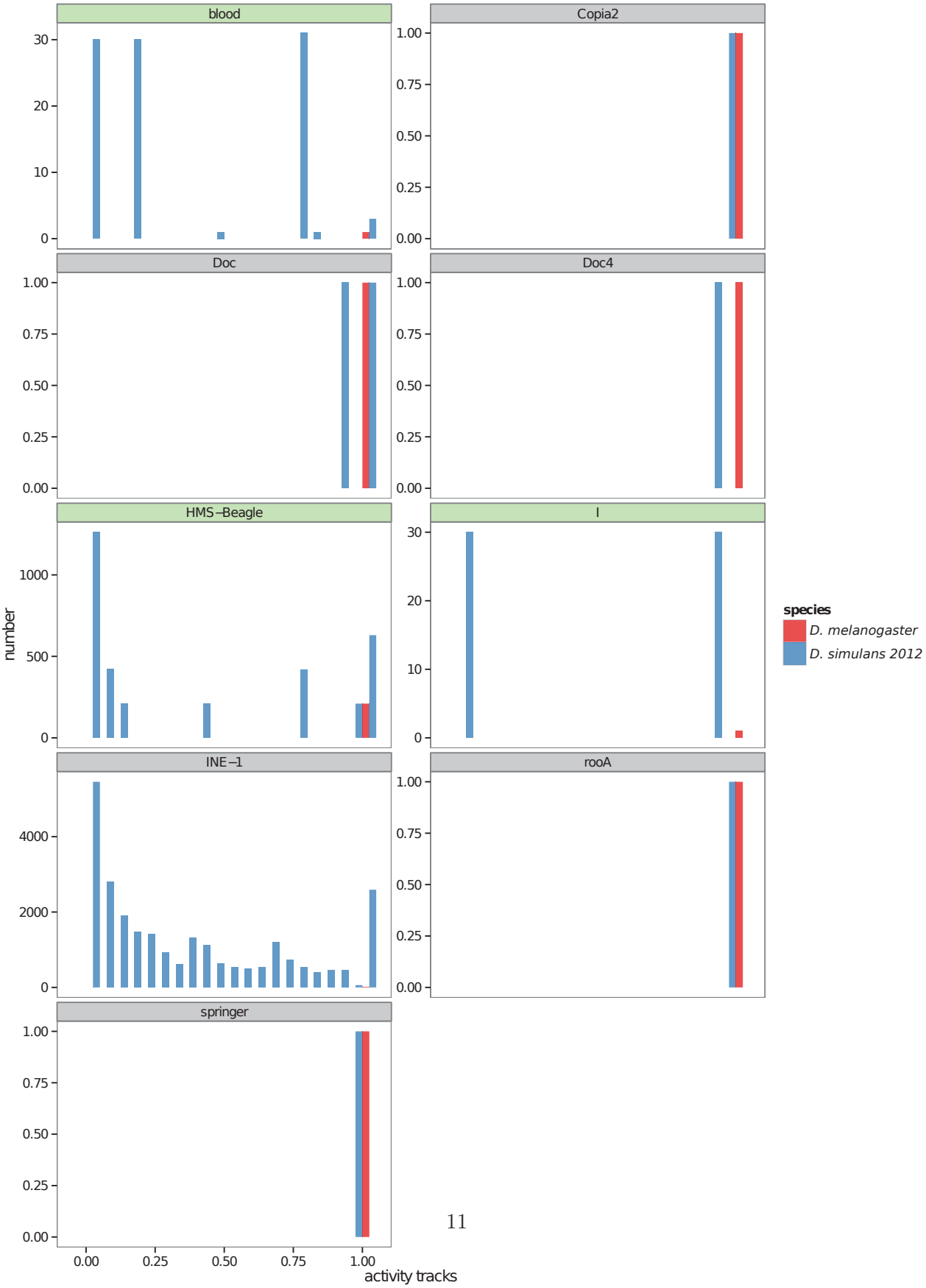
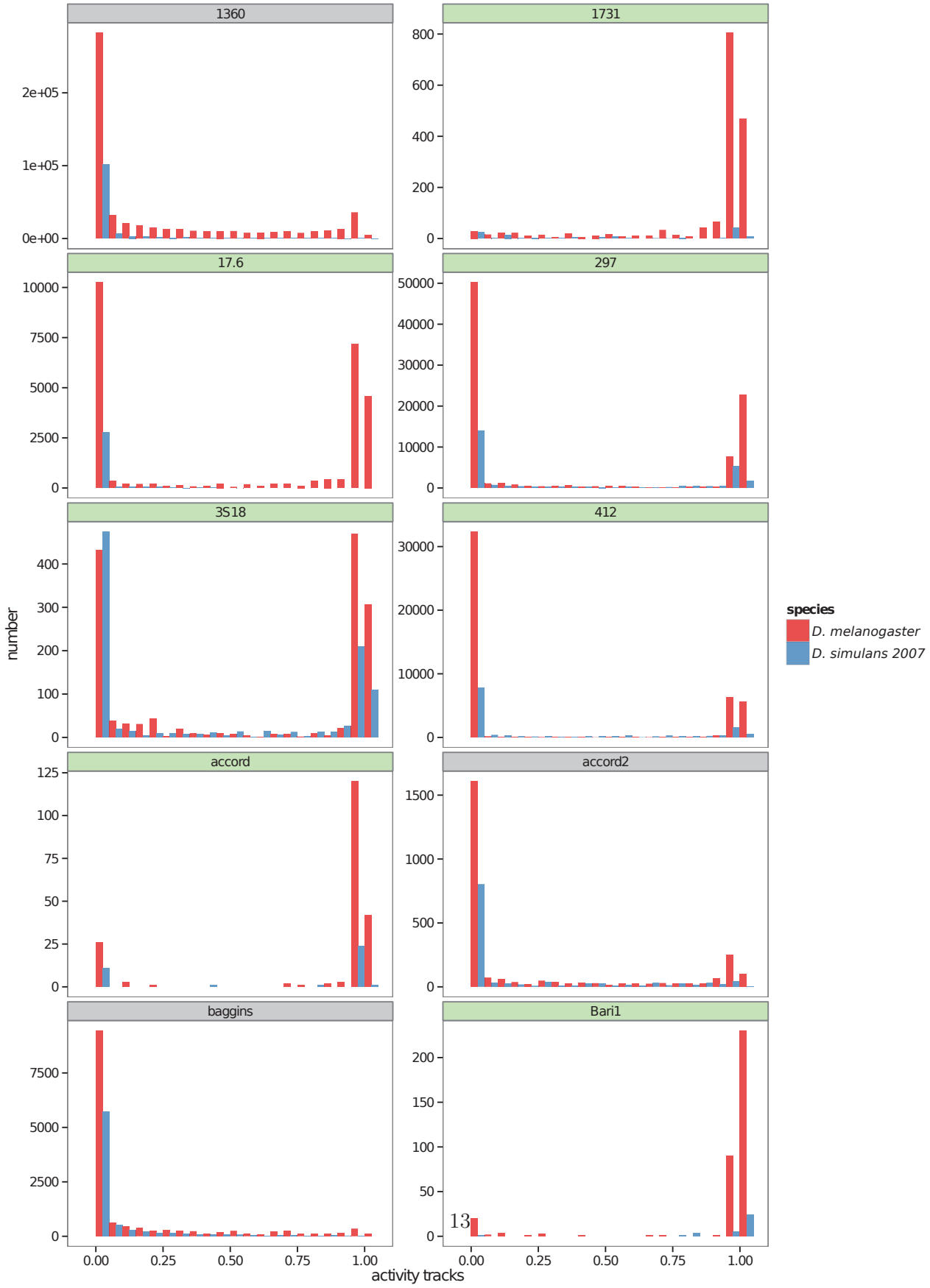
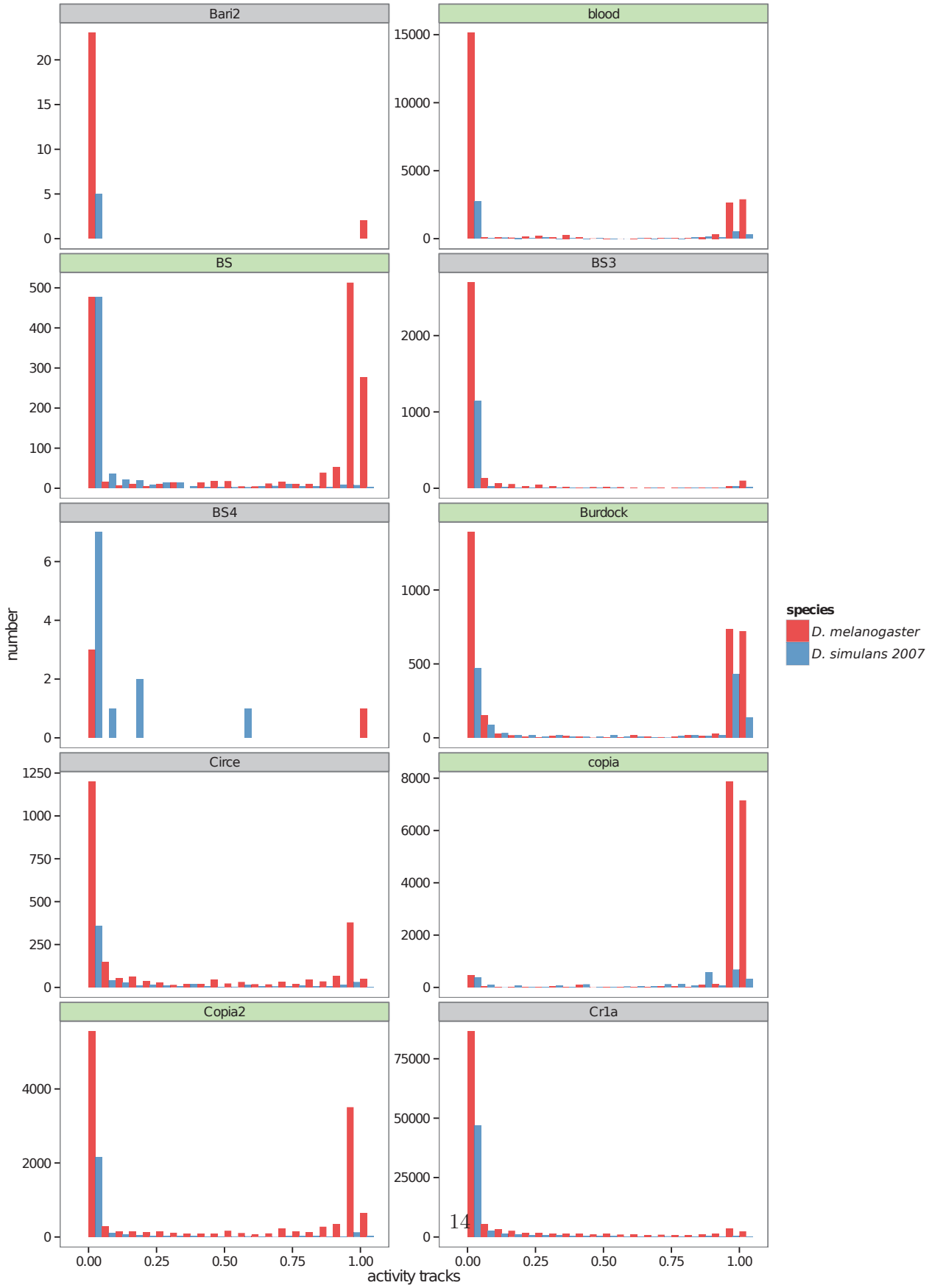
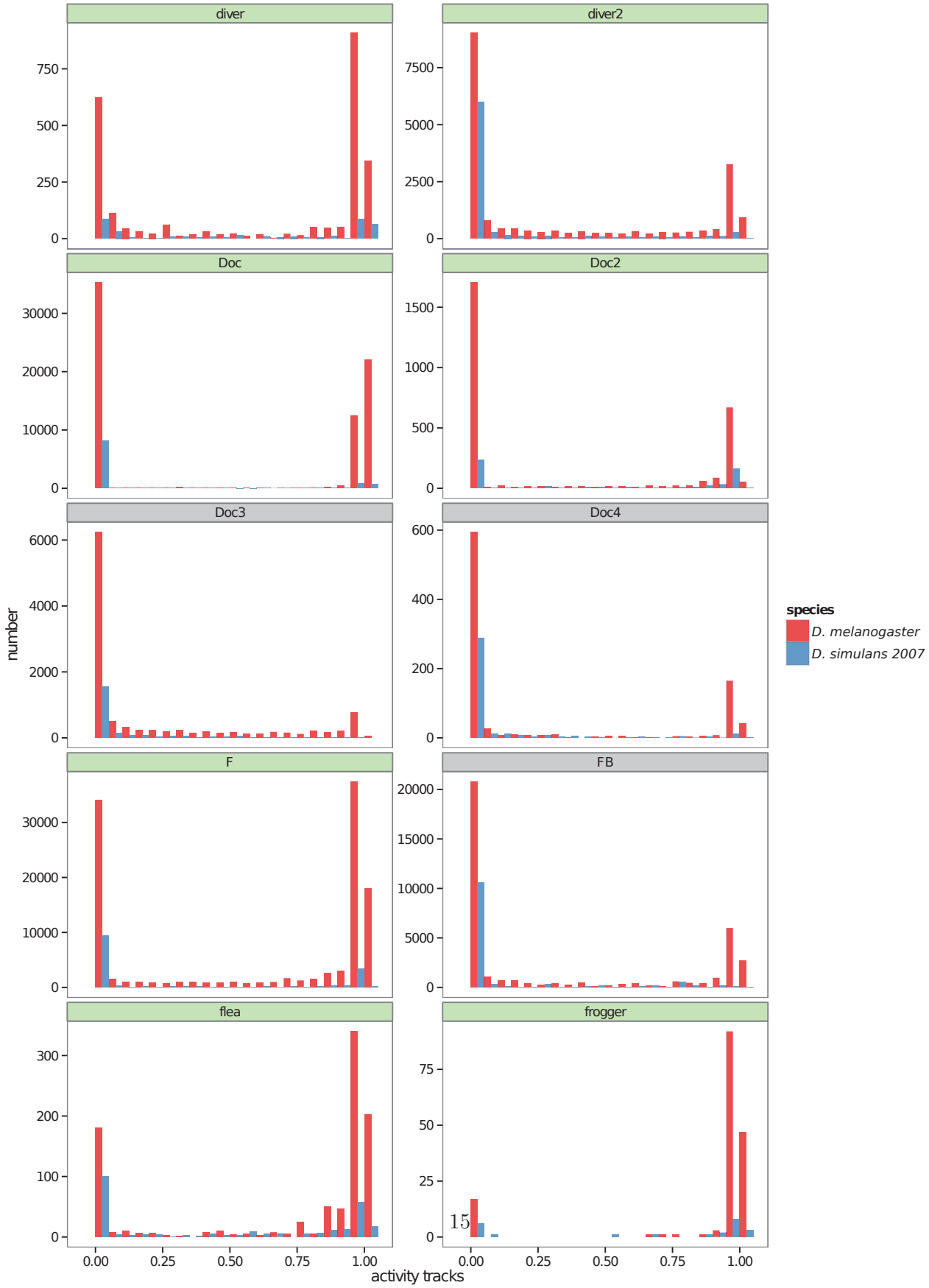

blood
Copia2
Doc
Doc4
HMS–Beagle
I
INE–1
rooA
springer
number
activity tracks
species
D. melanogaster
D. simulans 2012

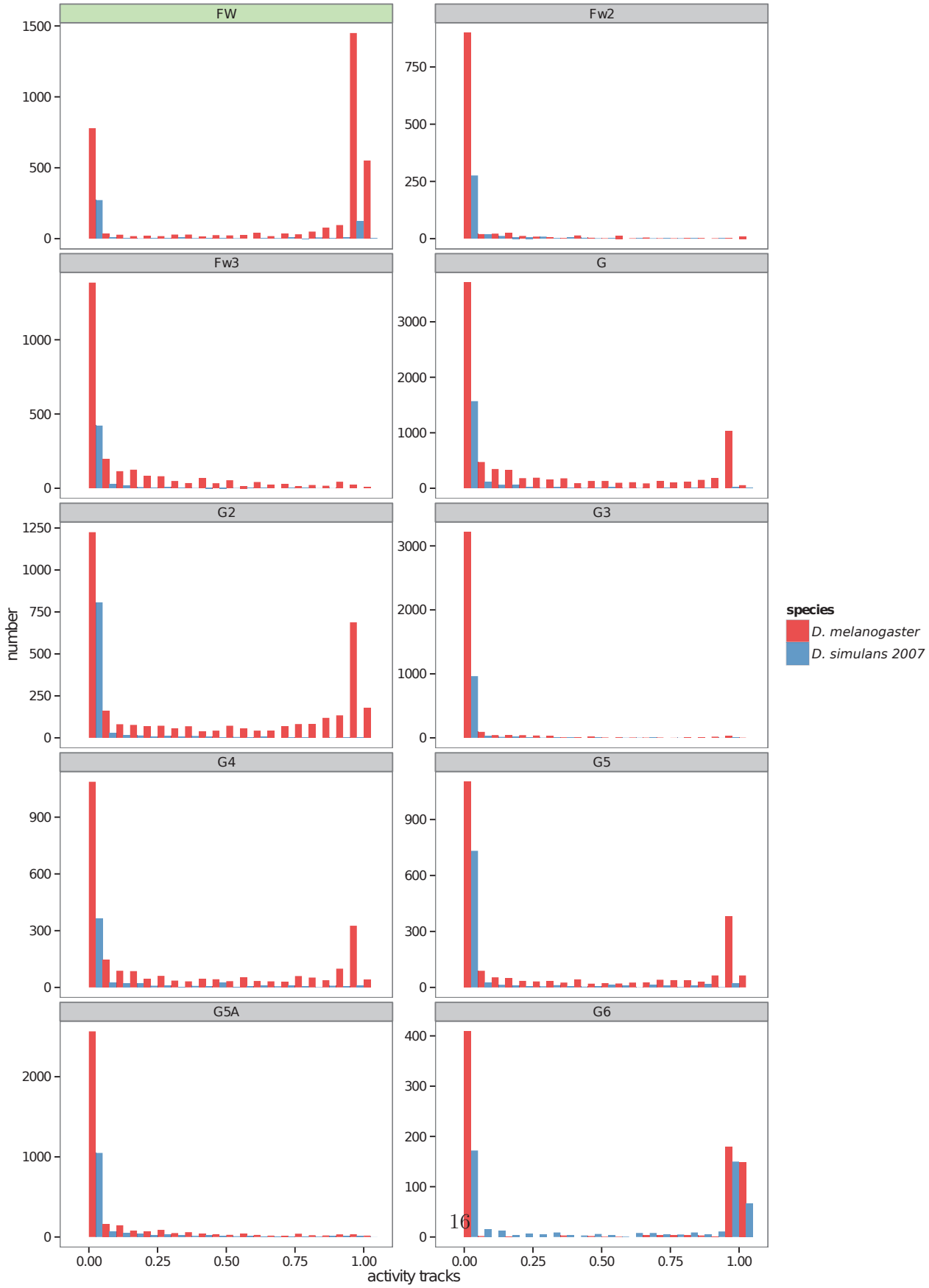
Figure S3

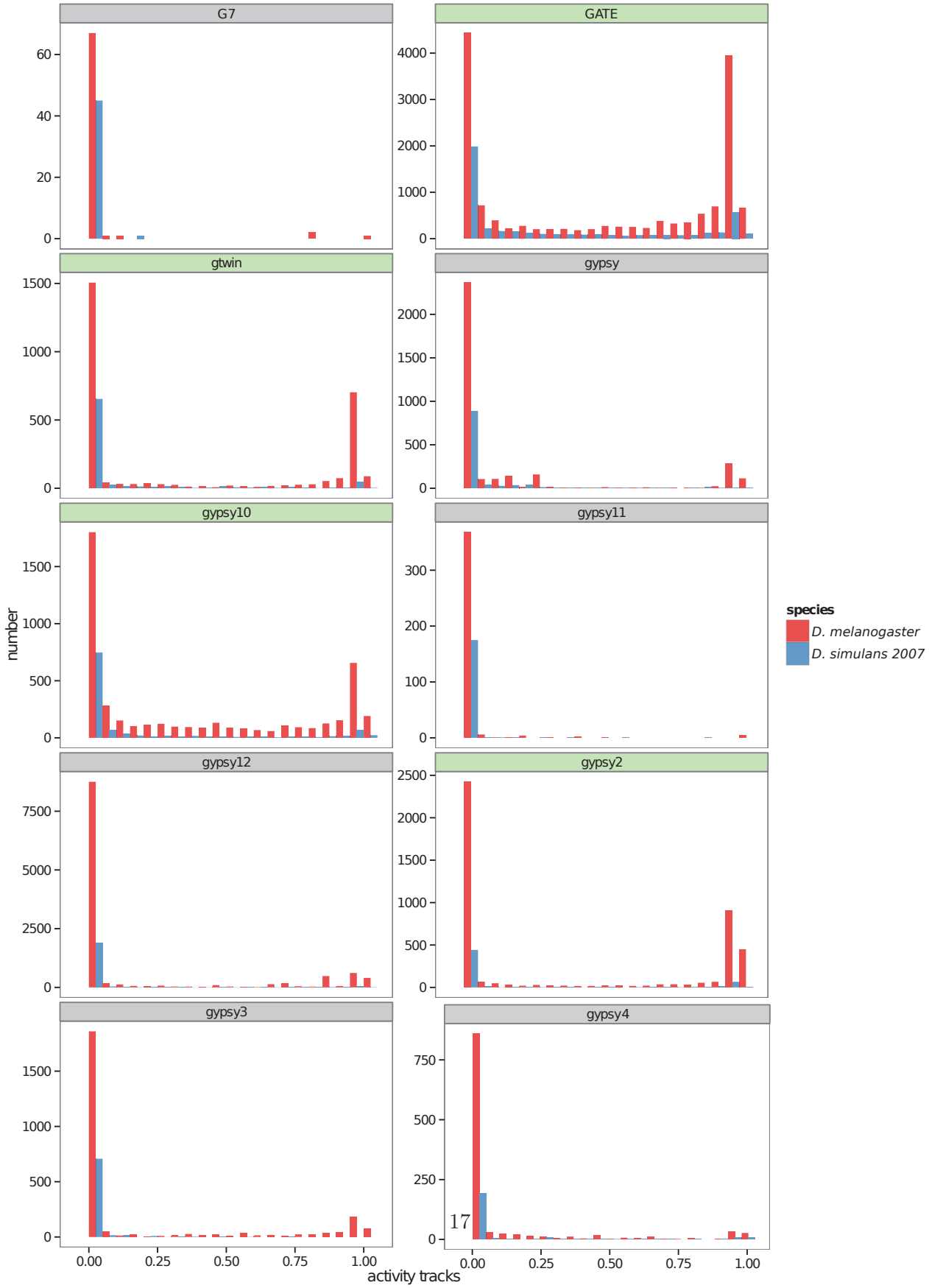

1360
1731
17.6
297
3S18
412
accord
accord2
baggins
Bari1
number
activity tracks
species
D. melanogaster
D. simulans 2007

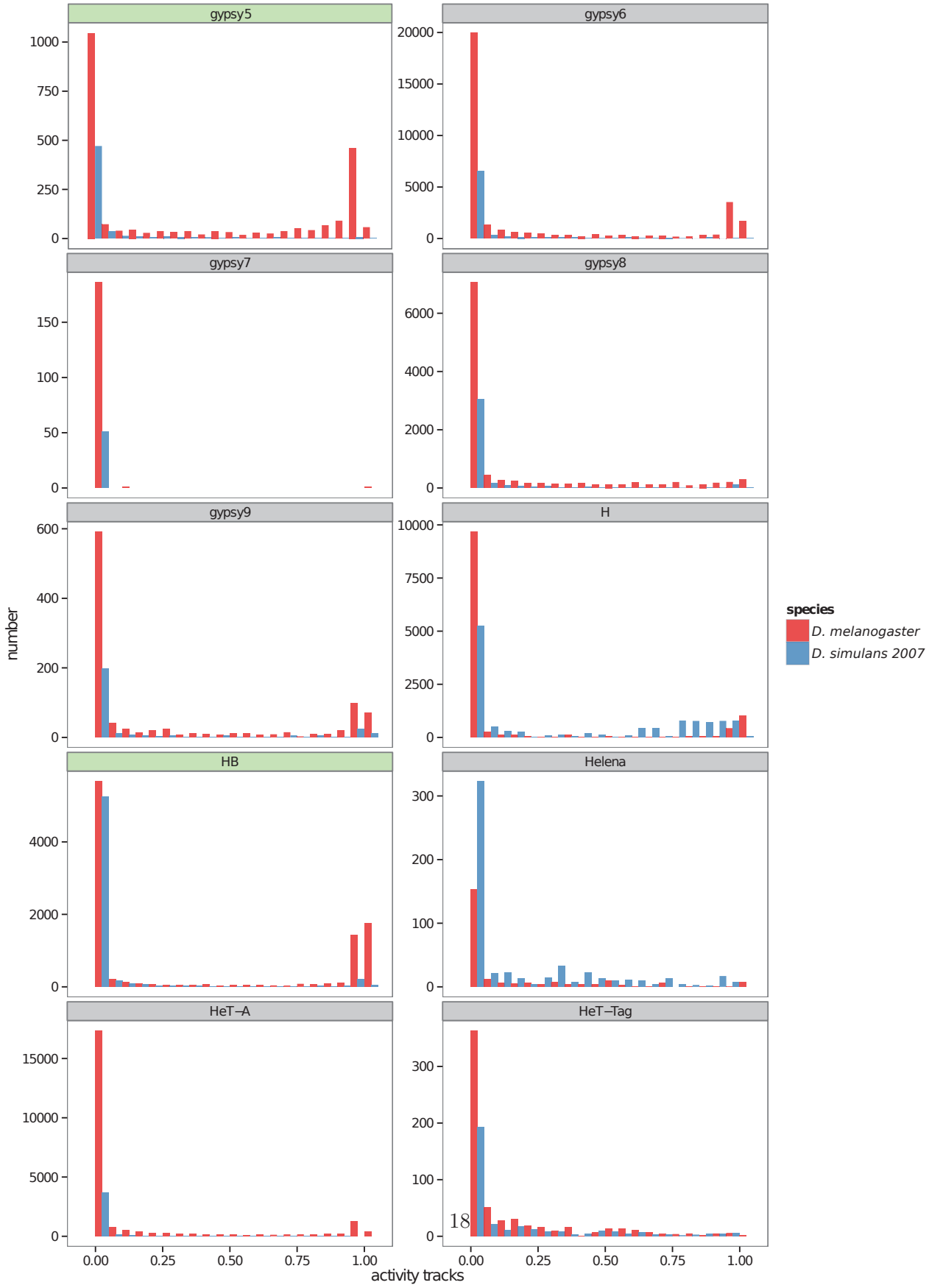

Bari2
blood
BS
BS3
BS4
Burdock
Circe
copia
Copia2
Cr1a
number
activity tracks
species
D. melanogaster
D. simulans 2007

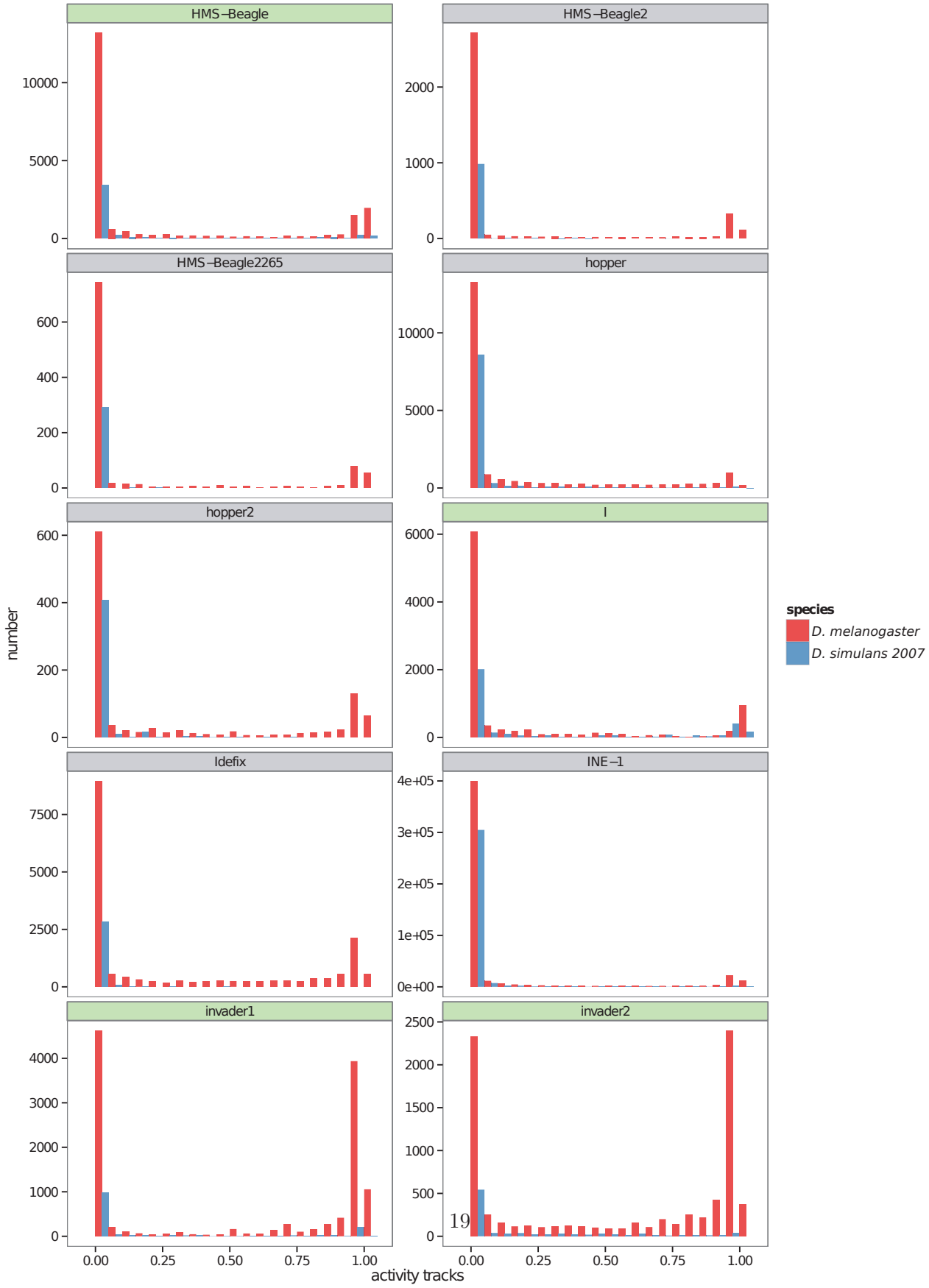

diver
diver2
Doc
Doc2
Doc3
Doc4
F
FB
flea
frogger
number
activity tracks
species
D. melanogaster
D. simulans 2007

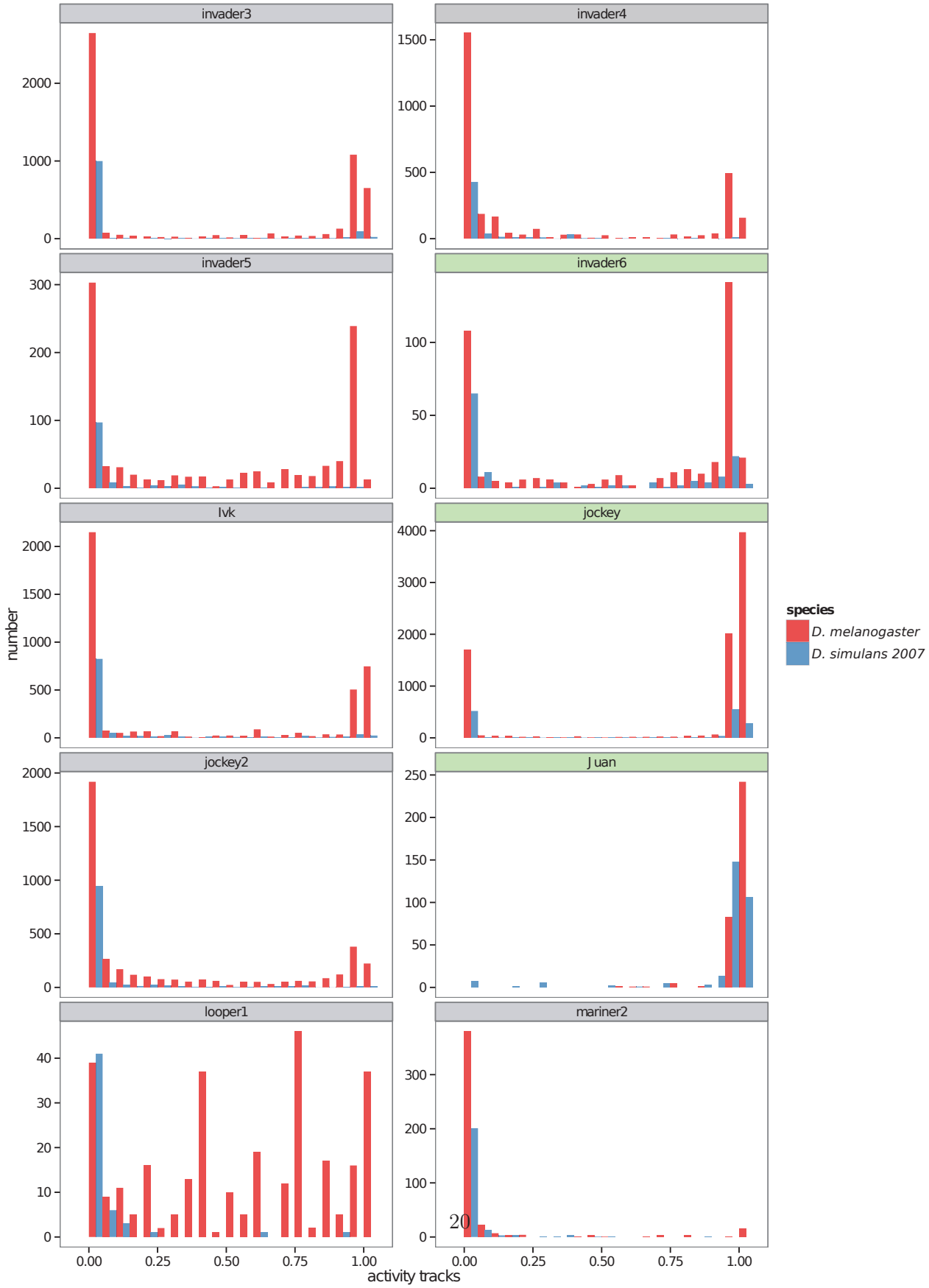
15

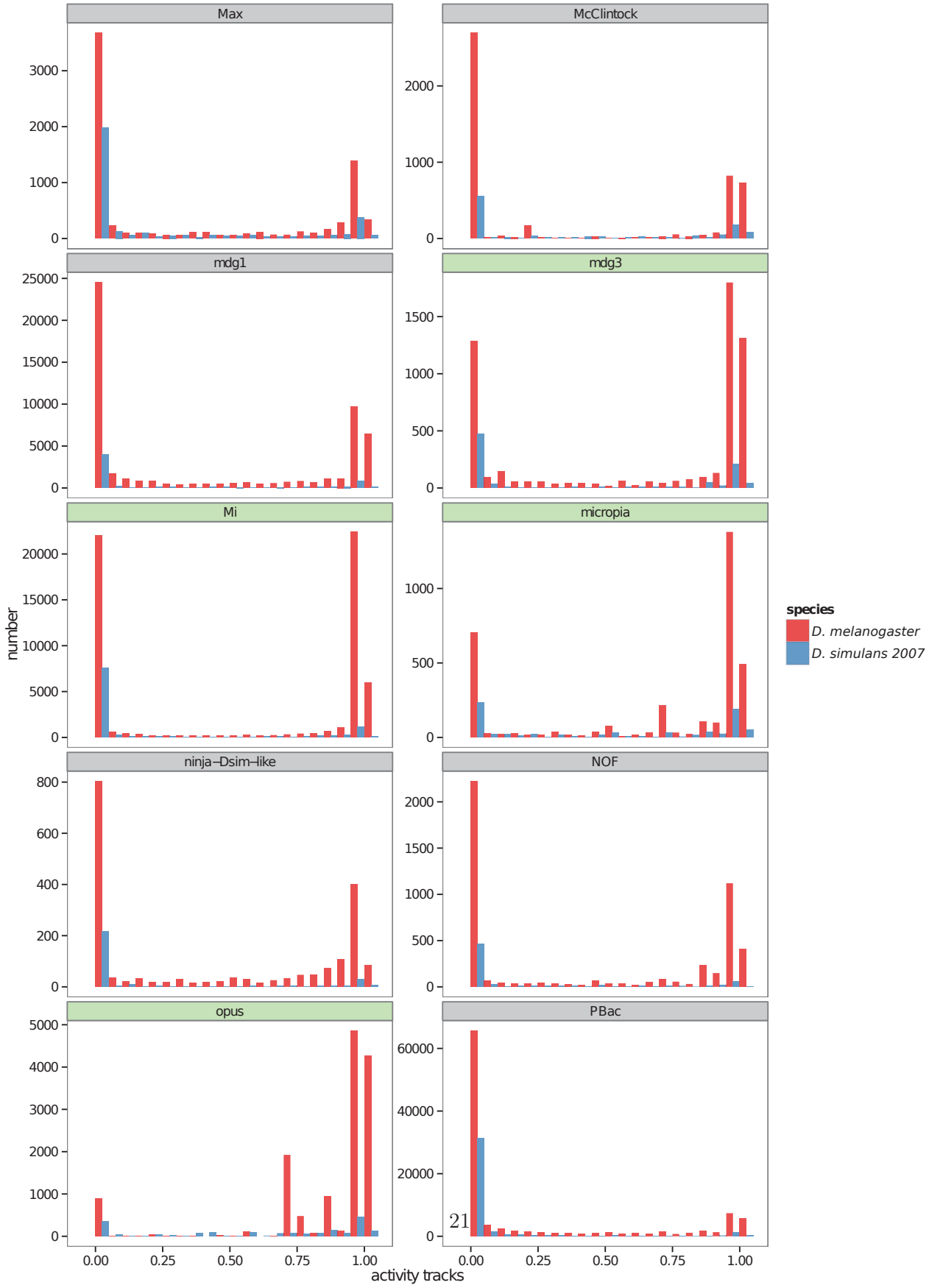

FW
Fw2
Fw3
G
G2
G3
G4
G5
G5A
G6
number
activity tracks
species
D. melanogaster
D. simulans 2007
16

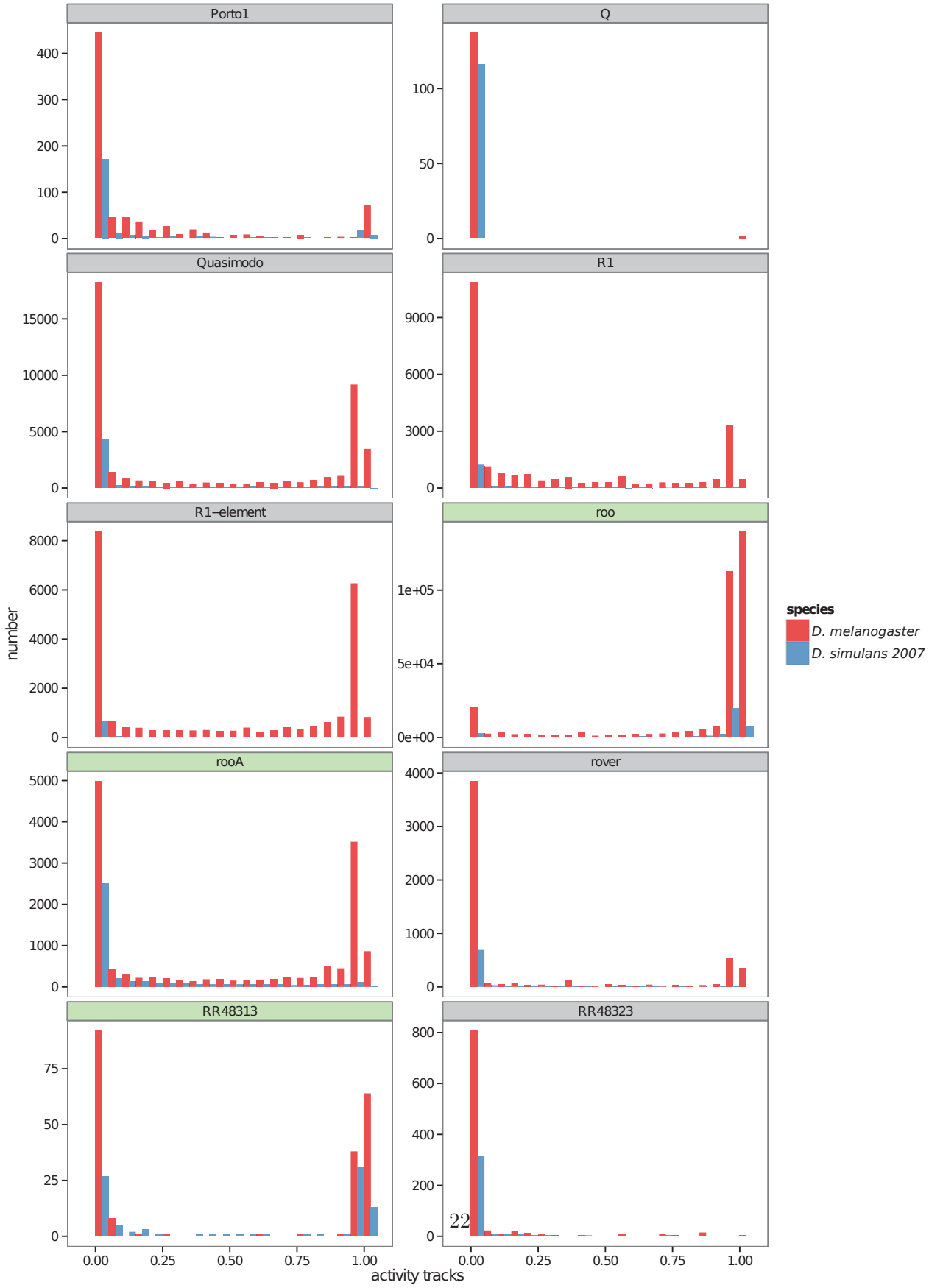

G7
GATE
gtwin
gypsy
gypsy10
gypsy11
gypsy12
gypsy2
gypsy3
gypsy4
number
activity tracks
species
D. melanogaster
D. simulans 2007
17

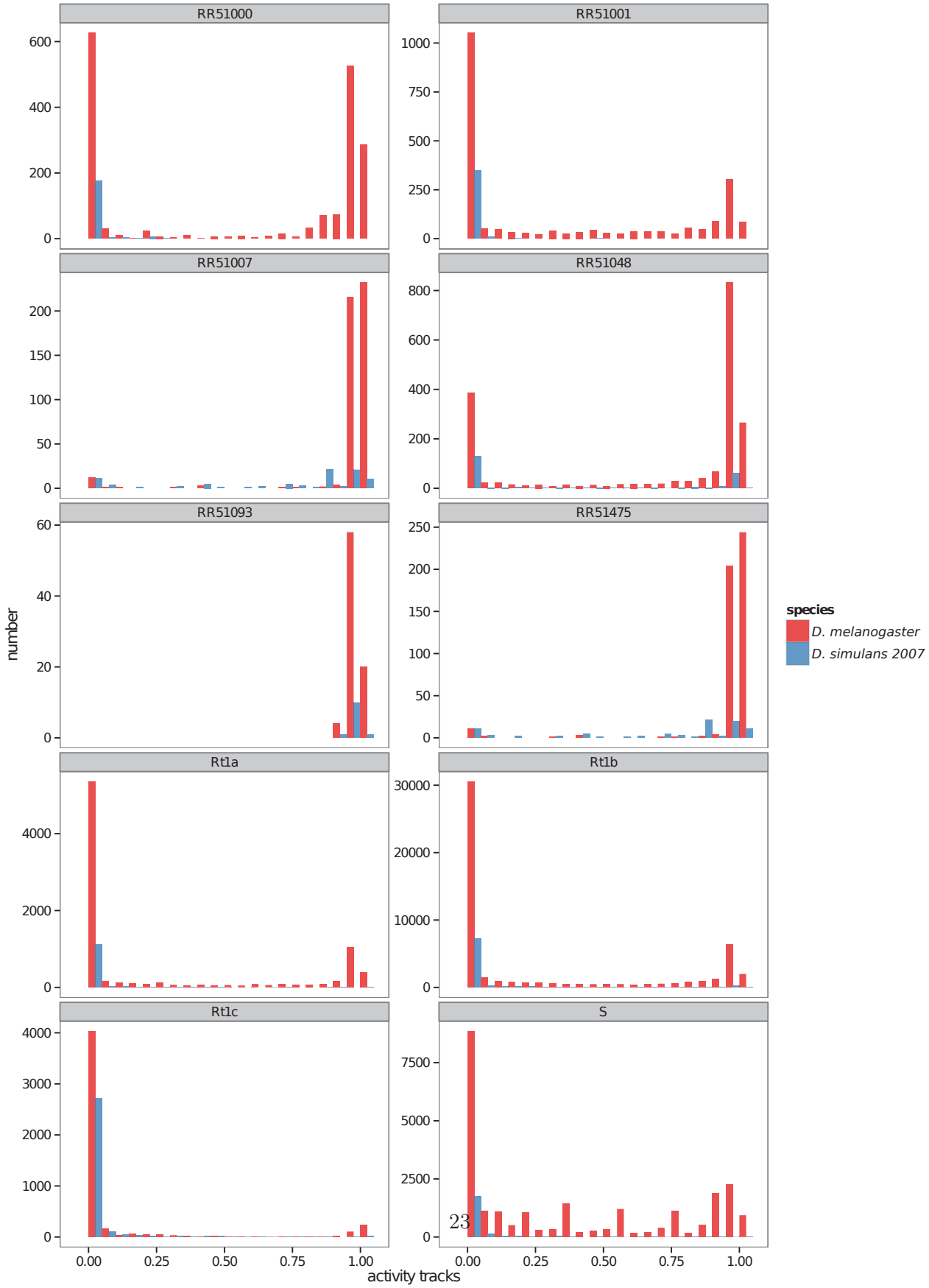

gypsy5
gypsy6
gypsy7
gypsy8
gypsy9
H
HB
Helena
HeT–A
HeT–Tag
number
activity tracks
species
D. melanogaster
D. simulans 2007

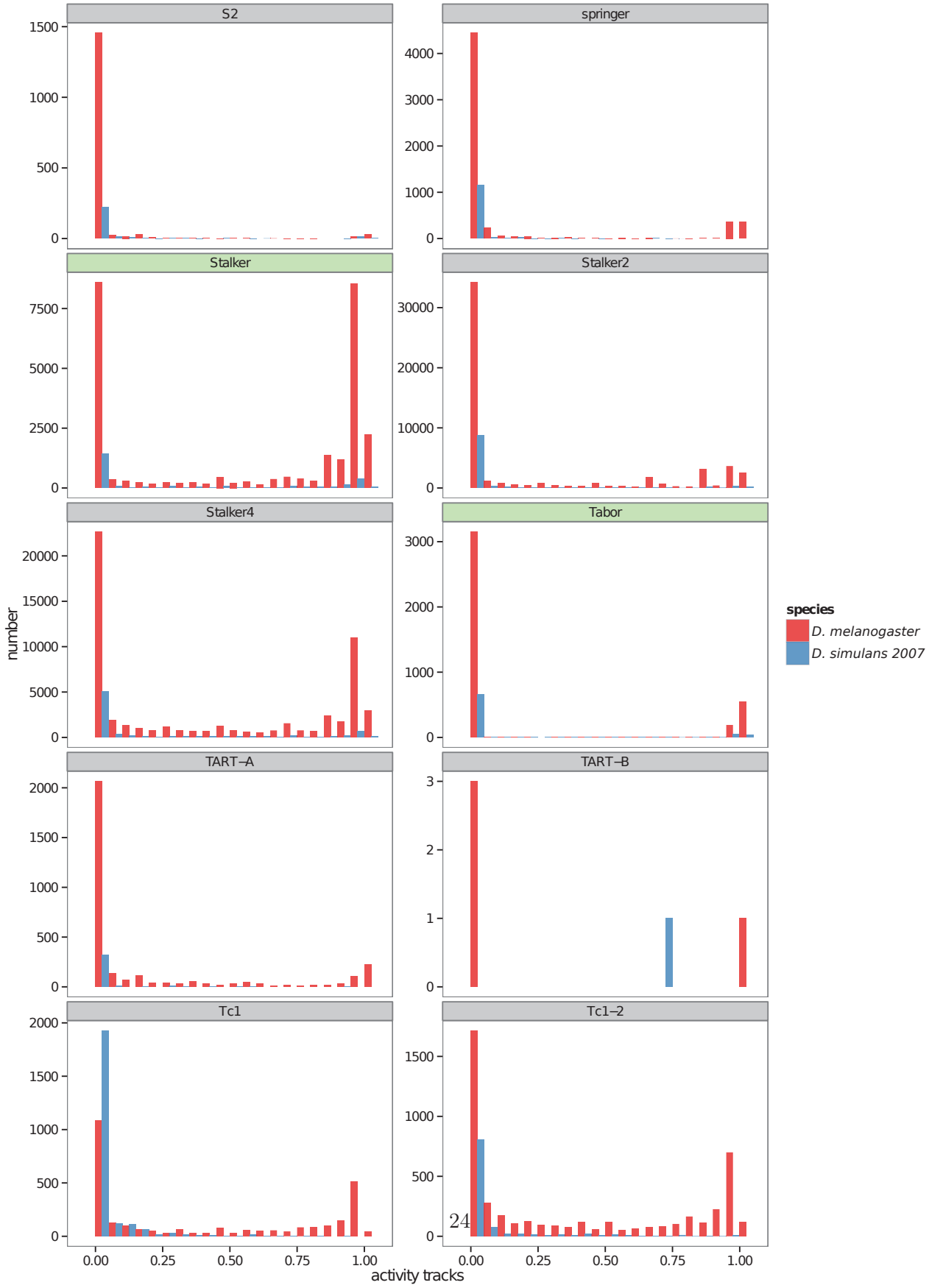

HMS–Beagle
HMS–Beagle2
HMS–Beagle2265
hopper
hopper2
I
Idefix
INE–1
invader1
invader2
number
activity tracks
species
D. melanogaster
D. simulans 2007

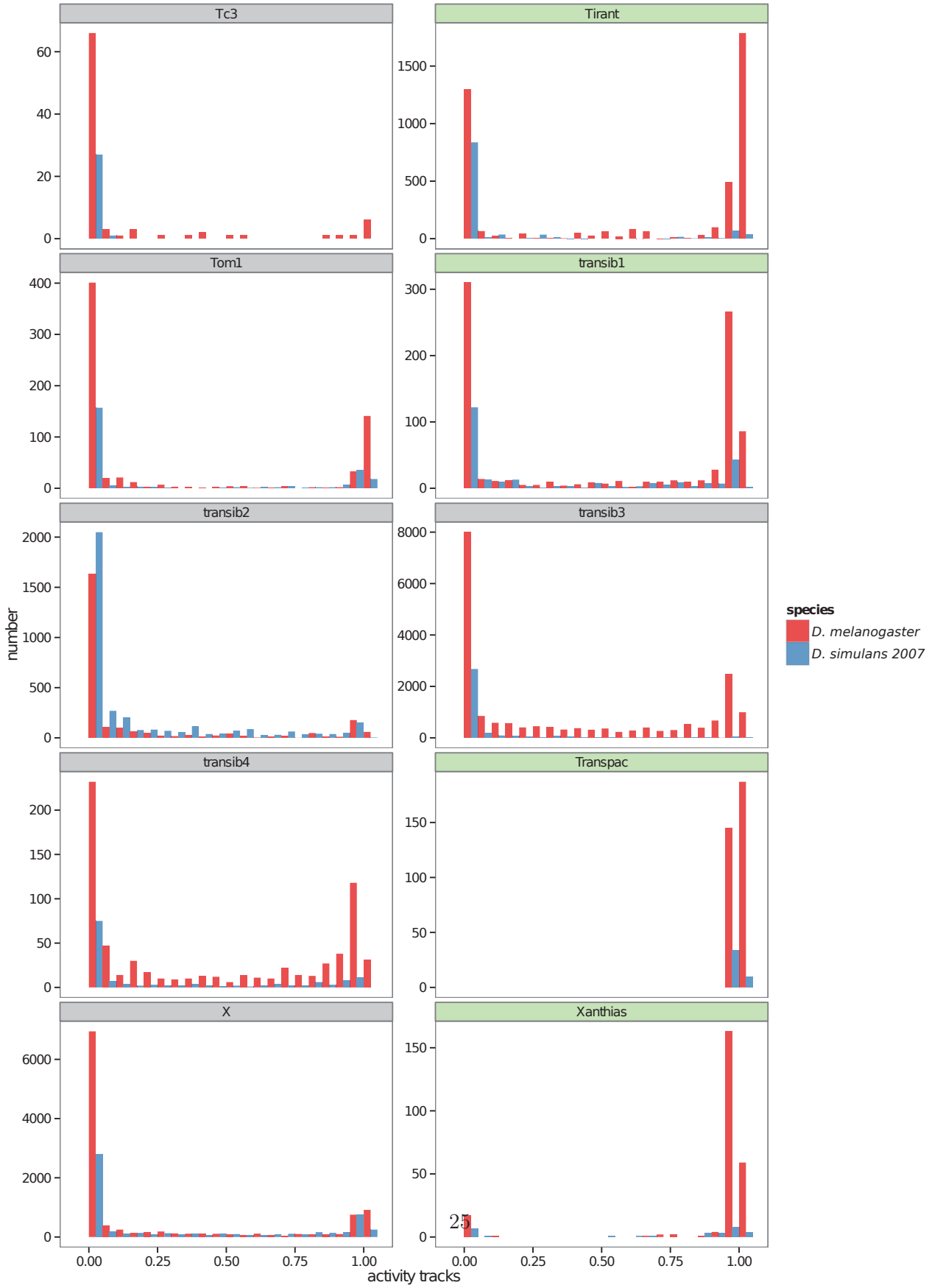

invader3
invader4
invader5
invader6
Ivk
jockey
jockey2
Juan
looper1
mariner2
number
activity tracks
species
D. melanogaster
D. simulans 2007
20


Max
McClintock
mdg1
mdg3
Mi
micropia
ninja–Dsim–like
NOF
opus
PBac
number
activity tracks
species
D. melanogaster
D. simulans 2007


Porto1
Q
Quasimodo
R1
R1−element
roo
rooA
rover
RR48313
RR48323
number
activity tracks
species
D. melanogaster
D. simulans 2007


RR51000
RR51001
RR51007
RR51048
RR51093
RR51475
Rt1a
Rt1b
Rt1c
S
number
activity tracks
species
D. melanogaster
D. simulans 2007


S2
springer
Stalker
Stalker2
Stalker4
Tabor
TART–A
TART–B
Tc1
Tc1–2
number
activity tracks
species
D. melanogaster
D. simulans 2007


Tc3
Tirant
Tom1
transib1
transib2
transib3
transib4
Transpac
X
Xanthias
number
activity tracks
species
D. melanogaster
D. simulans 2007

25

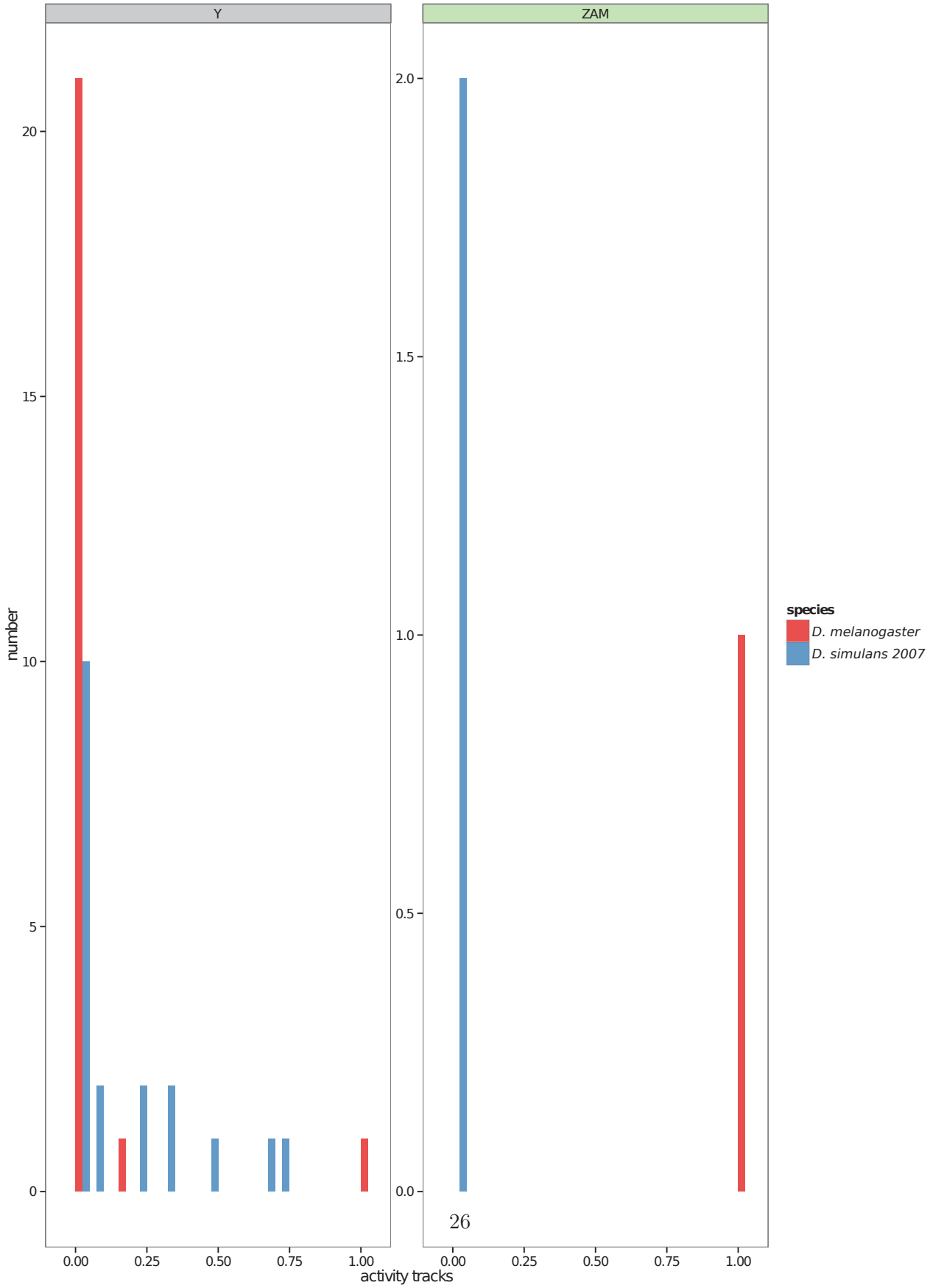
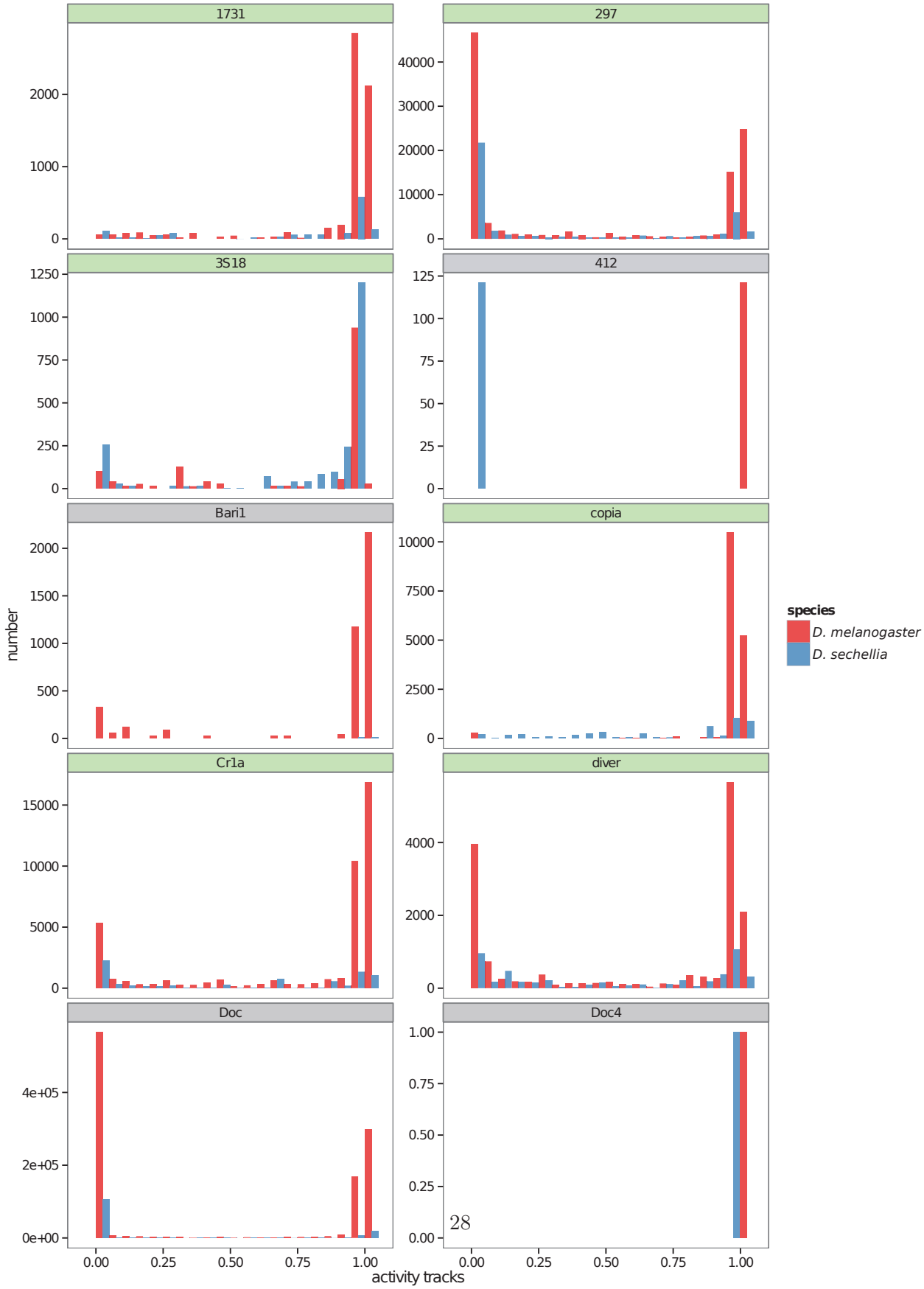
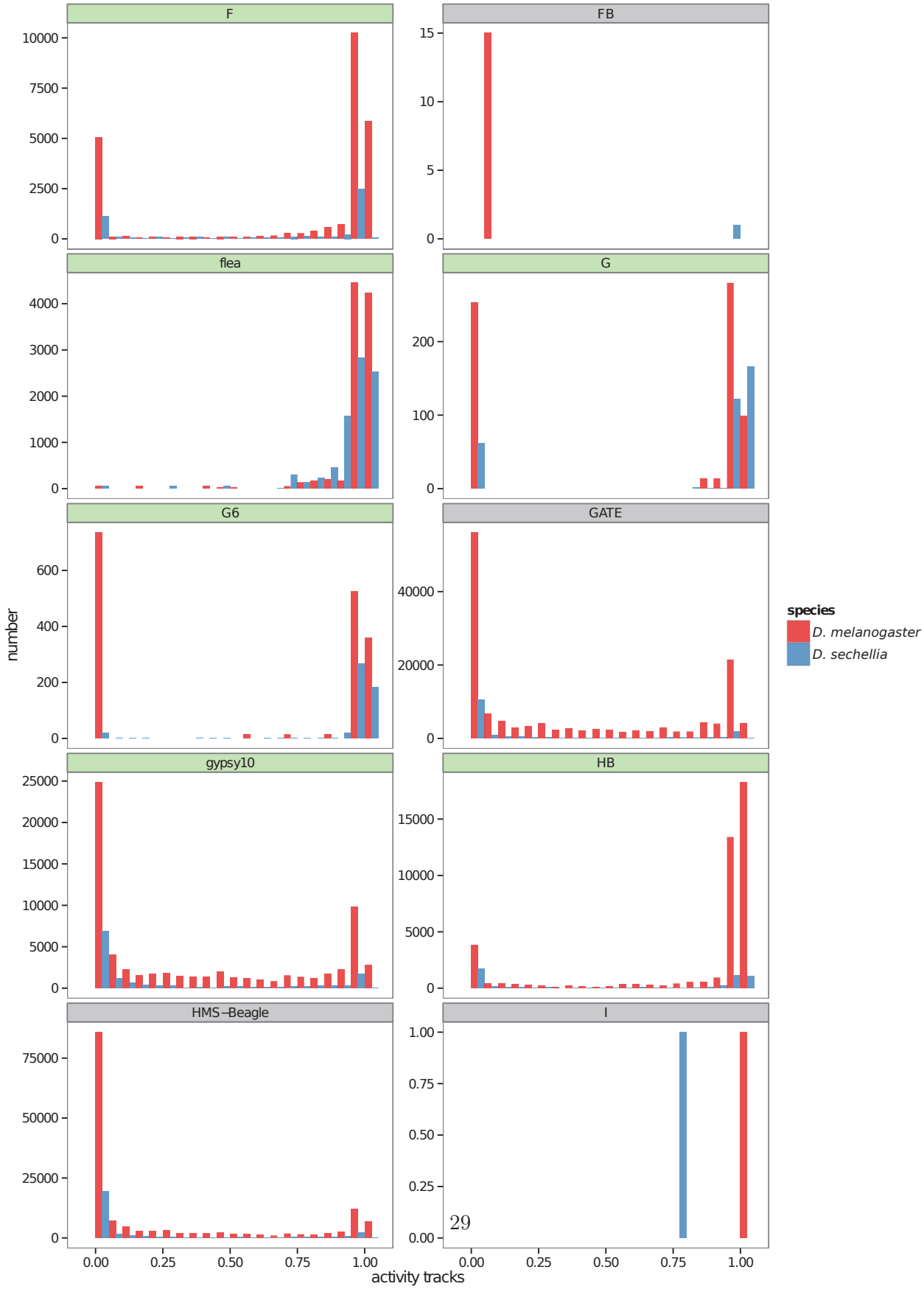
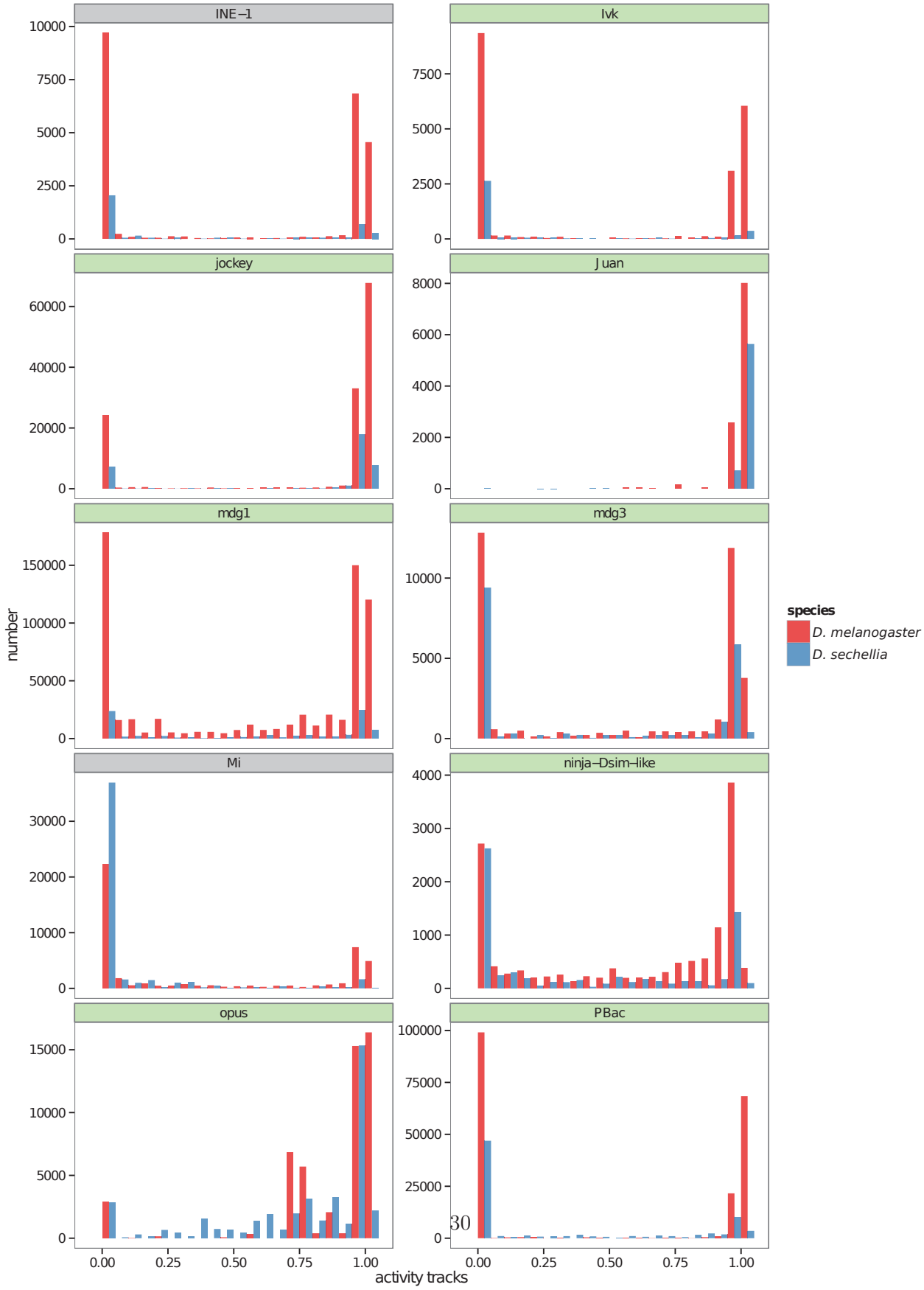
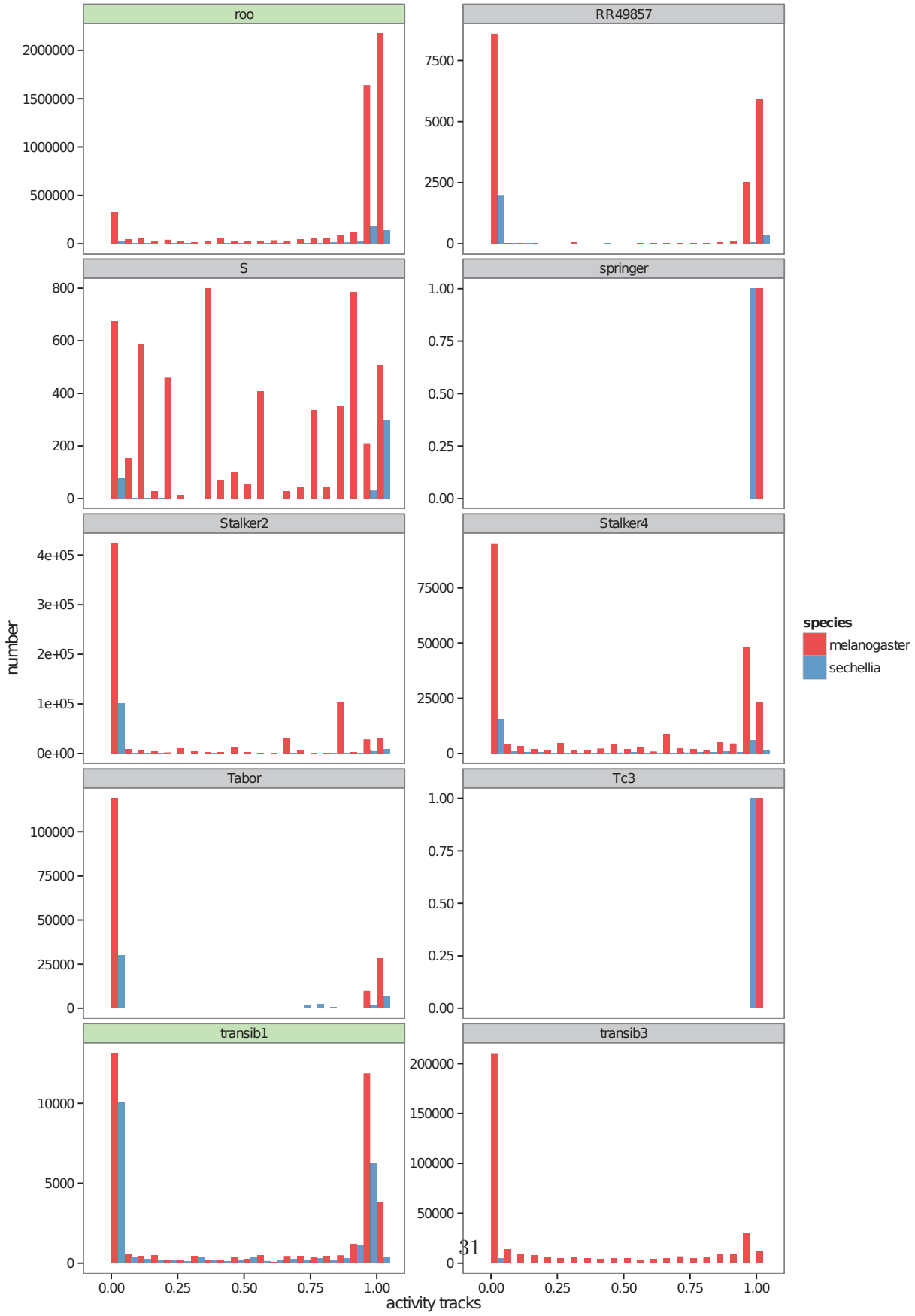

Y
ZAM
number
activity tracks
species
D. melanogaster
D. simulans 2007
26

Figure S4


1731
297
3S18
412
Bari1
copia
Cr1a
diver
Doc
Doc4
number
activity tracks
species
D. melanogaster
D. sechellia
28


F
FB
flea
G
G6
GATE
gypsy10
HB
HMS–Beagle
I
number
activity tracks
species
D. melanogaster
D. sechellia
29


INE−1
Ivk
jockey
Juan
mdg1
mdg3
Mi
ninja−Dsim−like
opus
PBac
number
activity tracks
species
D. melanogaster
D. sechellia


roo
RR49857
S
springer
Stalker2
Stalker4
Tabor
Tc3
transib1
transib3
number
activity tracks
species
melanogaster
sechellia

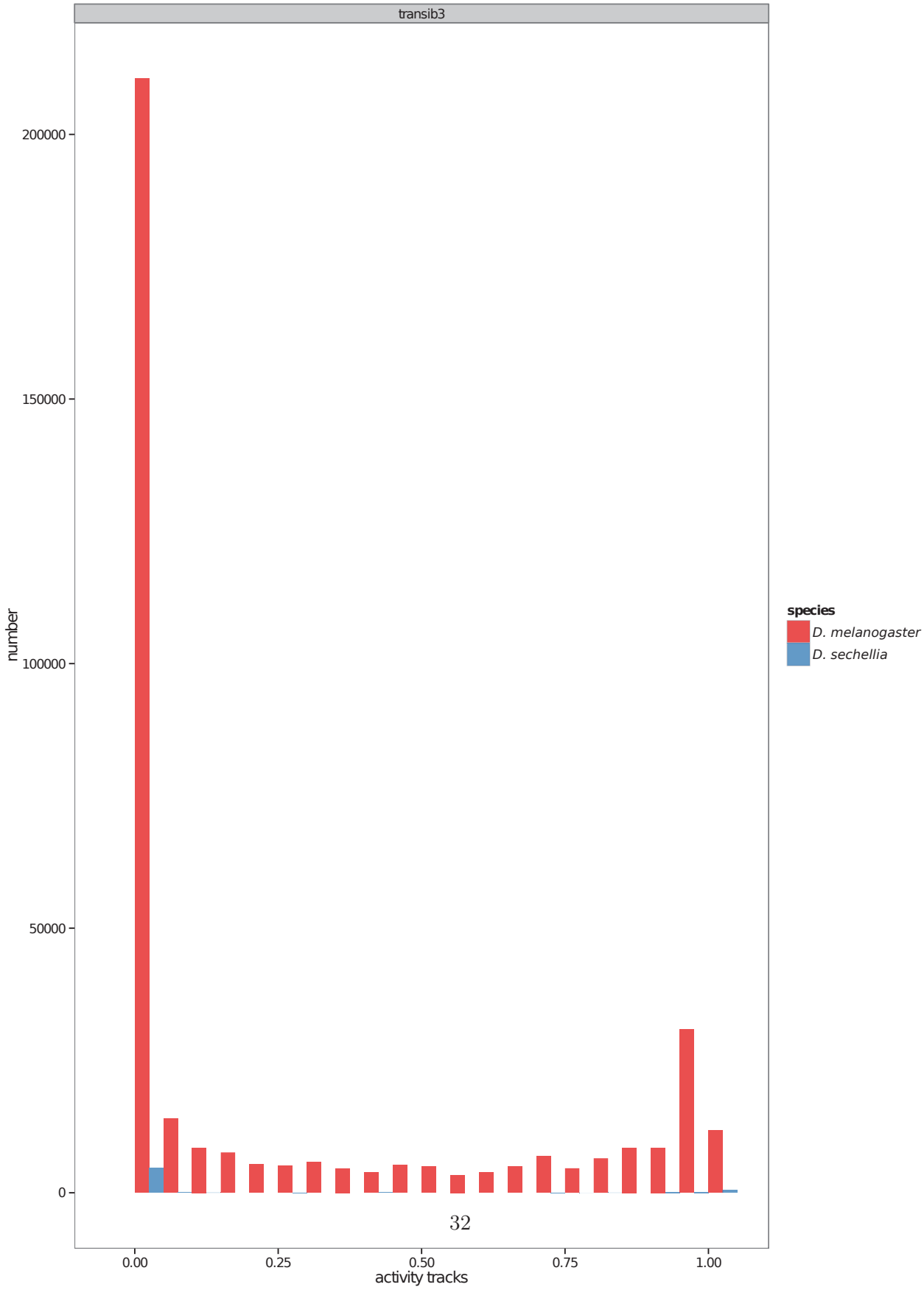
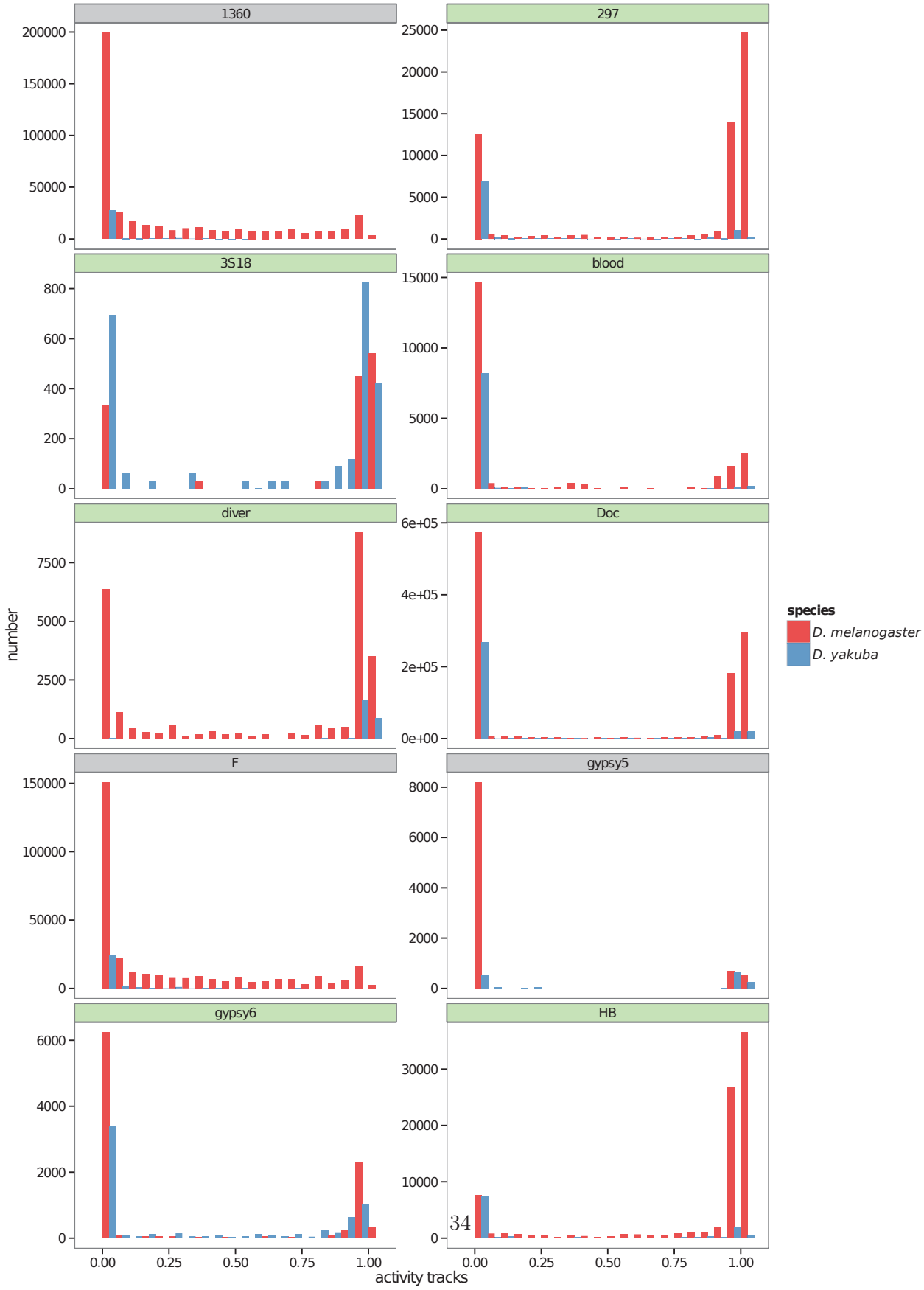
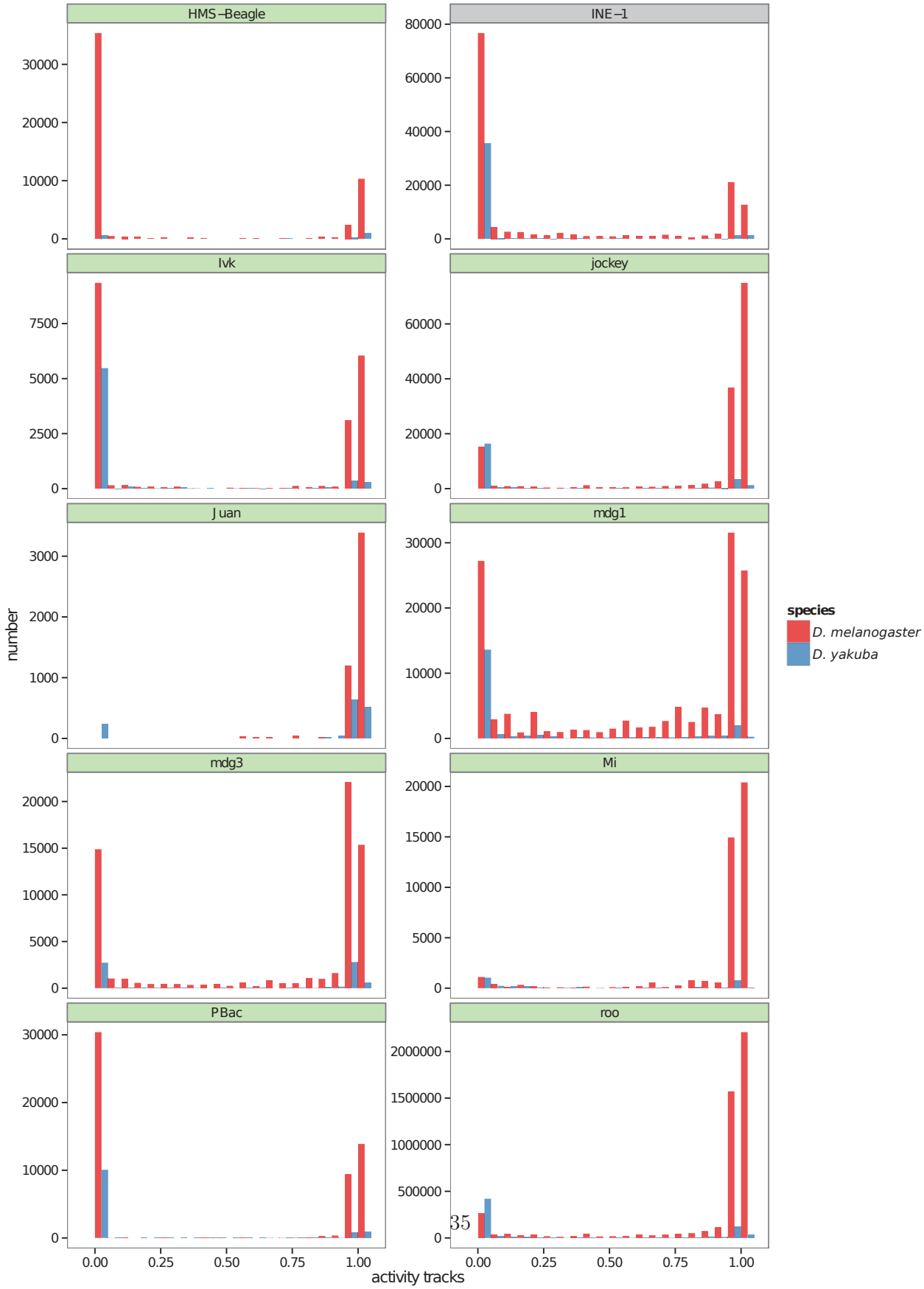

transib3
200000
150000
100000
50000
0
number
0.00
0.25
0.50
0.75
1.00
activity tracks
species
D. melanogaster
D. sechellia

Figure S5


1360
297
3S18
blood
diver
Doc
F
gypsy5
gypsy6
HB
number
activity tracks
0.00
0.25
0.50
0.75
1.00
species
D. melanogaster
D. yakuba


HMS−Beagle
INE−1
Ivk
jockey
Juan
mdg1
mdg3
Mi
PBac
roo
number
activity tracks
species
D. melanogaster
D. yakuba

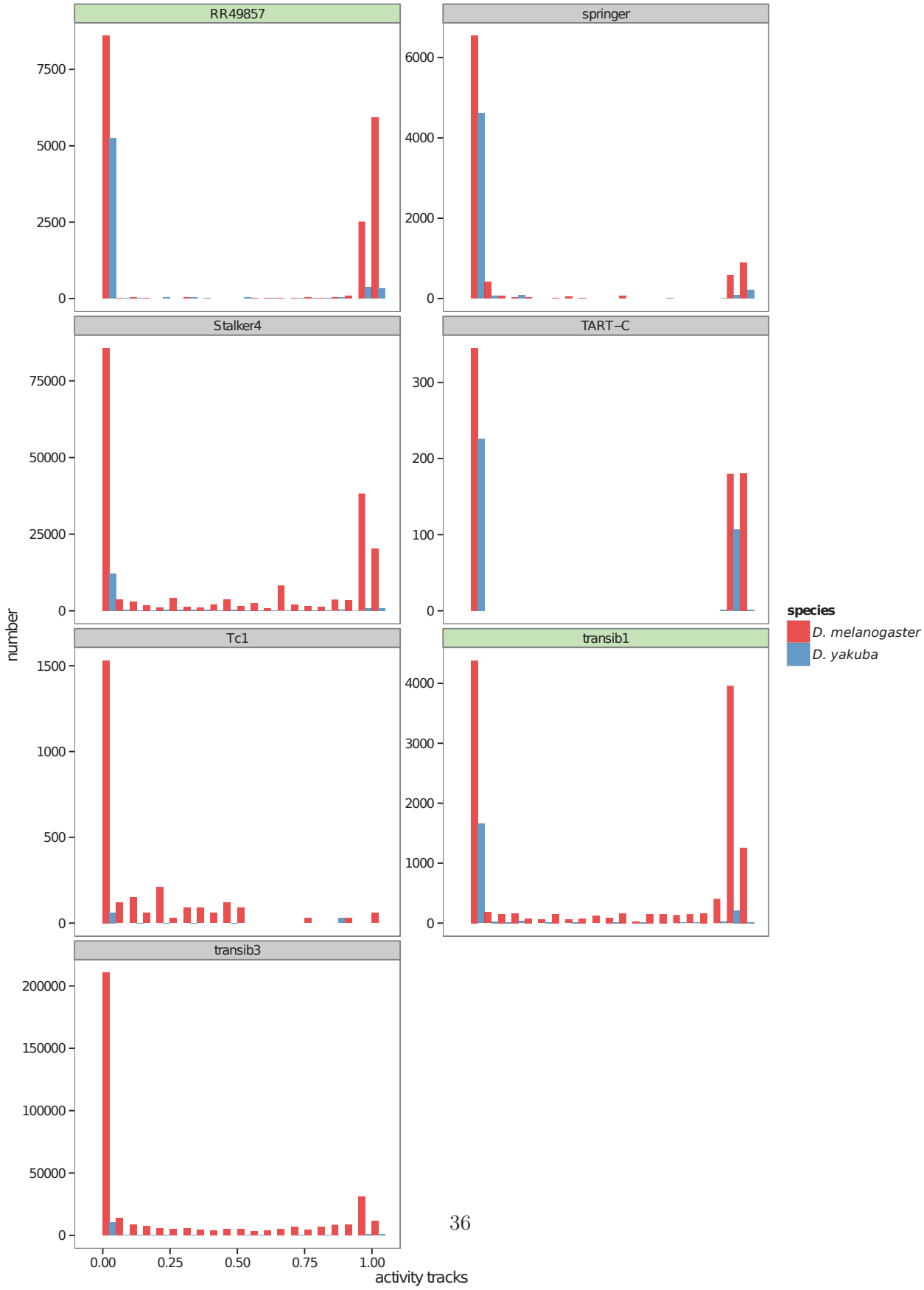

RR49857
springer
Stalker4
TART−C
Tc1
transib1
transib3
number
activity tracks
0.00
0.25
0.50
0.75
1.00
species
D. melanogaster
D. yakuba

Figure S6

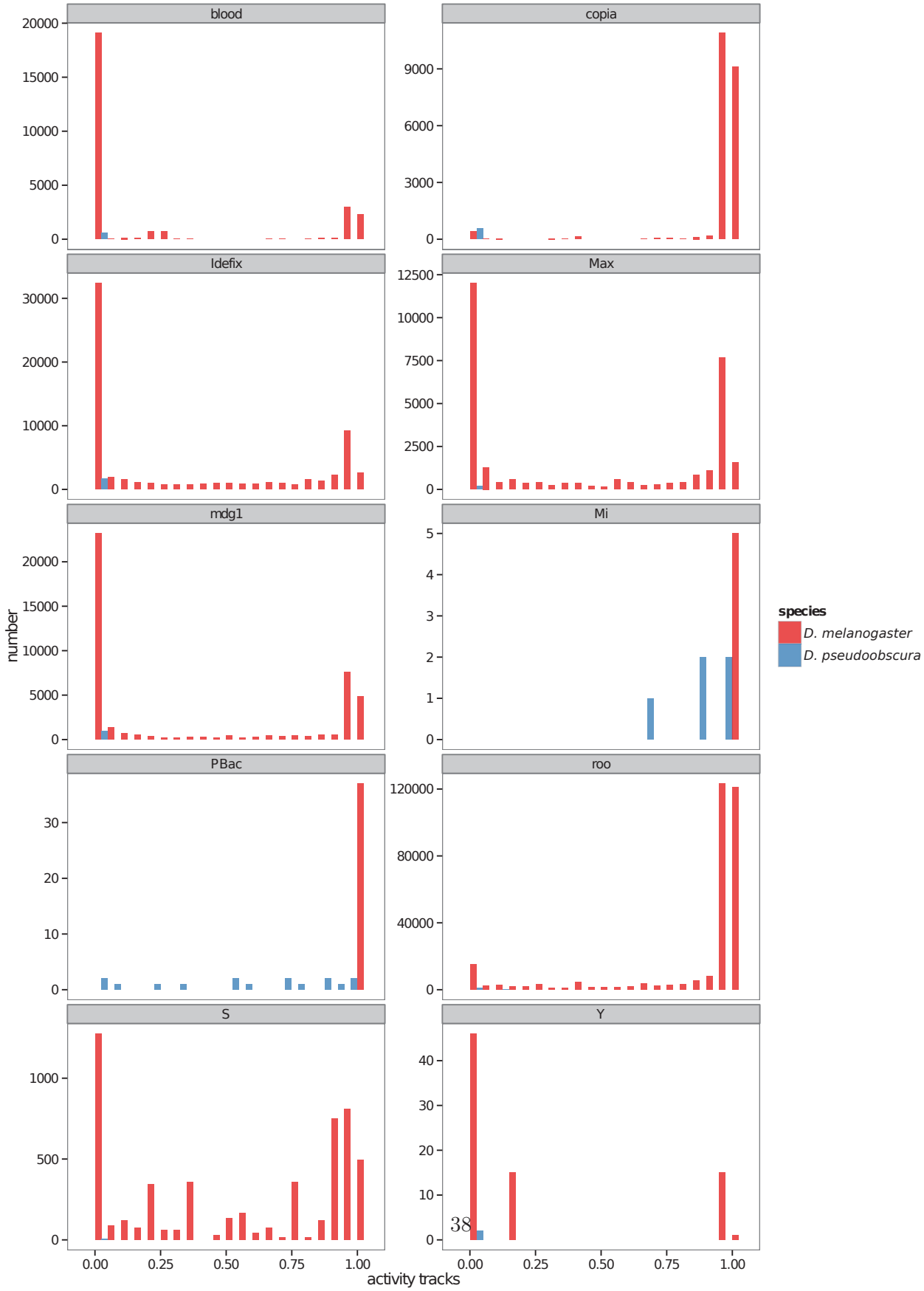
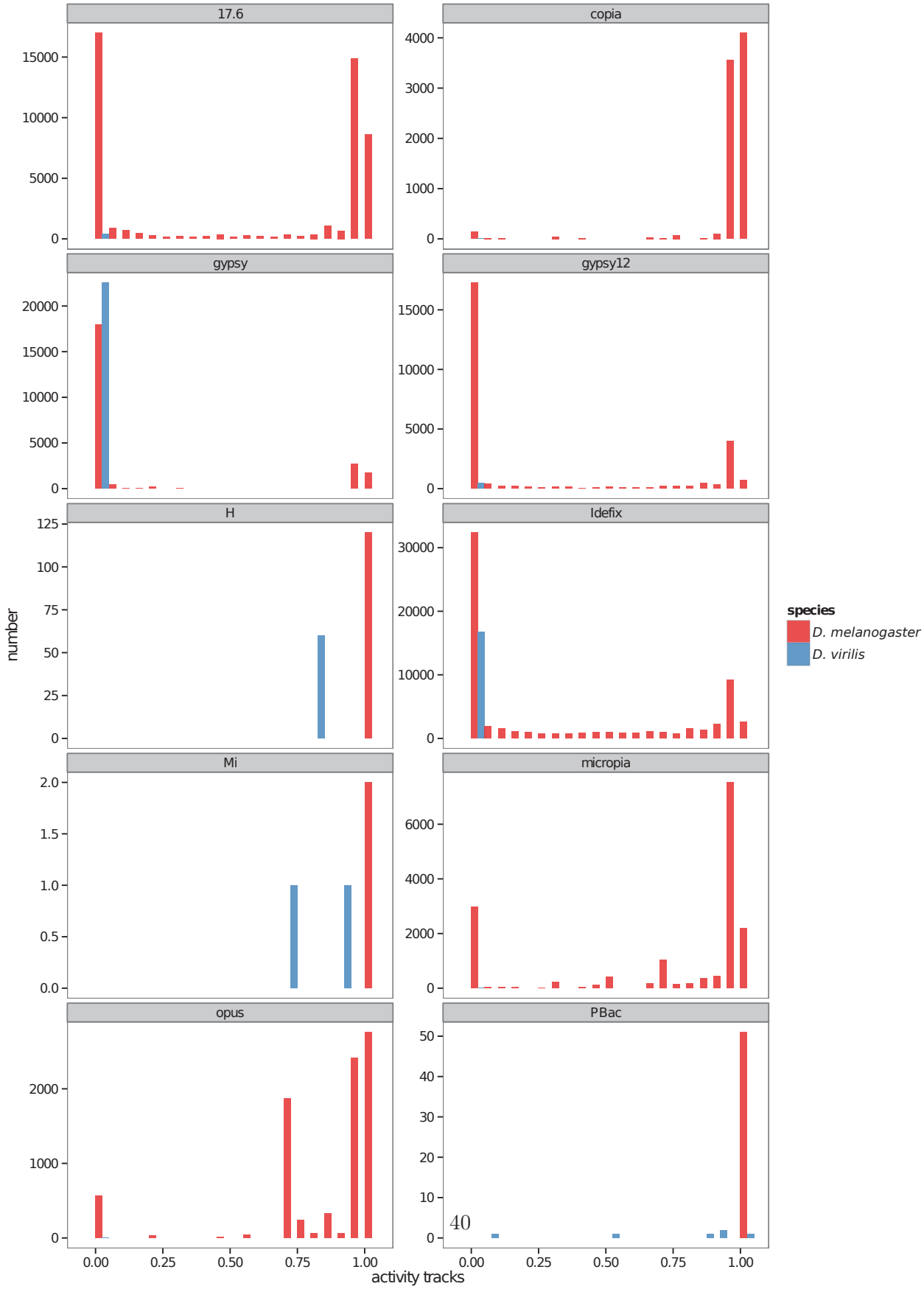

blood
copia
Idefix
Max
mdg1
Mi
PBac
roo
S
Y
number
activity tracks
species
D. melanogaster
D. pseudoobscura

Figure S7


17.6
copia
gypsy
gypsy12
H
Idefix
Mi
micropia
opus
PBac
number
activity tracks
species
D. melanogaster
D. virilis
40

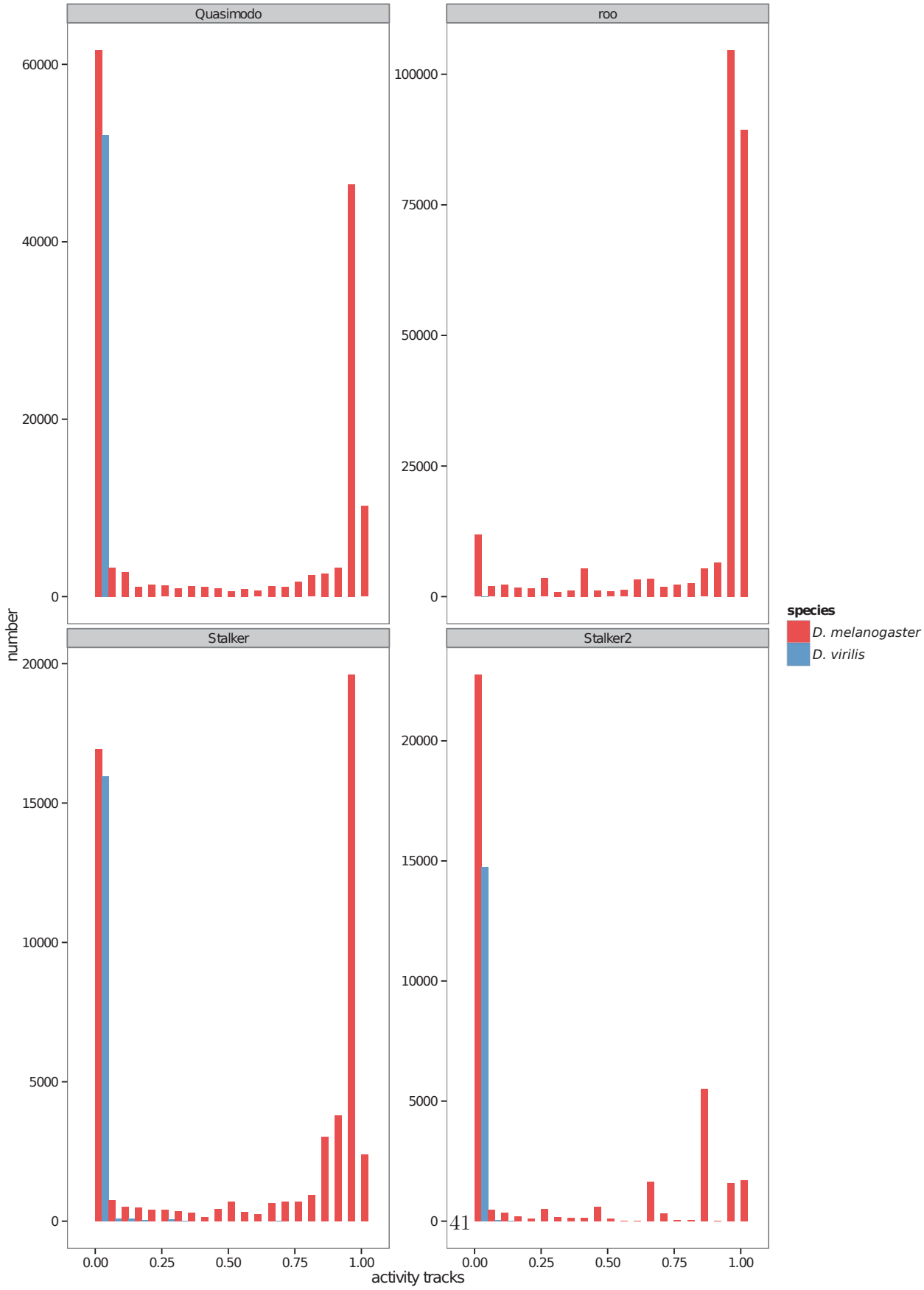

Quasimodo
roo
Stalker
Stalker2
number
activity tracks
species
D. melanogaster
D. virilis
60000
40000
20000
0
100000
75000
50000
25000
15000
10000
5000
0.00
0.25
0.50
0.75
1.00
41

Table S1: Annotation of the intergenic DNA fragments detected with the *D. melangoaster-D. simulans* analysis.

intergenic DNA	before filter (fragments)	after filter (fragments)
CRMs	244	222
DNA motif	0	0
five prime UTR	23725	1338
insertion site	4	2
insulator	338	176
mature peptide	0	0
miRNA	19	11
modified RNA base feature	6	5
ncRNA	512	177
oligonucleotide	10553	10192
origin of replication	1078	774
orthologous region	1018	664
point mutation	14	8
polyA site	4	4
pre miRNA	9	5
protein binding site	123	93
pseudogene	185	13
regulatory region	263	237
repeat region	4086	72
rescue fragment	179	148
RNAi reagent	7343	4580
rRNA	4	0
silencer	62	51
snoRNA	68	22
snRNA	195	0
tandem repeat	38	0
TF binding site	17312	12591
three prime UTR	2078	1904
tRNA	510	3
TSS	2055	1263
	72025	34555

Table S2: List of the transposables elements with an *activity track* consistent with an horizontal transfert between the considered genome and the one of *D. melanogaster*.

	D. simulans 2012	*D. simulans* 2007	*D. sechellia*	*D. yakuba*
Transposable	-	*1731*	*1731*	-
Elements	**-**	***17.6***	**-**	-
	297	***297***	***297***	*297*
(names in	***3S18***	***3S18***	*3S18*	***3S18***
bold were	**-**	***412***	-	-
previously	*accord*	*accord*	-	-
described	**-**	***Bari1***	-	-
in the	***blood***	***blood***	-	*blood*
literature)	*BS*	*BS*	-	-
	-	*Burdock*	-	**-**
	-	***copia***	*copia*	-
	-	*copia2*	-	-
	Cr1a	-	*Cr1a*	-
	-	*diver*	*diver*	***diver***
	-	*diver2*	-	-
	-	*Doc*	-	***Doc***
	-	*Doc2*	-	-
	-	***F***	*F*	-
	flea	***flea***	*flea*	-
	frogger	*frogger*	-	-
	-	*FW*	-	-
	GATE	*GATE*	-	-
	-	-	*G*	-
	-	-	*G6*	-
	-	*gtwin*	**-**	-
	-	***gypsy10***	*gypsy10*	**-**
	-	-	-	*gypsy6*
	gypsy12	-	-	-
	-	*gypsy2*	-	-
	gypsy4	-	-	-
	-	***gypsy5***	-	-
	HB	*HB*	*HB*	*HB*
	HMS-Beagle	***HMS-Beagle***	-	***HMS-Beagle***
	I	***I***	-	-
	-	*invader1*	-	-
	-	*invader2*	-	-
	-	***invader6***	-	**-**
	-	-	*Ivk*	*Ivk*
	jokey	***jokey***	*jokey*	*jokey*
	juan	*juan*	*juan*	*juan*
	mdg1	***mdg1***	*mdg1*	*mdg1*
	-	*mdg3*	*mdg3*	***mdg3***
	-	***Micropia***	**-**	-
	-	-	*ninja-Dsim-like*	-
	-	***opus***	***opus***	-
	PBac	-	*PBac*	*PBac*
	roo	***roo***	*roo*	*roo*
	-	*RR48313 (Max)*	-	-

	D. simulans 2012	*D. simulans* 2007	*D. sechellia*	*D. yakuba*
Transposable	***Stalker2***	***Stalker2***	-	-
Elements	***Tabor***	***Tabor***	-	-
	-	***Tirant***	-	-
	-	*transib1*	*transib1*	*transib1*
	-	-	-	*transib3*
	-	***Transpac***	-	-
	-	*Xanthias*	-	-
	-	***ZAM***	-	-
total	21	46	21	17

2 The unilateral side of multiple-testing : an ℓFDR application

The unilateral side of multiple-testing: an ℓFDR application
Supplementary File 1

Laurent Modolo, Alain Celisse, Emmanuelle Lerat and Franck Picard

October 14, 2014

Contents

1 Zero inflated Gaussian distribution.

Let us consider the ZI-Gaussian (ZIG) distribution $\phi_{\delta_0}(x, \alpha) = \alpha\delta_0(x) + (1-\alpha)\phi(x, 0, 1)$, where $\phi(\bullet, 0, 1)$ denotes the standard Gaussian pdf. The only parameter of this distribution is α which can be estimated using the complete data framework of the EM algorithm. Denoting by $V \sim \mathcal{B}(\alpha)$ the hidden variable such that when V_n equals 0, $X \sim \delta_0(\bullet)$ and when V_n equals 1, $X_n \sim \phi(\bullet, 0, \sigma_0)$. Thus the complete-data log-likelihood of a n-sample of ZI-Gaussian distribution is:

$$\log \mathcal{L}(\alpha; \mathbf{X}, \mathbf{V}) = \sum_{n=1}^{N} I(V_n = 0) \log\left(\alpha\delta_0(x_n)\right) + \sum_{n=1}^{N} I(V_n = 1) \log\left((1-\alpha)\phi(x_n, 0, 1)\right)$$

With $\nu_s(n) = P(V_n = s|\mathbf{X})$ by standard derivation, we have the estimators :

$$\widehat{\alpha} = \frac{1}{N}\sum_{n=1}^{N} \nu_0(n)$$

and the posterior expectation of V is

$$\nu_0(n) = \frac{\alpha\delta_0(x_n)}{\alpha\delta_0(x_n) + (1-\alpha)\phi(x_n, 0, 1)}$$

As $P(X = 0) = 0$ when $X \sim \mathcal{N}(0, 1)$, we can note that we always have :

$$\widehat{\alpha} \quad = \quad \frac{\sharp\{X = 0\}}{\sharp\{X\}}$$

2 Normal Mixture.

Let us consider the Normal Mixture distribution $\phi_L(x, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{c}) = \sum_{\ell=1}^{L} c_\ell \phi(x, \mu_\ell, \sigma_\ell)$, where $\phi(\bullet, \mu_\ell, \sigma_\ell)$ denotes the Gaussian pdf. The only parameters of this distribution are the vectors $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\omega}$ which can be estimated using the complete data framework of the EM algorithm. Denoting by $W_\ell \sim \mathcal{B}(c_\ell)$ the hidden variable such that when $W_{n\ell}$ equals 1, $X_n \sim \phi(x, \mu_\ell, \sigma_\ell)$. Thus the complete-data log-likelihood of a n-sample of Normal Mixture distribution is:

$$\log \mathcal{L}(\mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\sigma}; \mathbf{X}, \mathbf{W}) = \sum_{n=1}^{N} \sum_{\ell=1}^{L} I(W_{n\ell} = 1) \times \log(c_\ell) + I(W_{n\ell} = 1) \times \log(\phi(x, \mu_\ell, \sigma_\ell))$$

With $\omega_{s\ell}(n) = P(W_{n\ell} = s|\mathbf{X})$ and the constraint $\sum_{\ell=1}^{L} c_\ell = 1$, by standard derivation, we have the estimators:

$$\begin{aligned}
\widehat{c_\ell} &= \frac{\sum_{n=1}^{N} \omega_{1\ell}(n)}{N} \\
\widehat{\mu_\ell} &= \frac{\sum_{n=1}^{N} \omega_{1\ell}(n) \times x_n}{\sum_{n=1}^{N} \omega_{1\ell}(n)} \\
\widehat{\sigma_\ell^2} &= \frac{\sum_{n=1}^{N} \omega_{1\ell}(n)(x_n - \mu_\ell)^2}{\sum_{n=1}^{N} \omega_{1\ell}(n)}
\end{aligned}$$

and the posterior expectation of W is

$$\omega_{1\ell}(n) = \frac{c_\ell \phi(x_n, \mu_\ell, \sigma_\ell)}{\sum_{j=1}^{L} c_j \phi(x_n, \mu_j, \sigma_j)}$$

3 Mixture with a ZIG-component.

Let us consider a mixture of 2 distributions such that

$$f(x) = \kappa\phi_{\delta_0}(x,\alpha) + (1-\kappa)\phi_L\left(x,\boldsymbol{\mu},\boldsymbol{\sigma},\mathbf{c}\right)$$

we introduce the indicator variable $S \sim \mathcal{B}(\kappa)$ such that given $\{S_n = 0\}$ the conditional distribution of X_n is $\phi_{\delta_0}(\bullet,\alpha)$. To estimate the parameters of such a mixture, let us consider the complete-data likelihood:

$$\begin{aligned}\mathcal{L}(\kappa,\alpha,\mathbf{c},\boldsymbol{\mu},\boldsymbol{\sigma};\mathbf{X},\mathbf{S},\mathbf{V},\mathbf{W}) &= \prod_{n=1}^{N}\kappa^{I(S_n=0)}\alpha^{I(S_n=0,V_n=0)}(1-\alpha)^{I(V_n=1,S_n=0)}\phi(x_i,0,1)^{I(V_n=1,S_n=0)}\\ &\times \prod_{n=1}^{N}(1-\kappa)^{I(S_n=1)}\left(\prod_{\ell=1}^{L}\left(c_\ell\phi(x_n,\mu_\ell,\sigma_\ell)\right)^{I(W_{n\ell}=1,S_n=1)}\right)\end{aligned}$$

with $(\alpha\delta_0(x_n))^{I(S_n=0,V_n=0)} = \alpha^{I(S_n=0,V_n=0)}$ as $\nu_0(n) = 1$ if $x_n = 0$
and the constraint $\sum_{\ell=1}^{L} c_\ell = 1$

$$\begin{aligned}\log\mathcal{L}(\kappa,\alpha,\mathbf{c},\boldsymbol{\mu},\boldsymbol{\sigma};\mathbf{X},\mathbf{S},\mathbf{V},\mathbf{W}) &= \sum_{n=1}^{N} I(S_n=0)\log\kappa + \sum_{n=1}^{N} I(V_n=0,S_n=0)\log\alpha\\ &+ \sum_{n=1}^{N} I(V_n=1,S_n=0)\log(1-\alpha)\\ &+ \sum_{n=1}^{N} I(V_n=1,S_n=0)\log\phi(x_n,0,1) + \sum_{n=1}^{N} I(S_n=1)\log(1-\kappa)\\ &+ \sum_{n=1}^{N}\sum_{\ell=1}^{L} I(S_n=1,W_{n\ell}=1)\log\left(c_\ell\phi(x_n,\mu_\ell,\sigma_\ell)\right)\end{aligned}$$

The expected log-likelihood :

$$
\begin{aligned}
\mathbb{E}\left\{\log \mathcal{L}(\kappa, \alpha, \mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\sigma}; \mathbf{X}, \mathbf{S}, \mathbf{V}, \mathbf{W}) | \mathbf{X}\right\} &= \sum_{n=1}^{N} \mathbb{E}\left\{I(S_n = 0)|x_n\right\} \log \kappa \\
&+ \sum_{n=1}^{N} \mathbb{E}\left\{I(V_n = 0, S_n = 0)|x_n\right\} \log \alpha \\
&+ \sum_{n=1}^{N} \mathbb{E}\left\{I(V_n = 1, S_n = 0)|x_n\right\} \log(1-\alpha) \\
&+ \sum_{n=1}^{N} \mathbb{E}\left\{I(V_n = 1, S_n = 0)|x_n\right\} \log \phi(x_i, 0, 1) \\
&+ \sum_{n=1}^{N} \mathbb{E}\left\{I(S_n = 1)|x_n\right\} \log(1-\kappa) \\
&+ \sum_{n=1}^{N} \sum_{\ell=1}^{L} \mathbb{E}\left\{I(S_n = 1, W_{n\ell} = 1)|x_n\right\} \log\left(c_\ell \phi(x_n, \mu_\ell, \sigma_\ell)\right)
\end{aligned}
$$

3.1 E-Step.

$$
\begin{aligned}
\gamma_s(n) &= \mathbb{E}\left\{I(S_n = s)|\mathbf{X}\right\} = P\left(S_n = s|\mathbf{X}\right) \\
&= \frac{P\left(S_n = s\right) f(x_n|S_n = s)}{\sum_{k=0}^{1} P\left(S_n = k\right) f(x_n|S_n = k)} \\
\gamma_0(n) &= \frac{\kappa \phi_{\delta_0}(x_n, \alpha)}{\kappa \phi_{\delta_0}(x_n, \alpha) + (1-\kappa)\phi_L\left(x_n, \mu_{1:L}, \sigma^2_{1:L}, c_{1:L}\right)}
\end{aligned}
$$

$$
\begin{aligned}
\gamma_0(n)\nu_s(n) &= \mathbb{E}\left\{I(S_n = 0, V_n = s|\mathbf{X}\right\} = P\left(S_n = 0, V_n = s|\mathbf{X}\right) \\
&= \frac{P\left(S_n = 0, V_n = s\right) f(x_n|S_n = 0, V_n = s)}{P\left(S_n = 0\right) f(x_n|S_n = 0)} \\
&= \frac{P\left(S_n = 0|x_n\right) P\left(V_n = s|S_n = 0\right) f(x_n|S_n = 0, V_n = s)}{P\left(S_n = 0\right) f(x_n|S_n = 0)} \\
\gamma_0(n)\nu_0(n) &= \gamma_0(n) \times \frac{\kappa\alpha\delta_0(x_n)}{\kappa\alpha\delta_0(x_n) + \kappa(1-\alpha)\phi(x_n, 0, 1)} \\
&= \gamma_0(n) \times \frac{\alpha\delta_0(x_n)}{\alpha\delta_0(x_n) + (1-\alpha)\phi(x_i, 0, 1)}
\end{aligned}
$$

$$
\begin{aligned}
\gamma_1(n)\omega_{s\ell}(n) &= \mathbb{E}\left\{I(S_n = 1, W_{n\ell} = s)|\mathbf{X}\right\} = P\left(S_n = 1, W_{n\ell} = s|\mathbf{X}\right) \\
&= \frac{P\left(S_n = 1, W_{n\ell} = s\right) f(x_n|S_n = 1, W_{n\ell} = S)}{P\left(S_n = 1\right) f(x_n|S_n = 1)} \\
&= \frac{P\left(S_n = 1|\mathbf{X}\right) P\left(W_{n\ell} = s|S_n = 1\right) f(x_n|S_n = 1, W_{n\ell} = s)}{P\left(S_n = 0\right) f(x_n|S_n = 0)} \\
\gamma_1(n)\omega_{1\ell}(n) &= \gamma_1(n) \times \frac{(1-\kappa)c_\ell \phi(x_n, \mu_\ell, \sigma_\ell)}{(1-\kappa)\sum_{j=1}^{L} c_j \phi(x_n, \mu_j, \sigma_j)} \\
&= \gamma_1(n) \times \frac{c_\ell \phi(x_n, \mu_\ell, \sigma_\ell)}{\sum_{j=1}^{L} c_j \phi(x_n, \mu_j, \sigma_j)}
\end{aligned}
$$

3.2 M-Step.

$$
\begin{aligned}
\widehat{\kappa} &= \frac{\sum_{n=1}^{N} \gamma_0(n)}{N} \\
\widehat{\alpha} &= \frac{\sum_{n=1}^{N} \gamma_0(n)\nu_0(n)}{\sum_{n=1}^{N} \gamma_0(n)} \\
\widehat{\mu_\ell} &= \frac{\sum_{n=1}^{N} \gamma_1(n) \times \omega_{1\ell}(n) \times x_n}{\sum_{n=1}^{N} \gamma_1(n) \times \omega_{1\ell}(n)} \\
\widehat{\sigma_\ell^2} &= \frac{\sum_{n=1}^{N} \gamma_1(n) \times \omega_{1\ell}(n) \times (x_n - \widehat{\mu_\ell})^2}{\sum_{n=1}^{N} \gamma_1(n) \times \omega_{1\ell}(n)} \\
\widehat{c_\ell} &= \frac{\sum_{n=1}^{N} \gamma_1(n) \times \omega_{1\ell}(n)}{\sum_{n=1}^{N} \gamma_1(n)}
\end{aligned}
$$

4 Mixture with a ZIG-componen t and an HMM.

Let us consider a mixture of 2 distributions such that

$$f(x) = \kappa\phi_{\delta_0}(x,\alpha) + (1-\kappa)\phi_L\left(x,\boldsymbol{\mu},\boldsymbol{\sigma},\mathbf{c}\right)$$

we introduce the indicator variable $S_n \in \{0,1\}$, with S following a Markov chain of order 1 with κ the emission probability, A the transition matrix, with generator $a_{ij} = P\left(S_{n+1} = j | S_n = i\right)$ the transition matrix, and $f(x)$ the emission probability. Such that given $\{S_n = 0\}$ the conditional distribution of X_n is $\phi_{\delta_0}(\bullet,\alpha)$ and that given $\{S_n = 1\}$ the conditional distribution of X_n is $\phi_L\left(\bullet,\boldsymbol{\mu},\boldsymbol{\sigma},\mathbf{c}\right)$.

$$P\left(\mathbf{X},\mathbf{S}\right) = \prod_{i=0}^{1}\left[P\left(S_1 = i\right)P\left(x_1|S_1 = i\right)\right]\prod_{n=2}^{N}\prod_{i=0}^{1}\prod_{j=0}^{1}\left[P\left(S_n = j|S_{n-1} = i\right)P\left(x_n|S_n = j\right)\right]$$

With

$$\begin{aligned} P\left(x_n|S_n = 0\right) &= \phi_{\delta_0}(x_n,\alpha) \\ P\left(x_n|S_n = 1\right) &= \phi_L\left(x_n,\boldsymbol{\mu},\boldsymbol{\sigma},\mathbf{c}\right) \end{aligned}$$

And the constraints:

$$\begin{aligned} \sum_{i=0}^{1}\pi_i &= 1 \\ \sum_{j=0}^{1}a_{ij} &= 1, \text{ for } i \in \{0,1\} \\ \sum_{\ell=1}^{L}c_\ell &= 1 \end{aligned}$$

To estimate the parameters of such a mixture, let us consider the complete-data likelihood:

$$\begin{aligned}
\mathcal{L}(\kappa, \alpha, \mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\sigma}; \mathbf{X}, \mathbf{S}, \mathbf{V}, \mathbf{W}) &= \prod_{i=0}^{1} \left[P\left(S_1 = i\right) P\left(x_1 | S_1 = i\right)\right]^{I[S_1=i]} \\
&\times \prod_{n=2}^{N} \prod_{i=0}^{1} \prod_{j=0}^{1} \left[P\left(S_n = j | S_{n-1} = i\right) P\left(x_n | S_n = j\right)\right]^{I[S_{n-1}=i, S_n=j]} \\
&= \left(P\left(S_n = 0\right) P\left(x_1 | S_1 = 0\right)\right)^{I(S_1=0)} \\
&\times \left(P\left(S_n = 1\right) P\left(x_1 | S_1 = 1\right)\right)^{I(S_1=1)} \\
&\times \prod_{n=2}^{N} \left(P\left(S_n = 0 | S_{n-1} = 1\right)^{I(S_{n-1}=1)} P\left(S_n = 0 | S_{n-1} = 0\right)^{I(S_{n-1}=0)}\right)^{I(S_n=0)} \\
&\times \prod_{n=2}^{N} \left(P\left(S_n = 1 | S_{n-1} = 1\right)^{I(S_{n-1}=1)} P\left(S_n = 1 | S_{n-1} = 0\right)^{I(S_{n-1}=0)}\right)^{I(S_n=1)} \\
&\times \prod_{n=2}^{N} P\left(x_n | S_n = 1\right)^{I(S_n=1)} P\left(x_n | S_n = 0\right)^{I(S_n=0)}
\end{aligned}$$

$$\begin{aligned}
\log \mathcal{L}(\kappa, \alpha, \mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\sigma}; \mathbf{X}, \mathbf{S}, \mathbf{V}, \mathbf{W}) &= I(S_1 = 0) log(\kappa) + I(S_1 = 1) \log(1 - \kappa) \\
&+ \sum_{n=1}^{N} I(S_n = 0) \log \phi_{\delta_0}(x_n, \alpha) \\
&+ \sum_{n=1}^{N} I(S_n = 1) \log \phi_L\left(x_n, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{c}\right) \\
&+ \sum_{n=2}^{N} I(S_n = 0, S_{n-1} = 1) \log P\left(S_n = 0 | S_{n-1} = 1\right) \\
&+ \sum_{n=2}^{N} I(S_n = 0, S_{n-1} = 0) \log P\left(S_n = 0 | S_{n-1} = 0\right) \\
&+ \sum_{n=2}^{N} I(S_n = 1, S_{n-1} = 1) \log P\left(S_n = 1 | S_{n-1} = 1\right) \\
&+ \sum_{n=2}^{N} I(S_n = 1, S_{n-1} = 0) \log P\left(S_n = 1 | S_{n-1} = 0\right)
\end{aligned}$$

Taking the expected log-likelihood :

$$\begin{aligned}
\mathbb{E}\left\{\log \mathcal{L}(\kappa, \alpha, \mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\sigma}; \mathbf{X}, \mathbf{S}, \mathbf{V}, \mathbf{W})|\mathbf{X}\right\} &= \mathbb{E}\left\{I(S_1 = 0)|\mathbf{X}\right\} log(\kappa) + \mathbb{E}\left\{I(S_1 = 1)|\mathbf{X}\right\} \log(1-\kappa) \\
&+ \sum_{n=1}^{N} \mathbb{E}\left\{I(S_n = 0)|\mathbf{X}\right\} \log \phi_{\delta_0}(x_n, \alpha) \\
&+ \sum_{n=1}^{N} \mathbb{E}\left\{I(S_n = 1)|\mathbf{X}\right\} \log \phi_L\left(x_n, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{c}\right) \\
&+ \sum_{n=2}^{N} \mathbb{E}\left\{I(S_n = 0, S_{n-1} = 1)|\mathbf{X}\right\} \log P\left(S_n = 0|S_{n-1} = 1\right) \\
&+ \sum_{n=2}^{N} \mathbb{E}\left\{I(S_n = 0, S_{n-1} = 0)|\mathbf{X}\right\} \log P\left(S_n = 0|S_{n-1} = 0\right) \\
&+ \sum_{n=2}^{N} \mathbb{E}\left\{I(S_n = 1, S_{n-1} = 1)|\mathbf{X}\right\} \log P\left(S_n = 1|S_{n-1} = 1\right) \\
&+ \sum_{n=2}^{N} \mathbb{E}\left\{I(S_n = 1, S_{n-1} = 0)|\mathbf{X}\right\} \log P\left(S_n = 1|S_{n-1} = 0\right)
\end{aligned}$$

4.1 E-Step.

we define $x_{1:N} = \{x_1, \ldots, x_N\}$, $x_{1:n} = \{x_1, \ldots, x_n\}$ and $x_{n+1:N} = \{x_{n+1}, \ldots, x_N\}$.

$$\begin{aligned}
P\left(S_n = i, x_{1:N}\right) &= P\left(x_{n+1:N}|S_n = i\right) P\left(S_n = i, x_{1:n}\right) \\
P(S_n = i, x_{1:n}) &= \sum_{j=0}^{1} P(x_n|S_n = i) P(S_n = i|S_{n-1} = j) P(S_{n-1} = j|x_{1:n-1}) \\
\alpha_i(n) &= \sum_{j=0}^{1} P(x_n|S_n = i) \times a_{ji} \times \alpha_j(n-1) \\
\alpha_i(1) &= \sum_{j=0}^{1} P(x_1|S_1 = i) P(S_1 = i) \\
P(x_{n+1:N}|S_n = i) &= \sum_{j=0}^{1} P(x_{n+2:N}|S_{n+1} = j) P(x_{n+1}|S_{n+1} = j) P(S_{n+1} = j|S_n = i) \\
\beta_i(n) &= \sum_{j=0}^{1} \beta_j(n+1) \times P(x_{n+1}|S_{n+1} = j) \times a_{ij} \\
\beta_i(n) &= 1
\end{aligned}$$

$$
\begin{aligned}
\gamma_i(n) &= \mathbb{E}\left\{I(S_n = i)|x_{1:N}\right\} = P\left(S_n = i|x_{1:N}\right) \\
&= \frac{\alpha_i(n)\beta_i(n)}{\sum_{j=0}^{1} \alpha_j(n)\beta_j(n)} \\
\\
\xi_{ij}(n) &= \mathbb{E}\left\{I(S_{n-1} = i, S_n = j)|x_{1:N}\right\} = P\left(S_{n-1} = i, S_n = j|x_{1:N}\right) \\
&= \frac{\alpha_i(n-1)a_{ij}P\left(x_n|S_n = j\right)\beta_j(n)}{\sum_{r=0}^{1}\sum_{s=0}^{1} \alpha_r(n-1)a_{rs}P\left(x_n|S_n = s\right)\beta_s(n)} \\
&= \frac{\gamma_i(n)a_{ij}P\left(x_n|S_{n+1} = j\right)\beta_j(n+1)}{\beta_i(n)} \\
\gamma_0(n)\nu_0(n) &= \mathbb{E}\left\{I(S_n = 0, V_n = 0)|x_n\right\} = P\left(S_n = 1, V_n = 0|x_n\right) \\
&= \gamma_0(n) \times \frac{\alpha\delta_0(x_n)}{\alpha\delta_0(x_n) + (1-\alpha)\phi(x_n, 0)} \\
\\
\gamma_1(n)\omega_{1\ell}(n) &= \mathbb{E}\left\{I(S_n = 1, W_{n\ell} = 1)|x_n\right\} = P\left(S_n = 1, W_{n\ell} = 1|x_n\right) \\
&= \gamma_1(n) \times \frac{c_\ell\phi(x_n, \mu_\ell, \sigma_\ell)}{\sum_{j=1}^{L} c_j\phi(x_n, \mu_j, \sigma_j)}
\end{aligned}
$$

4.2 M-Step.

$$
\begin{aligned}
\widehat{\pi_i} &= \gamma_i(1) \\
\widehat{a_{ij}} &= \frac{\sum_{n=1}^{N-1} \xi_{ij}(n)}{\sum_{n=1}^{N-1} \gamma_i(n)} \\
\widehat{\kappa} &= \frac{1}{N} \sum_{n=1}^{N} \gamma_0(n) \\
\widehat{\alpha} &= \frac{\sum_{n=1}^{N} \gamma_0(n) \times \nu_0(n)}{\sum_{n=1}^{N} \gamma_0(n)} \\
\widehat{\mu_\ell} &= \frac{\sum_{n=1}^{N} \gamma_1(n) \times \omega_{1\ell}(n) \times x_n}{\sum_{n=1}^{N} \gamma_1(n) \times \omega_{1\ell}(n)} \\
\widehat{\sigma_\ell^2} &= \frac{\sum_{n=1}^{N} \gamma_1(n) \times \omega_{1\ell}(n) \times (x_n - \widehat{\mu_\ell})^2}{\sum_{n=1}^{N} \gamma_1(n) \times \omega_{1\ell}(n)} \\
\widehat{c_\ell} &= \frac{\sum_{n=1}^{N} \gamma_1(n) \times \omega_{1\ell}(n)}{\sum_{n=1}^{N} \gamma_1(n)}
\end{aligned}
$$

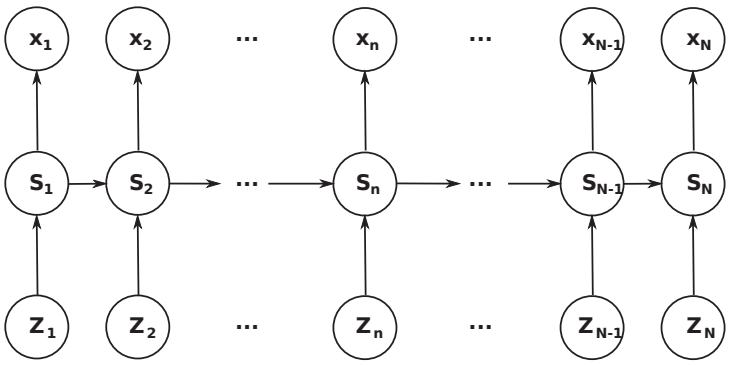


Figure 1: Bayesian network representation of a non-homogeneous HMM.

5 ZIG NHMM

The LIS theorems wrote by Sun and Cai (2009) show that the same level of FDR control can be obtain by the LIS procedure than by the FDR procedure if the observation X_n are conditionally independent according to the hidden state S_n.

$$P(X_n = x|S_n = s) = f_i(x)$$

with $f_i(x)$ the density function of the data where $S_n = i$. In the case where $S_n \in \{0, 1\}$ we are going to use $f_0(x) = \phi_{\delta_0}(x, \alpha)$ and $f_1(x) = \phi_L(x, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{c})$.

In the cases where the dependency structure is different from an homogeneous first order Markov chain, the departure from the dependency model will result in a relaxation of the FDR control. It is why instead of using an homogeneous HMMs we can use the non-homogeneous HMM (NHMM) framework of Hughes and Cuttorp to try to capture most of the dependency structure between the tests.

This model assume that we have a vector $\mathbf{Z}_{1:N} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_N)$ of covariables with $\mathbf{Z}_n$ a vector of D covariables associated with X_n. We can see from the Figure 1 that two assumptions are made :

$$\begin{aligned} P(s_n|s_{1:n}, \mathbf{z}_{1:n}, x_{1:n}) &= \begin{cases} P(s_n|s_{n-1}, \mathbf{z}_n) & n \geq 2 \\ P(s_1|\mathbf{z}_1) & n = 1 \end{cases} \\ P(x_n|s_n, \mathbf{z}_{1:N}, x_{n-1}) &= P(x_n|s_n) \end{aligned}$$

In this model the value of the emission probability and the transition matrix are function of $\mathbf{Z}$

$$\begin{aligned} \pi_j(\mathbf{z}) &= P(S_1 = j|\mathbf{Z}_1 = \mathbf{z}) \\ a_{ij}(\mathbf{z}) &= P(S_n = j|S_{n-1} = i, \mathbf{Z}_n = \mathbf{z}) \end{aligned}$$

As we work with probability (i.e. defined in $[0, 1]$) Hughes and Cuttorp chose to employ multi-

nomial logistic regression to parametrize the hidden state transition.

$$\begin{aligned}\pi_j(\mathbf{z}) &= \frac{\exp\left(\lambda_j + \boldsymbol{\rho}_j^n \times \mathbf{z}\right)}{\sum_{k=1}^{K} \exp\left(\lambda_k + \boldsymbol{\rho}_k^n \times \mathbf{z}\right)} \\ a_{ij}(\mathbf{z}) &= \frac{\exp\left(\sigma_{ij} + \boldsymbol{\rho}_j^n \times \mathbf{z}\right)}{\sum_{k=1}^{K} \exp\left(\sigma_{ik} + \boldsymbol{\rho}_k^n \times \mathbf{z}\right)}\end{aligned}$$

with λ_j, $\sigma_{ij} \in \mathbb{R}$ and $\boldsymbol{\rho}_j \in \mathbb{R}^D$. With $\boldsymbol{\omega}_j$ the set of transition parameters for the state j we have to set $\boldsymbol{\omega}_0 = 0$ to guarantee the uniqueness of the parameters.

Note: A homogeneous HMMs can be seen as a particular case of NHMM where $\boldsymbol{\rho} = 0$.

$$\boldsymbol{\rho} = 0 \Leftrightarrow P(S_n = j|S_{n-1} = i, \mathbf{Z}_n = \mathbf{z}) = P(S_n = j|S_{n-1} = i)$$

With $\boldsymbol{\Theta}$ the set of NHMM parameters, the joint probability of the data and the hidden states is:

$$\begin{aligned}P(x_{1:N}, s_{1:N}|\mathbf{z}_{1:N}, \boldsymbol{\Theta}) &= P(s_1|\mathbf{z}_1, \boldsymbol{\Theta}) \prod_{n=2}^{N} P(s_n|s_{n-1}, \mathbf{z}_n, x_{n-1}, \boldsymbol{\Theta}) \prod_{n=1}^{N} P(x_n|s_n, \mathbf{z}_n, x_{n-1}, \boldsymbol{\Theta}) \\ &= P(s_1|\mathbf{z}_1, \boldsymbol{\Theta}) \prod_{n=2}^{N} P(s_n|s_{n-1}, \mathbf{z}_n, \boldsymbol{\Theta}) \prod_{n=1}^{N} P(x_n|s_n, \boldsymbol{\Theta}) \\ &= \prod_{j=0}^{K} \pi_j(\mathbf{z}_1)^{I(S_1=j)} \prod_{n=2}^{N} \prod_{i=0}^{K} \prod_{j=0}^{K} a_{ij}(\mathbf{z}_n)^{I(S_{n-1}=i, S_n=j)} \prod_{n=1}^{N} \prod_{i=0}^{K} f_i(x_n)^{I(S_n=i)}\end{aligned}$$

The log-likelihood of the NHMM model is given by:

$$\begin{aligned}\log P(x_{1:N}, s_{1:N}|\mathbf{z}_{1:N}, \boldsymbol{\Theta}) &= \sum_{j=0}^{K} I(S_1 = j) \log \pi_j(\mathbf{z}_1) \\ &+ \sum_{n=2}^{N} \sum_{i=0}^{K} \sum_{j=0}^{K} I(S_{n-1} = i, S_n = j) \log a_{ij}(\mathbf{z}_n) \\ &+ \sum_{n=1}^{N} \sum_{i=0}^{K} I(S_n = i) \log f_i(x_n)\end{aligned}$$

with the constraints

$$\begin{aligned}\sum_{j=0}^{K} \pi_j(\mathbf{Z}_1) &= 1 \\ \sum_{j=0}^{K} a_{ij}(\mathbf{Z}_n) &= 1 \text{ ,for } i \in \{0, \ldots, K\} \text{ and } n \geq 2\end{aligned}$$

The expected log-likelihood is:

$$\begin{aligned}\mathbb{E}\left\{\log P\left(x_{1:N}, s_{1:N}|\mathbf{z}_{1:N}, \mathbf{\Theta}\right)|\mathbf{X}, \mathbf{Z}\right\} &= \sum_{j=0}^{K}\mathbb{E}\left\{I\left(S_1 = j\right)|\mathbf{X}, \mathbf{Z}\right\}\log \pi_j\left(\mathbf{z}_1\right) \\ &+ \sum_{n=2}^{N}\sum_{i=0}^{K}\sum_{j=0}^{K}\mathbb{E}\left\{I\left(S_{n-1} = i, S_n = j\right)|\mathbf{X}, \mathbf{Z}\right\}\log a_{ij}\left(\mathbf{z}_n\right) \\ &+ \sum_{n=1}^{N}\sum_{i=0}^{K}\mathbb{E}\left\{I\left(S_n = i\right)|\mathbf{X}, \mathbf{Z}\right\}\log f_i\left(x_n\right)\end{aligned}$$

as for the homogeneous case, we can use the forward-backward equations to compute $P\left(s_n|x_{1:N}, \mathbf{z}_{1:N}\right)$:

$$\begin{aligned}\alpha_i\left(n\right) &= P\left(S_n = i, x_{1:n}|\mathbf{z}_{1:n}\right) \\ \beta_i\left(n\right) &= P\left(x_{n+1:N}|S_n = i, x_n, \mathbf{z}_{n+1:N}\right)\end{aligned}$$

with

$$\begin{aligned}\alpha_i\left(1\right) &= P\left(S_1 = i|\mathbf{z}_1\right) \times P\left(x_1|S_1 = i\right) \\ \alpha_i\left(n+1\right) &= P\left(x_{n+1}|S_{n+1} = i, \mathbf{z}_{n+1}\right)\sum_{j=0}^{K}P\left(S_{n+1} = i|S_n = j, \mathbf{z}_{n+1}\right)\alpha_j\left(n\right) \\ \beta_i\left(N\right) &= 1 \\ \beta_i\left(n\right) &= \sum_{j=0}^{K}P\left(S_n = i|S_{n+1} = j, \mathbf{z}_{n+1}\right)P\left(x_{n+1}|S_{n+1} = j, \mathbf{z}_n\right)\beta_j\left(n+1\right)\end{aligned}$$

Note: The likelihood of the data sequence $P\left(X_{1:N}|Z_{1:N}\right)$ can be computed by

$$\begin{aligned}P\left(x_{1:N}|z_{1:N}\right) &= \sum_{i=0}^{K}P\left(S_N = i, x_{1:N}|z_{1:N}\right) \\ &= \sum_{i=0}^{K}\alpha_i(N)\end{aligned}$$

5.1 E-Step.

$$\begin{aligned}\gamma_i(n) &= \mathbb{E}\left\{I(S_n = i)|x_{1:N}\mathbf{z}_{1:N}\right\} = P\left(S_n = i|x_{1:N}, \mathbf{z}_{1:N}\right) \\ &= \frac{\alpha_i(n)\beta_i(n)}{\sum_{j=0}^{K}\alpha_j(n)\beta_j(n)} \\ &= \frac{\alpha_i(n)\beta_i(n)}{\sum_{j=0}^{K}\alpha_j(N)}\end{aligned}$$

$$
\begin{aligned}
\xi_{ij}(n) &= \mathbb{E}\left\{I(S_{n-1}=i,S_n=j)|x_{1:N},\mathbf{z}_{1:N}\right\} = P\left(S_{n-1}=i,S_n=j|x_{1:N},\mathbf{z}_{1:N}\right) \\
&= \frac{\alpha_i(n-1)a_{ij}P\left(x_n|S_n=j\right)\beta_j(n)}{\sum_{r=0}^{K}\sum_{s=0}^{K}\alpha_r(n-1)a_{rs}P\left(x_n|S_n=s\right)\beta_s(n)} \\
&= \frac{\alpha_i(n-1)a_{ij}P\left(x_n|S_n=j\right)\beta_j(n)}{\sum_{j=0}^{K}\alpha_j(N)}
\end{aligned}
$$

In the case where $K=\{0,1\}$ and $f(x)=\kappa\phi_{\delta_0}(x,\alpha)+(1-\kappa)\phi_L\left(x,\boldsymbol{\mu},\boldsymbol{\sigma},\mathbf{c}\right)$, with the additional hidden variables V and W_ℓ as label for the compartments of respectively $\phi_{\delta_0}(x,\alpha)$ and the $\phi_L\left(x,\boldsymbol{\mu},\boldsymbol{\sigma},\mathbf{c}\right)$, we have:

$$
\begin{aligned}
\gamma_0(n)\nu_0(n) &= \mathbb{E}\left\{I(S_n=0,V_n=0)|x_{1:N},\mathbf{z}_{1:N}\right\} = P\left(S_n=1,V_n=0|x_{1:N},\mathbf{z}_{1:N}\right) \\
&= \gamma_0(n)\times\frac{\alpha\delta_0(x_n)}{\alpha\delta_0(x_n)+(1-\alpha)\phi(x_n,0)} \\
\gamma_1(n)\omega_{1\ell}(n) &= \mathbb{E}\left\{I(S_n=1,W_{n\ell}=1)|x_{1:N},\mathbf{z}_{1:N}\right\} = P\left(S_n=1,W_{n\ell}=1|x_{1:N},\mathbf{z}_{1:N}\right) \\
&= \gamma_1(n)\times\frac{\omega_\ell\phi(x_n,\mu_\ell,\sigma_\ell)}{\sum_{j=1}^{L}\omega_j\phi(x_n,\mu_j,\sigma_j)}
\end{aligned}
$$

with the contraint $\sum_{\ell=1}^{L}c_\ell=1$

5.2 M-Step.

$$
\begin{aligned}
\widehat{\kappa} &= \frac{1}{N}\sum_{n=1}^{N}\gamma_0(n) \\
\widehat{\alpha} &= \frac{\sum_{n=1}^{N}\gamma_0(n)\times\nu_0(n)}{\sum_{n=1}^{N}\gamma_0(n)} \\
\widehat{\mu_\ell} &= \frac{\sum_{n=1}^{N}\gamma_1(n)\times\omega_{1\ell}(n)\times x_n}{\sum_{n=1}^{N}\gamma_1(n)\times\omega_{1\ell}(n)} \\
\widehat{\sigma_\ell^2} &= \frac{\sum_{n=1}^{N}\gamma_1(n)\times\omega_{1\ell}(n)\times(x_n-\widehat{\mu_\ell})^2}{\sum_{n=1}^{N}\gamma_1(n)\times\omega_{1\ell}(n)}
\end{aligned}
$$

$$\widehat{c_\ell} \;=\; \frac{\sum_{n=1}^{N} \gamma_1(n) \times \omega_{1\ell}(n)}{\sum_{n=1}^{N} \gamma_1(n)}$$

Unfortunately, the set of transition parameters $\mathbf{\Omega} = (\boldsymbol{\omega}_0, \ldots, \boldsymbol{\omega}_K)$ have non-linear partial derivatives which can not be equated to zero analytically. We can use a conjugate gradient algorithm to update $\mathbf{\Omega}$ iteratively such that $Q(\mathbf{\Omega}_{t+1}) \geq Q(\mathbf{\Omega}_t)$.

$$\begin{aligned} Q(\mathbf{\Omega}_t) &= \sum_{j=0}^{K} \mathbb{E}\left\{I(S_1 = j) \,|\mathbf{X}, \mathbf{Z}\right\} \log \pi_j(\mathbf{z}_1) \\ &+ \sum_{n=2}^{N} \sum_{i=0}^{K} \sum_{j=0}^{K} \mathbb{E}\left\{I(S_{n-1} = i, S_n = j) \,|\mathbf{X}, \mathbf{Z}\right\} \log a_{ij}(\mathbf{z}_n) \\ &= \sum_{j=0}^{K} \gamma_j(1) \log \pi_j(\mathbf{z}_1) + \sum_{n=2}^{N} \sum_{i=0}^{K} \sum_{j=0}^{K} \xi_{ij}(n) \log a_{ij}(\mathbf{z}_n) \end{aligned}$$

For identifiability we have $\boldsymbol{\omega}_0 = 0$ (i.e. for $K = \{0, 1\}$, $\lambda_0 = \sigma_{00} = \sigma_{10} = \boldsymbol{\rho}_0 = 0$).

$$\boldsymbol{\Omega}_{t+1} = \boldsymbol{\Omega}_t + \nu_t \boldsymbol{\Phi}_t \text{ where } \nu_t = \arg\max_{\nu} Q\left(\boldsymbol{\Omega}_t + \nu \boldsymbol{\Phi}_t\right)$$

We have to compute, the gradient $\boldsymbol{\Phi}_t = \nabla Q\left(\boldsymbol{\Omega}_t\right)$ (which is the vector of first partial derivatives).

$$\nabla Q\left(\boldsymbol{\Omega}_t\right) = \left(\frac{\partial Q}{\partial \lambda_j}, \frac{\partial Q}{\partial \sigma_{ij}}, \frac{\partial Q}{\partial \boldsymbol{\rho}_j}\right)$$

In the case where $K = \{0, 1\}$, we have :

$$\begin{aligned}
\frac{\partial Q}{\partial \lambda_1} &= \gamma_1(1) - \pi_1(\mathbf{z}_1) \\
\frac{\partial Q}{\partial \sigma_{01}} &= \sum_{n=2}^{N} \left[\xi_{01}(n) - \gamma_0(n-1) a_{01}(n)\right] \\
\frac{\partial Q}{\partial \sigma_{11}} &= \sum_{n=2}^{N} \left[\xi_{11}(n) - \gamma_1(n-1) a_{11}(n)\right] \\
\frac{\partial Q}{\partial \boldsymbol{\rho}_1} &= \left[\gamma_1(1) - \pi_1(\mathbf{z}_1)\right] \mathbf{z}_1 + \sum_{n=2}^{N} \sum_{r=0}^{K} \left[\xi_{r1} - \gamma_r(n-1) a_{r1}(n)\right] \mathbf{z}_n
\end{aligned}$$

We use the Polak-Ribiere variation of the conjugate gradient algorithm to update $\boldsymbol{\Phi}$:

$$\begin{aligned}
\boldsymbol{\Phi}_0 &= -\nabla Q\left(\boldsymbol{\Omega}_0\right) \\
\boldsymbol{\Phi}_{t+1} &= \nabla Q\left(\boldsymbol{\Omega}_t\right) - \gamma_t \boldsymbol{\Phi}_t \\
\gamma_t &= \max\left\{\frac{\nabla Q\left(\boldsymbol{\Omega}_{t+1}\right)^{\mathrm{T}} \left(\nabla Q\left(\boldsymbol{\Omega}_{t+1}\right) - \nabla Q\left(\boldsymbol{\Omega}_t\right)\right)}{\nabla Q\left(\boldsymbol{\Omega}_t\right)^{\mathrm{T}} \nabla Q\left(\boldsymbol{\Omega}_t\right)}, 0\right\}
\end{aligned}$$

And the Newton-Raphson algorithm to perform a line search to find ν_t :

$$\begin{aligned}
\nu_0 &= 0 \\
\nu_{t+1} &= \nu_t - \frac{\frac{dQ(\boldsymbol{\Omega}_t + \nu \boldsymbol{\Phi}_t)}{d\nu_t}}{\frac{d^2 Q(\boldsymbol{\Omega}_t + \nu \boldsymbol{\Phi}_t)}{d\nu_t^2}}
\end{aligned}$$

In the case where $K = \{0, 1\}$, we have :

$$\begin{aligned}
\frac{dQ\left(\mathbf{\Omega}_t + \nu\mathbf{\Phi}_t\right)}{d\nu_t} &= \sum_{j=0}^{1}\left(\mathbf{\Phi}_{\lambda_j} + \mathbf{\Phi}_{\boldsymbol{\rho}_j^n} \times \mathbf{z}_1\right)\left(\gamma_j(1) - \pi_j(\mathbf{z}_1)\right) \\
&+ \sum_{n=2}^{N}\sum_{i=0}^{1}\sum_{j=0}^{1}\left(\mathbf{\Phi}_{\sigma_{ij}} + \mathbf{\Phi}_{\boldsymbol{\rho}_j^n} \times \mathbf{z}_n\right)\left(\xi_{ij}(n) - \gamma_i(n-1)a_{ij}(\mathbf{z}_n)\right) \\
\frac{d^2Q\left(\mathbf{\Omega}_t + \nu\mathbf{\Phi}_t\right)}{d\nu_t^2} &= -\sum_{j=0}^{1}\left(\mathbf{\Phi}_{\lambda_j} + \mathbf{\Phi}_{\boldsymbol{\rho}_j^n} \times \mathbf{z}_1\right)^2 \pi_j(\mathbf{z}_1)\left(1 - \pi_j(\mathbf{z}_1)\right) \\
&- \sum_{n=2}^{N}\sum_{i=0}^{1}\sum_{j=0}^{1}\left(\mathbf{\Phi}_{\sigma_{ij}} + \mathbf{\Phi}_{\boldsymbol{\rho}_j^n} \times \mathbf{z}_n\right)^2 a_{ij}(\mathbf{z}_n)\left(1 - a_{ij}(\mathbf{z}_n)\right)\gamma_i(n-1)
\end{aligned}$$

3 UrQt : an efficient and fast software for the Unsupervised Quality trimming of NGS data

Quality analysis of the 6 NGS samples

August 1, 2014

Contents

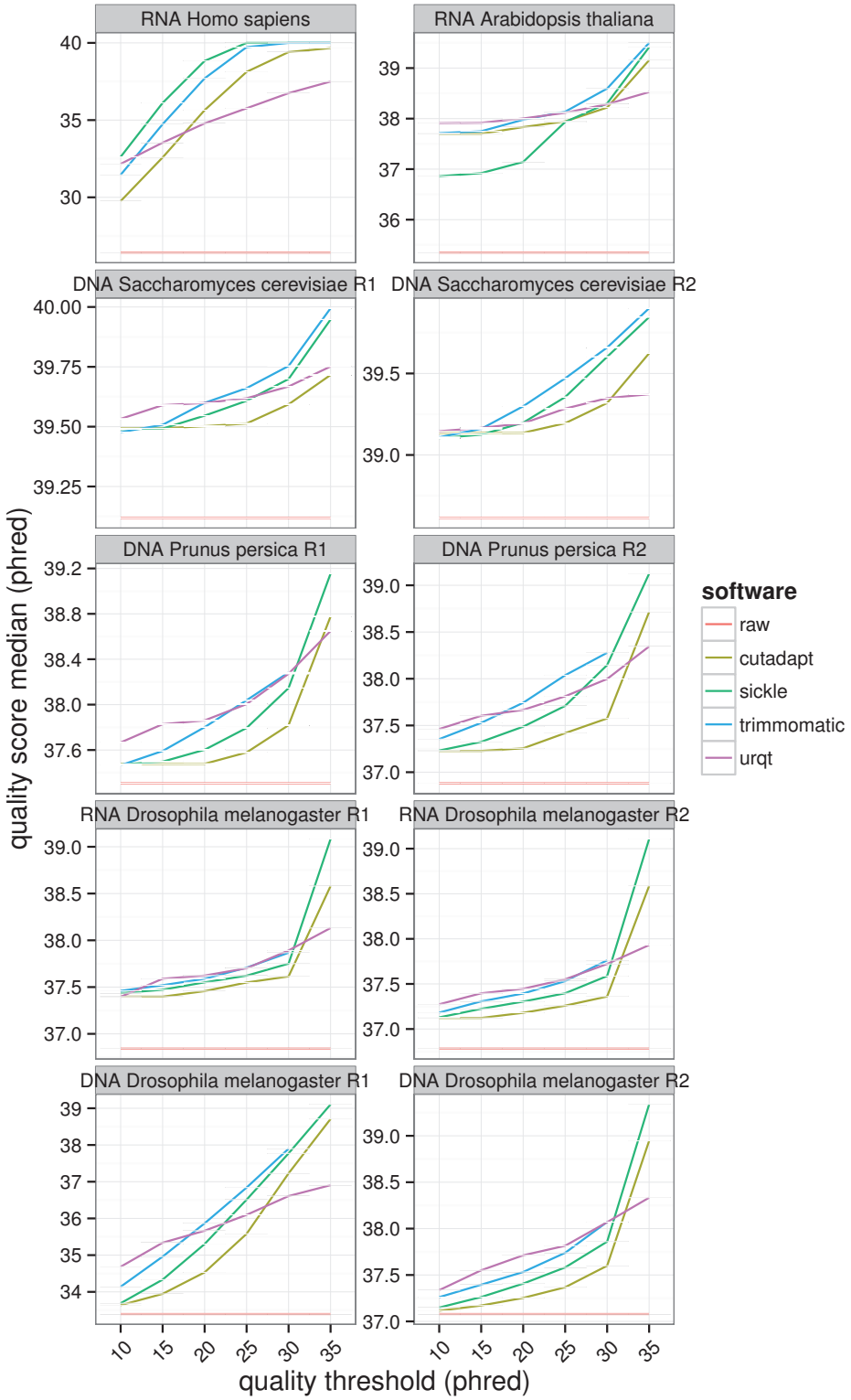


FIGURE 1 : Supplementary figure S2. Performances of different trimming algorithm in terms of the median quality (phred) of the resulting trimmed data set for different quality thresholds. The choice of t correspond to the parameter $-t$ for UrQt, $-q$ for Cutadapt and Sickle and *SLIDINGWINDOW :4 :t* for Trimmomatic. The black line corresponds to raw (untrimmed) data and R1 and R2 correspond to the two ends of paired-end data. This figure complete the Figure **??** with the six data sets

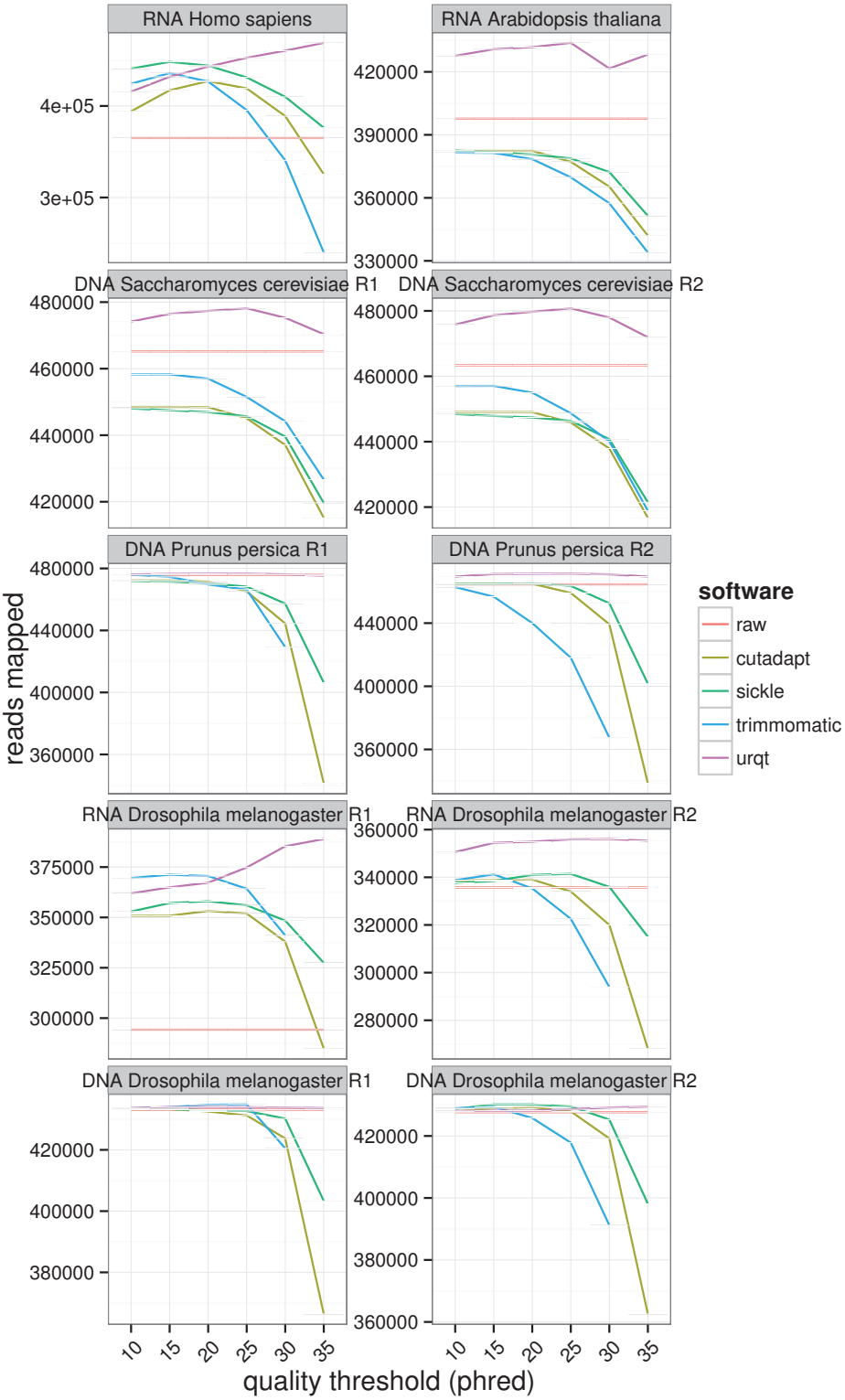


FIGURE 2 : Supplementary figure S2. Mapping performances for different quality threshold. The choice of t correspond to the parameter $-t$ for UrQt, $-q$ for Cutadapt and Sickle and *SLIDINGWINDOW :4 :t* for Trimmomatic. The black line corresponds to raw (untrimmed) data and R1 and R2 correspond to the two ends of paired-end data. This figure complete the Figure **??** with the six data sets

Table 1: **NGS data sets used for testing**

Accession number	Species	sample type	paired-end	read size (bp)	reference genome
SRR002073	*Homo sapiens*	RNA	no	33	hg19
SRR420813	*Arabidopsis thaliana*	RNA	no	83	TAIR10
SRX150254	*Prunus persica*	DNA	yes	100	1.22
SRR452441	*Saccharomyces cerevisiae*	DNA	yes	100	EF4
SRR988074	*Drosophila melanogaster*	DNA	yes	101	5.41
SRR919326	*Drosophila melanogaster*	RNA	yes	101	5.41

1 *Drosophila melanogaster* DNA sample (SRR988074)

1.1 Per base sequence quality

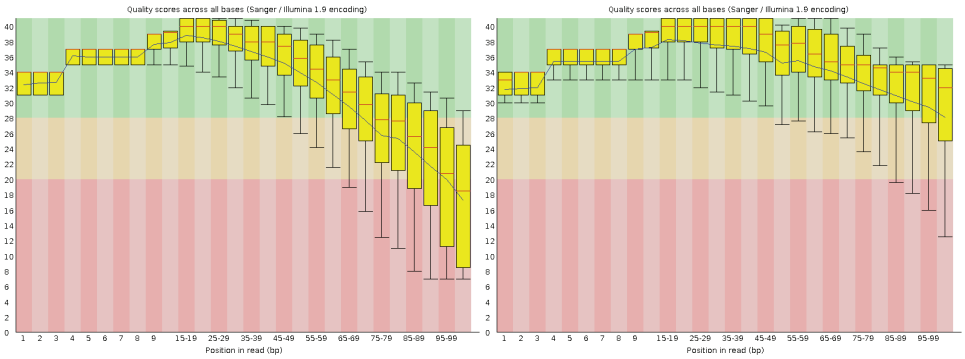


Figure 1: *Drosophila melanogaster* DNA: per base sequence quality. Left: R1, right: R2.

1.2 Per sequence quality scores

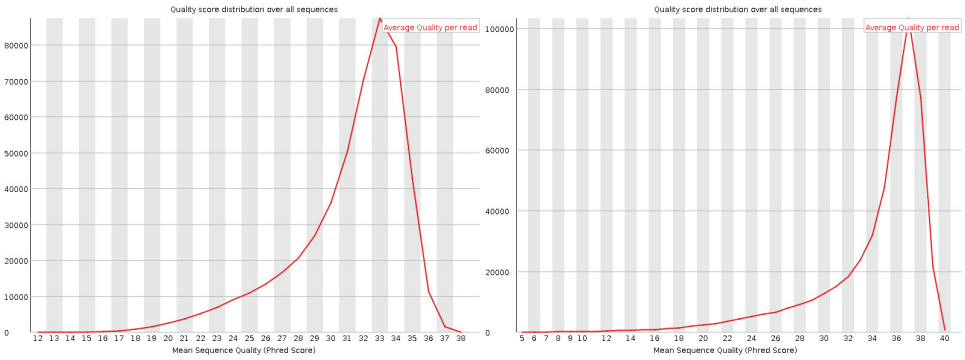


Figure 2: *Drosophila melanogaster* DNA: per sequence quality scores. Left: R1, right: R2.

1.3 Per base sequence content

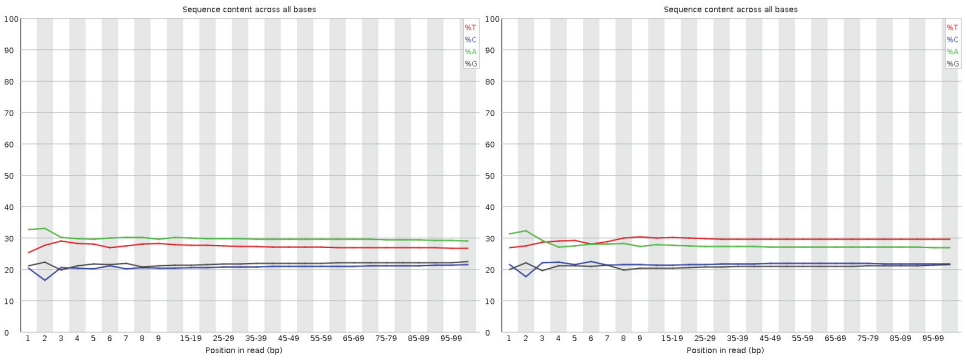


Figure 3: *Drosophila melanogaster* DNA: per base sequence content. Left: R1, right: R2.

1.4 Sequence Length Distribution

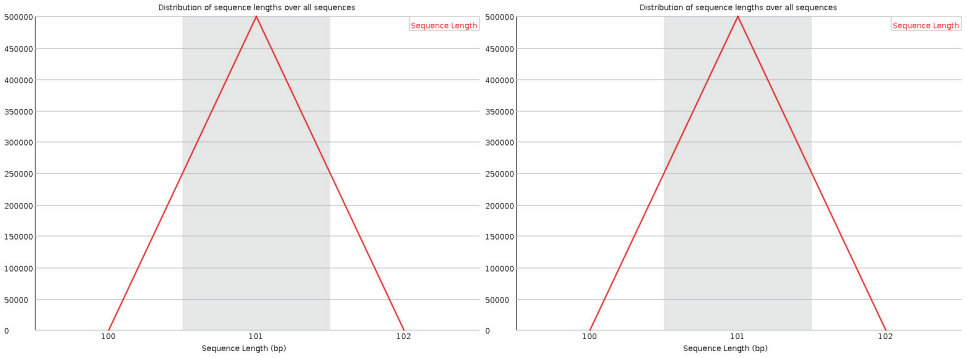


Figure 4: *Drosophila melanogaster* DNA: sequence Length Distribution. Left: R1, right: R2.

2 *Drosophila melanogaster* RNA sample (SRR919326)

2.1 Per base sequence quality

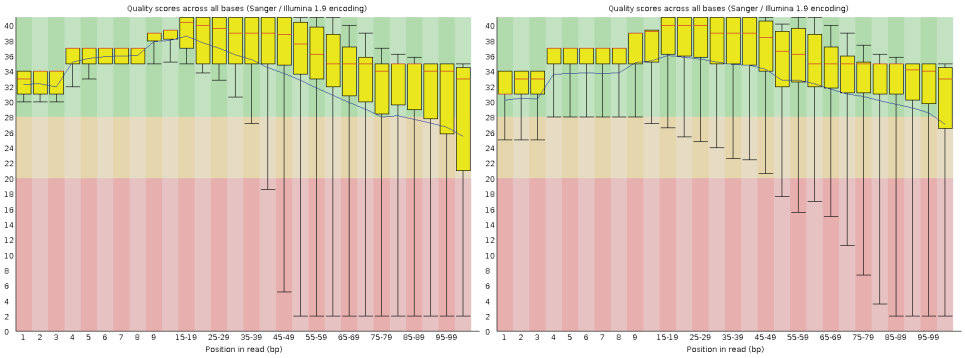


Figure 5: *Drosophila melanogaster* RNA: per base sequence quality. Left: R1, right: R2.

2.2 Per sequence quality scores

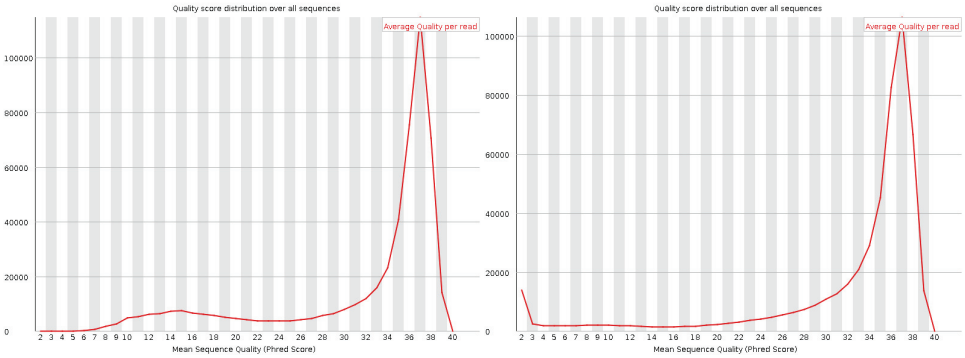


Figure 6: *Drosophila melanogaster* RNA: per sequence quality scores. Left: R1, right: R2.

2.3 Per base sequence content

2.4 Sequence Length Distribution

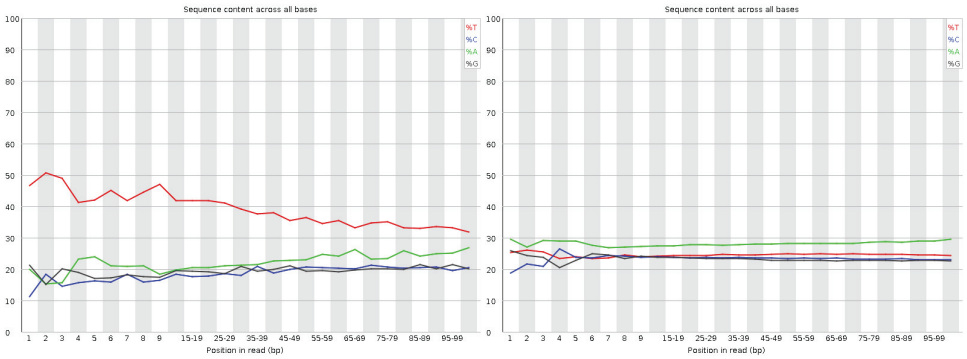


Figure 7: *Drosophila melanogaster* RNA: per base sequence content. Left: R1, right: R2.

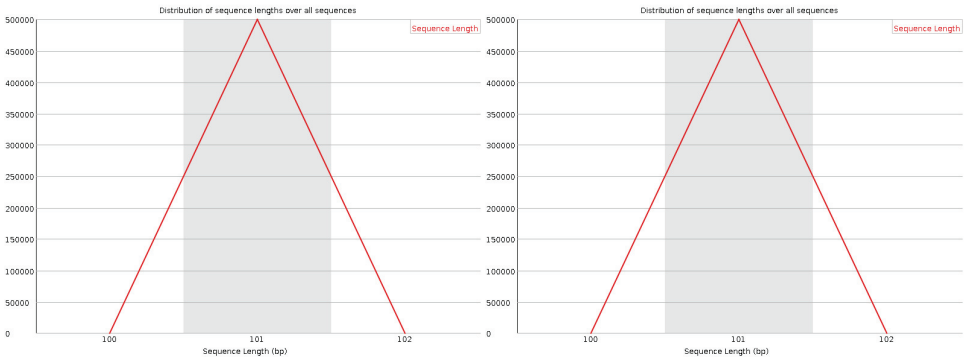


Figure 8: *Drosophila melanogaster* RNA: sequence Length Distributiony. Left: R1, right: R2.

3 *Homo sapiens* RNA sample (SRR002073)

3.1 Per base sequence quality

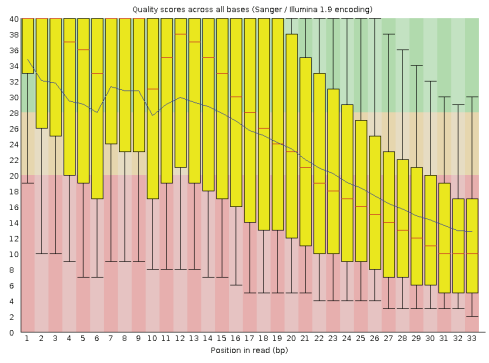


Figure 9: *Homo sapiens* RNA: per base sequence quality

3.2 Per sequence quality scores

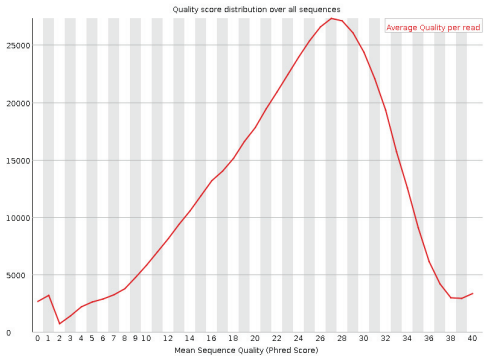


Figure 10: *Homo sapiens* RNA: per sequence quality scores

3.3 Per base sequence content

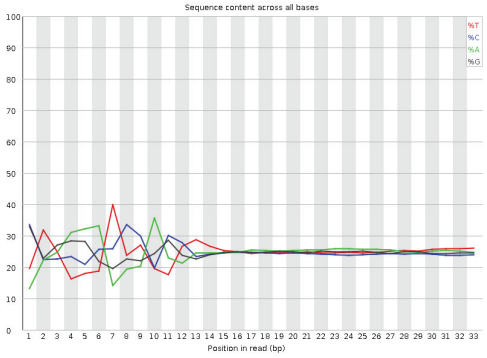


Figure 11: *Homo sapiens* RNA: per base sequence content

3.4 Sequence Length Distribution

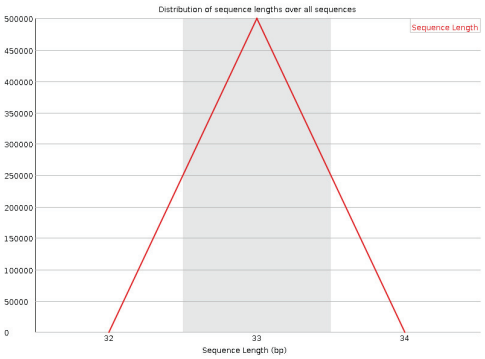


Figure 12: *Homo sapiens* RNA: sequence Length Distribution

4 *Arabidopsis thaliana* RNA sample (SRR420813)

4.1 Per base sequence quality

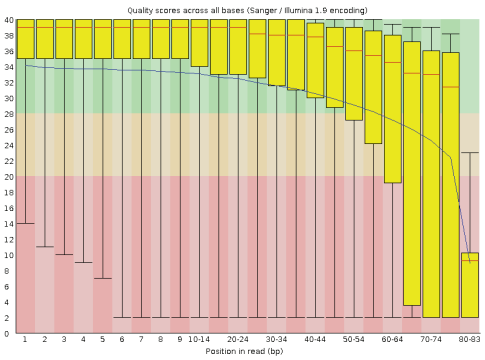


Figure 13: *Arabidopsis thaliana* RNA: per base sequence quality

4.2 Per sequence quality scores

4.3 Per base sequence content

4.4 Sequence Length Distribution

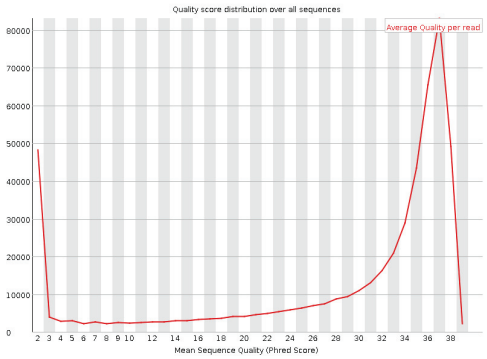


Figure 14: *Arabidopsis thaliana* RNA: per sequence quality scores

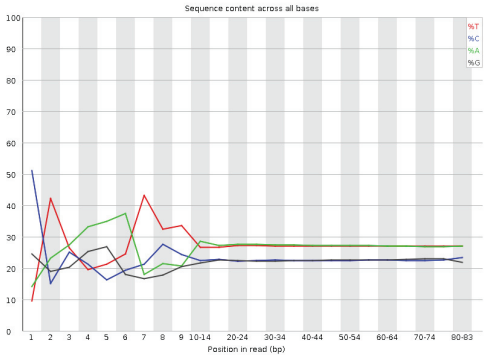


Figure 15: *Arabidopsis thaliana* RNA: per base sequence content

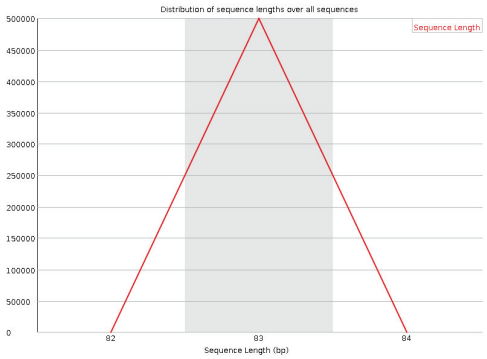


Figure 16: *Arabidopsis thaliana* RNA: sequence Length Distribution

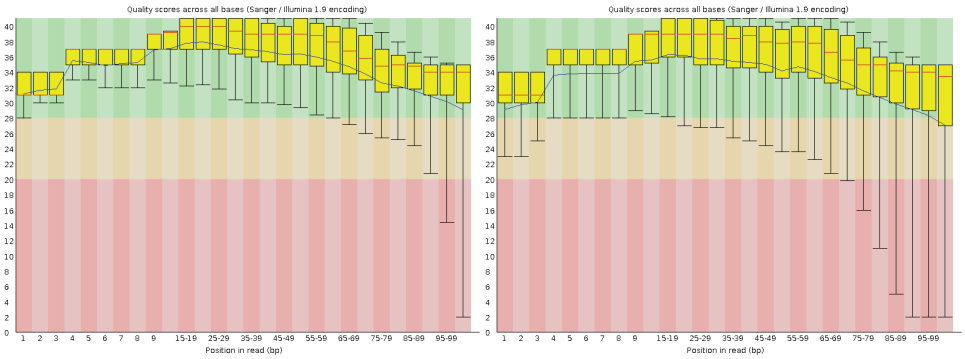


Figure 17: *Prunus persica* DNA: per base sequence quality. Left: R1, right: R2.

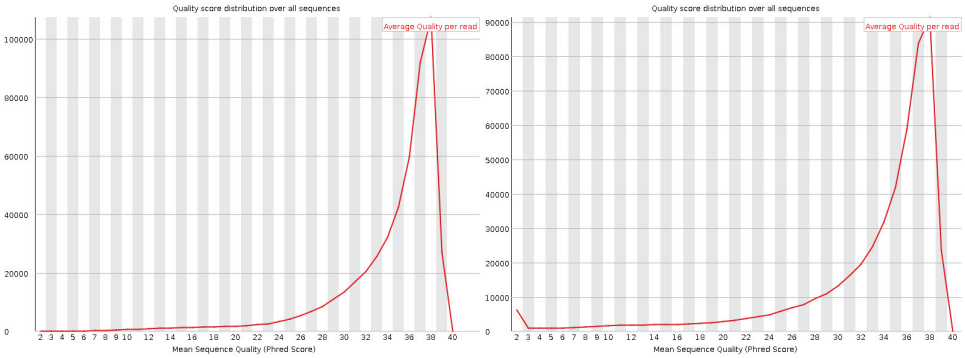


Figure 18: *Prunus persica* DNA: per sequence quality scores. Left: R1, right: R2.

5 *Prunus persica* DNA sample (SRX150254)

5.1 Per base sequence quality

5.2 Per sequence quality scores

5.3 Per base sequence content

5.4 Sequence Length Distribution

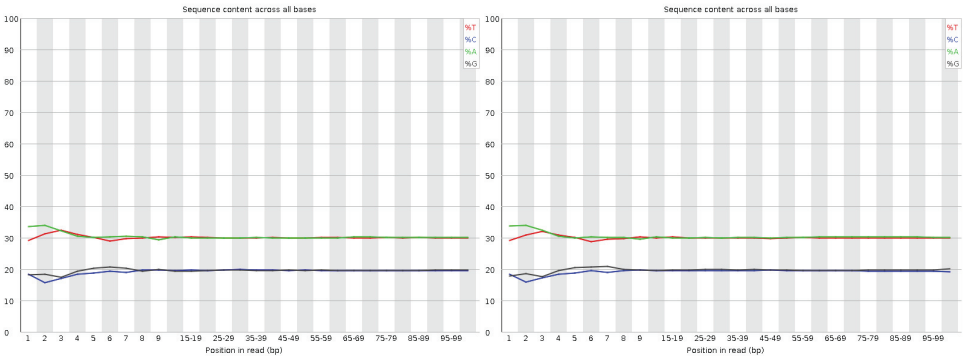


Figure 19: *Prunus persica* DNA: per base sequence content. Left: R1, right: R2.

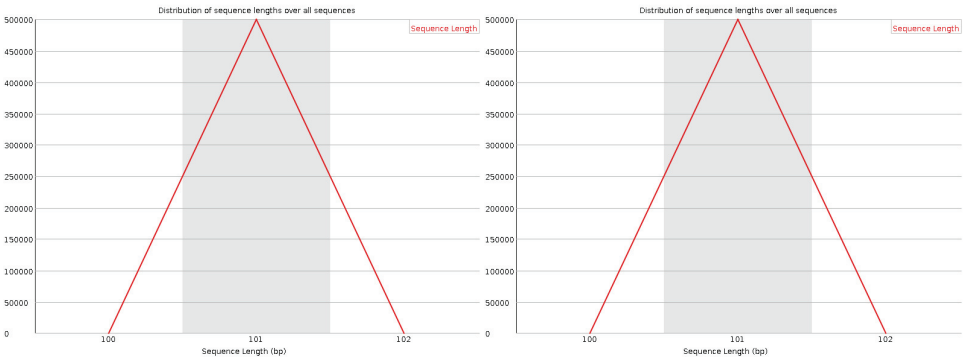


Figure 20: *Prunus persica* DNA: per sequence Length Distribution. Left: R1, right: R2.

6 *Saccharomyces cerevisiae* DNA sample (SRR452441)

6.1 Per base sequence quality

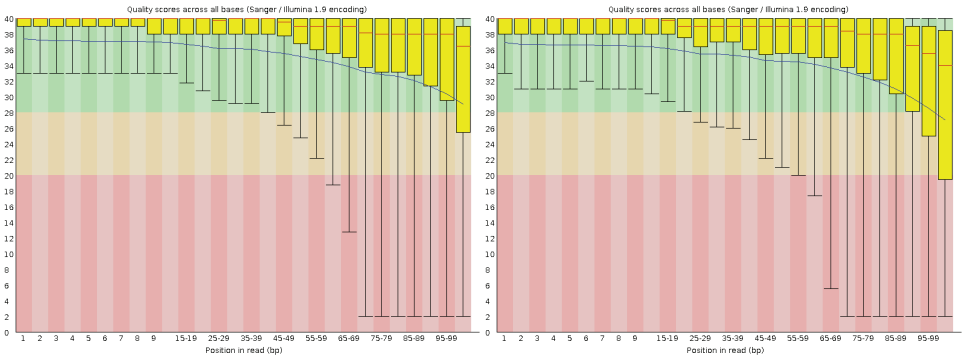


Figure 21: *Saccharomyces cerevisiae* DNA: per base sequence quality. Left: R1, right: R2.

6.2 Per sequence quality scores

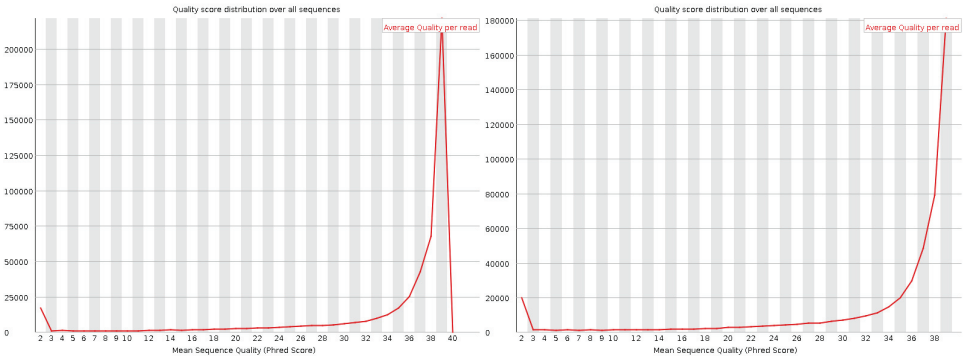


Figure 22: *Saccharomyces cerevisiae* DNA: per sequence quality scores. Left: R1, right: R2.

6.3 Per base sequence content

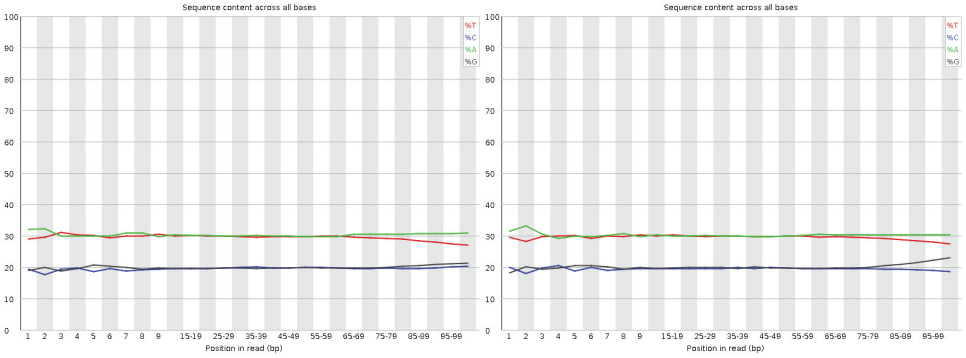


Figure 23: *Saccharomyces cerevisiae* DNA: per base sequence content. Left: R1, right: R2.

6.4 Sequence Length Distribution

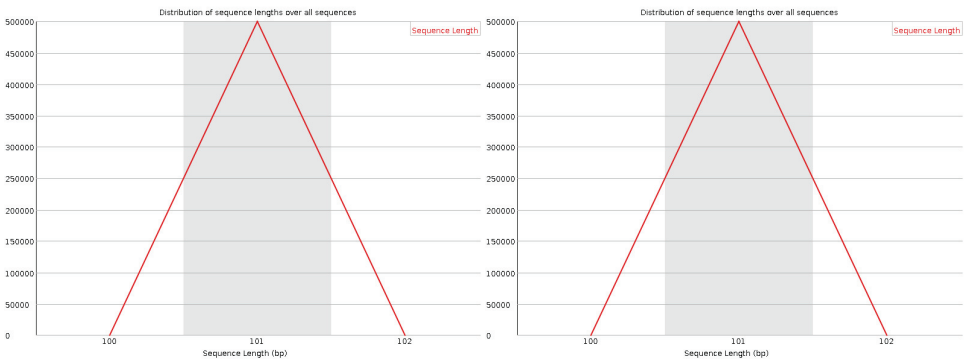


Figure 24: *Saccharomyces cerevisiae* DNA: per sequence Length Distribution. Left: R1, right: R2.

4 De novo assembly and annotation of the Asian Tiger Mosquito (*Aedes albopictus*) Repeatome from raw genomic reads with dnaPipeTE and comparative analysis with the yellow fever mosquito (*Aedes aegypti*)

Supplementary material

November 14, 2014

De novo assembly and annotation of the Asian Tiger Mosquito (*Aedes albopictus*) Repeatome from raw genomic reads with dnaPipeTE and comparative analysis with the yellow fever mosquito (*Aedes aegypti*)

1 Test of dnaPipeTE efficiency

In order to assess the method performances and boundaries, we tested dnaPipeTE on a set of reference genomes for which both fully assembled genome and suitable NGS sequences were available. We then compared the genome proportion of the main families and the TE landscapes made either with dnaPipeTE, or with RepeatMasker on assembled genomes. RepeatMasker analysis were already performed by the A. Smit team (Institute for Systems Biology), and fully available online at: http://repeatmasker.org/genomicDatasets/RMGenomicDatasets.html

1.1 Datasets

When available, we used the same strain in dnaPipeTE analysis than the sequenced one. We downloaded NGS datasets from the Short Read Archive (http://www.ncbi.nlm.nih.gov/sra), and cleaned the datasets using fastx-toolkit with the parameters used described in the Method section. For the smallest genome sizes we used an initial sample size of 0.25X, while we use 0.1X for the biggest ones (*Ae. aegypti* and *Homo sapiens*). We used different sample size since preliminary results showed that increasing sample size wont increase substantially the total amount of repeat found while it exponentially increased the computation time. We only used the R1 end of each file (single end) for the dnaPipeTE runs. The following table (Supp. Table 1) summarizes dataset specifications.

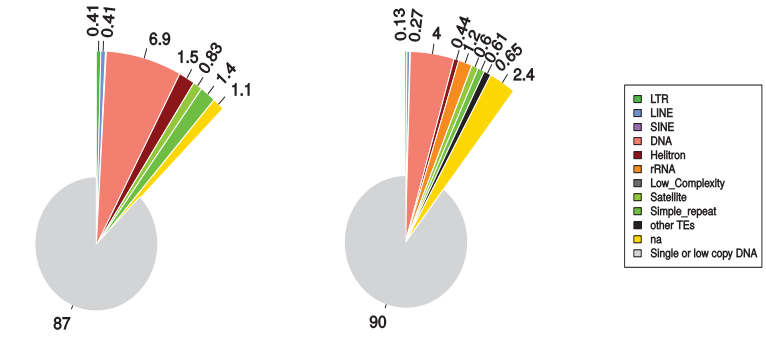
Table 1: Species and dataset used for dnaPipeTE tests. Genomes size are taken from whole genome assemblies and will be found at http://repeatmasker.org/. Unless otherwise stated, datasets used for dnaPipeTE are the R1 reads from 101 bp Illumina HiSeq2000 sequencing

Species	Genome size	Assembled strain	RepeatMasker Analysis	dnaPipeTE strain	dnaPipeTE samples size (assembly + blast sample)	NCBI SRA archive
Drosophila melanogaster	162 Mbp	*w1118*	*http://repeatmasker.org /species/dm.html*	*w1118*	2x 0,25X + 0,25 X	SRR988075[a]
Anopheles gambiae	263 Mbp	PEST	*http://repeatmasker.org /species/anoGam.html*	*Unknown*	2x 0,25X + 0,25 X	ERR554052[b]
Caenorhabditis elegans	100 Mbp	N2	http://repeatmasker.org/ species/ce.html	N2	2x 0,25X + 0,25 X	DRR008444[c]
Ciona intestinalis	141 Mbp	HK	http://repeatmasker.org/ species/ci.html	T	2x 0,25X + 0,25 X	DRR018354[d]
Gasterosteus aculeatus	447 Mbp	Bear Paw	http://repeatmasker.org/ species/gasAcu.html	Bear Paw	2x 0,25X + 0,25 X	SRR070080[e]
Homo sapiens	3 Gbp	hg38	http://repeatmasker.org/ species/hg.html	NA18912	2x0,1X + 0,1X	SRR350153[f]
Aedes aegypti	1,3 Gbp	Liverpool	--	Liverpool	2x0,1X + 0,1X (82 bp Illumina Hiseq2000 reads)	SRR871496[g]

[a] Beijing Institute of Genomics, CAS, 2013
[b] Anopheles Genome Variation Project, 2014
[c] Center for Genetic Resource Information, Comparative Genomics Laboratory, National Institute of Genetics, Research Organization of Information and Systems, 2014
[d] Marine Genomics Unit, Okinawa Institute of Science and Technology
[e] Broad Institute, 2006
[f] 1000 Genomes Project, 2008
[g] Virginia Tech, 2013

1.2 Supplementary figures

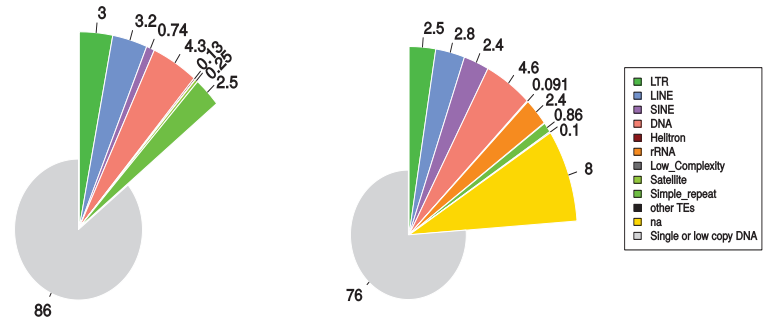
Caenorhabditis elegans



RepeatMasker on assembled genome Ce10 strain N2

dnaPipeTE on 0.75X 100 pb Illumina reads strain N2

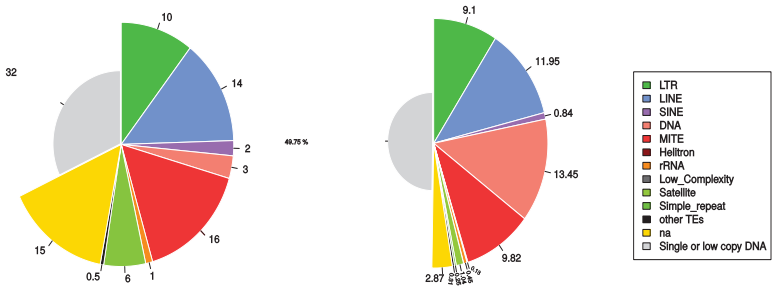
Anopheles gambiae



RepeatMasker on assembled genome anoGam1 strain PEST

dnaPipeTE on 0.75X 100 pb Illumina reads (sample location unknown)

Aedes aegypti



Nene et al. 2007 - WGS sequencing strain Liverpool

dnaPipeTE on 0.3X 82 pb Illumina reads strain Liverpool

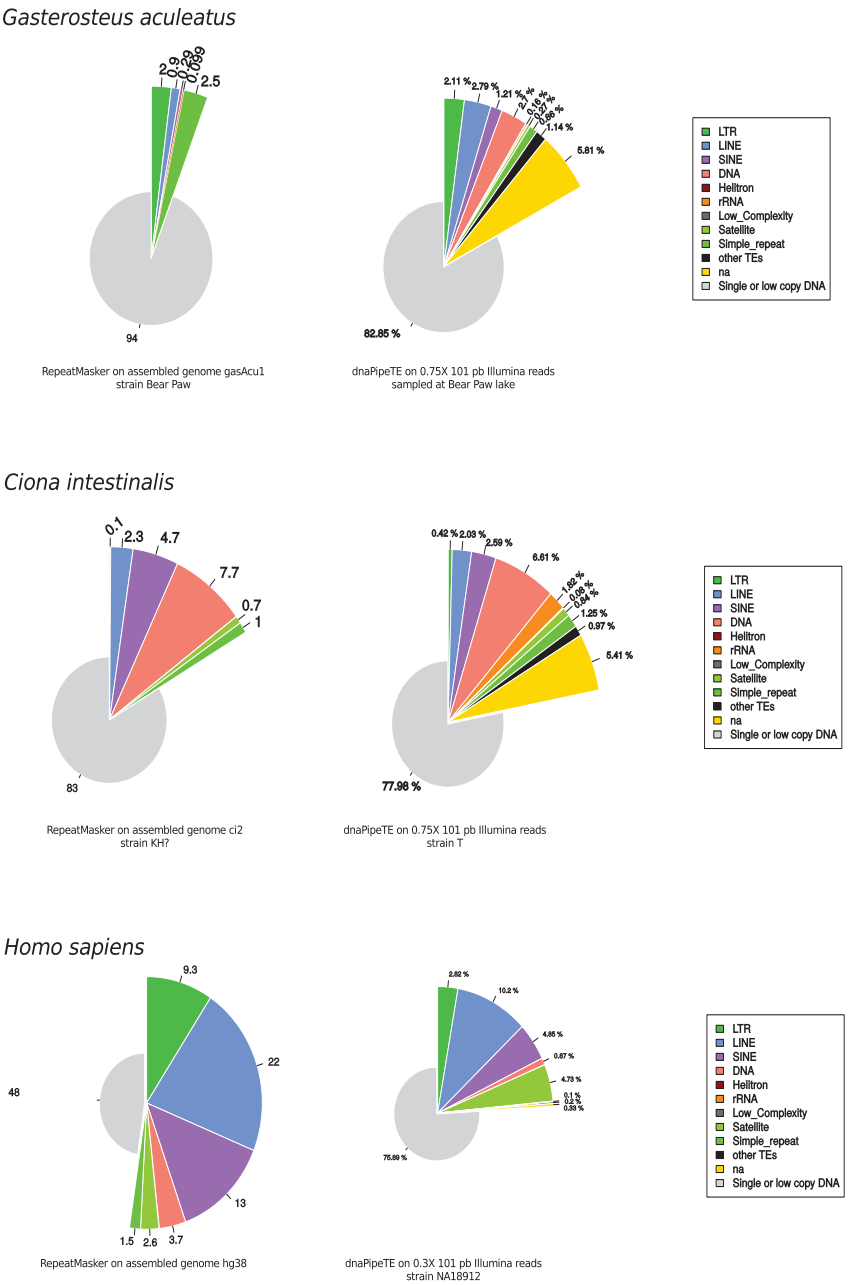


Figure 1: Estimation of the repeated content with dnaPipeTE and comparison with whole assembled genome analysis. Pairs of piecharts summarize the overall amount of repeats classes either using RepeatMasker on the whole assembled genome (left) [except for *Ae. aegypti*, data from the TE content analysis performed by Nene et al. 2007], either dnaPipeTE on single end Illumina reads (right). Values are given in percentage of the genome content.

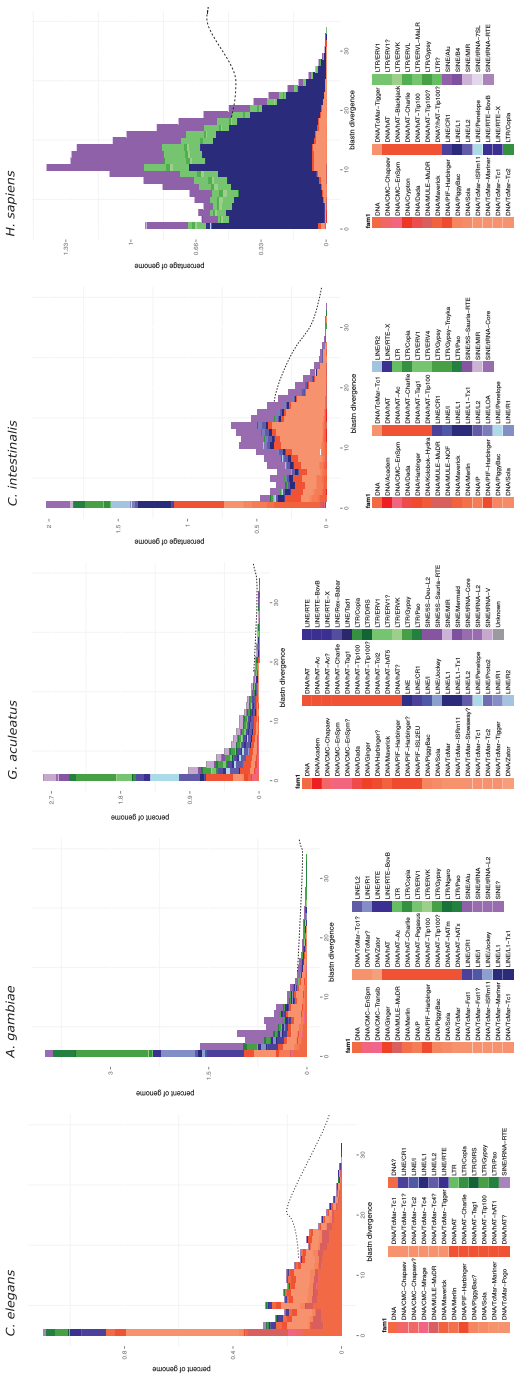


Figure 2: TE age distribution (TE landscapes) from blastn divergence between reads and their dnaPipeTE contigs. Barplots represent the total amount of reads from TEs (in genome percentage) according to their divergence from their matching dnaPipeTE contig, using blastn. Dashed black line represent the shape of the distribution observed using RepeatMasker on assembled genomes.

1.3 Discussion

In most cases, dnaPipeTE was able to find most of the described TE and other repeats classes in the NGS datasets. Globally, the estimation of the total amount of repeat using dnaPipeTE appear relevant, regarding to the estimation made from the assembled genomes. Estimation of the TE content was really good for samples using exactly the same strain in both analysis (*D. melanogaster* and *C. elegans*). Estimations were also very good in *C. intestinalis* and *A. gambiae* for wich the strains were different or unidentified; we noticed that in both species, dnaPipeTE identified new repeats.

Looking at the results from *G. aculeatus*, we found much more repeats with dnaPipeTE than the RepeatMasker analysis did on the assembled genome. New annotations mostly came from other fishes, thus we are quite confident that they are not false positive. However it is possible that the current G. aculeatus did not include all the repeats, depending on the sequencing and assembly method used and / or, although they came from the same place (Bear Paw lake), that the NGS sequenced sample is divergent from the reference sequenced individual.

In *Ae. aegypti*, we note that if the total amount of Class I repeats (MITEs and DNA) are close (19% vs 24%) however dnaPipeTE found less MITEs TEs than expected. MITEs are short, without coding sequences TEs derived from DNA transposons. It is thus possible that most of them have been identified as DNA, without better available annotation.

In the Human genome however, dnaPipeTE did not manage to estimate accurately the repeated content. While inside the repeated content, the relative proportion of each TE classes is well estimated, we only found that it represent half of its actual size. This, is certainly due to the particular profile of the TE age in the Human genome (and many vertebrates) where TE are mostly ancient. Thus, there is less identity between reads from different copies of the same TE family and the assembly will fail to find the older families. This is the main limitation of methods based on low coverage datasets.

From the TE landscapes estimations, we showed that our method allow to catch variation un TE age distribution until at least 15% divergence. Below this threshold, the blastn method fail to match reads with more divergent contigs, and thus those TEs will be dropped out from the report. However, we can clearly disctinct caracteristic shapes between models that fits with the fully assembled genomes TE landscapes that are available at http://repeatmasker.org/genomicDatasets/RMGenomicDatasets.html In addition, we can note that the first bin in those graphs (0 to 1% divergence) is inflated comparing to RepeatMasker analysis on fully assembled genomes. This issue is discussed in the paper.

2 Sample size choice for *Ae. albopictus* and interest of multiple iteration

In order to maximize the N50 of assembly while not assembling too much none repeated genome content, we tested both dnaPipeTE with different sample size and Trinity iterations (see material and methods). We found that in *Ae. albopictus*, the best compromise was to choose a combination

of two iterations with a sample size of 0,1X. On figure S.X. Each combination of Trinity iteration and sample size was tested two times, except for 0,1X and 0,17X that have three repetitions.

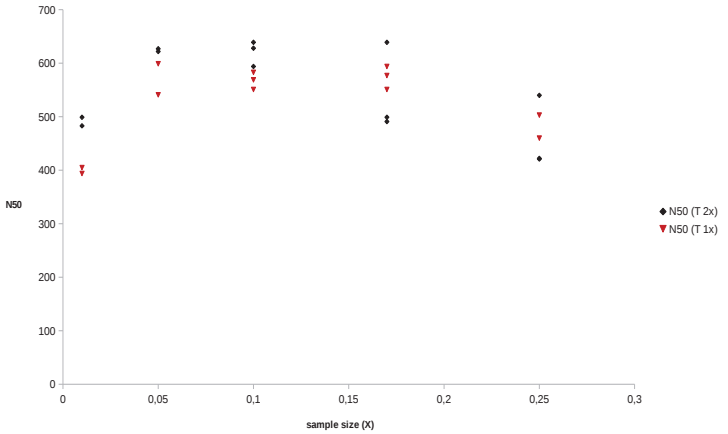


Figure 3: Assembly N50 after the first (T 1x) and the second (T 2x) Trinity iteration in dnaPipeTE for *Ae. albopictus*, according to sample size. For T 2x, two samples of the same size were used successively, according to the description made in material and methods

In addition we provide here the results from D. melanogaster for 1st and 2nd iteration of Trinity. After the first Trinity assembly on a 0,25X sample, the N50 was 1342 bp for 1302 contigs. Then adding a new sample of 0,25X to the assembled reads for the second iteration, the N50 rose to 2054 bp with 2590 contigs.

3 Supplementary data

Supplementary data 1 : annotated full-length contigs (dnaPipeTE_full_lengths_TE_albo.fasta)
Supplementary data 2: annotated partial contigs (dnaPipeTE_partial_TE_albo.fasta)

4 Blat results of *Ae. albopictus* assembled transcriptome on dnaPipeTE contigs (blast format)

Column Content 1 Transcriptome contig (query) 2 dnaPipeTE contigs (target) 3 Percentage indentity 4 Alignment size (bp) 5 # mismatches 6 # gap opening 7 Query start 8 Query end 9 Subject start 10 Subject end 11 E-value 12 Bit Score