HAL open science

Analyses et méthodes pour les données transcriptomiques issues d'espèces non modèles : Variation de l'expression des éléments transposables (et des gènes) et variants nucléotidiques

Hélène Lopez-Maestre

▶ **To cite this version:**

Hélène Lopez-Maestre. Analyses et méthodes pour les données transcriptomiques issues d'espèces non modèles : Variation de l'expression des éléments transposables (et des gènes) et variants nucléotidiques. Bio-Informatique, Biologie Systémique [q-bio.QM]. Université Claude Bernard Lyon 1, 2017. Français. NNT : . tel-01575640v2

HAL Id: tel-01575640

https://inria.hal.science/tel-01575640v2

Submitted on 7 Sep 2017 (v2), last revised 14 Sep 2017 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

N° d'ordre NNT : 2017LYSE1025

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

École Doctorale ED 342
Ecosystèmes Evolution Modélisation Microbiologie

Spécialité de doctorat : Biomath-Bioinfo Génomique évolutive

Soutenue publiquement le 15/02/2017, par :
Hélène Lopez-Maestre

Analyses et méthodes pour les données transcriptomiques issues d'espèces non modèles

variation de l'expression des éléments transposables (et des gènes) et variants nucléotidiques

---

Devant le jury composé de :

Malika Ainouche, Professeure des universités, Université Rennes 1	Rapporteure/Présidente
Eric Rivals, Directeur de recherche - CNRS, LIRMM	Rapporteur
François Sabot, Chargé de Recherche, UMR-DIADE IRD	Examinateur
Vieira Cristina, Professeure des universités, UCBL1	Directrice de thèse
Lacroix Vincent, Maitre de conférence, UCBL1	Co-directeur de thèse

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. le Professeur Frédéric FLEURY**
Président du Conseil Académique	M. le Professeur Hamda BEN HADID
Vice-président du Conseil d'Administration	M. le Professeur Didier REVEL
Vice-président du Conseil Formation et Vie Universitaire	M. le Professeur Philippe CHEVALIER
Vice-président de la Commission Recherche	M. Fabrice VALLÉE
Directrice Générale des Services	Mme Dominique MARCHAND

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur G.RODE
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. X. PERROT
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. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. THEVENARD
Département Chimie Biochimie	Directeur : Mme C. FELIX
Département GEP	Directeur : M. Hassan 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 : M. le Professeur 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. le Professeur E.PERRIN
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

Résumé

Le développement de la seconde génération de séquenceurs haut débit a généralisé l'accès à l'étude du transcriptome via le protocole RNAseq. Celui-ci permet d'obtenir à la fois la séquence et l'abondance des transcrits d'un échantillon. De nombreuses méthodes bioinformatiques ont été et sont encore développées pour permettre l'analyse des données issues du RNAseq et en tirer le maximum d'information. Ce type d'analyse est notamment possible sans utiliser de génome de référence, et donc pour les espèces modèles ou non-modèles, grâce à des méthodes d'assemblage.

Durant ma thèse, j'ai principalement travaillé à partir de données RNA-seq issues d'espèces non modèles. Je me suis intéressée dans un premier temps à l'impact de l'hybridation inter spécifique sur la stabilité des génomes chez les hybrides issus des croisements réciproques de *D. mojavensis* et *D. arizonae*. Nos résultats ne montrent pas une dérégulation globale, mais plutôt quelques gènes et éléments transposables qui sont spécifiquement dérégulés. La pipeline d'analyse mis en place ici sera réutilisée pour l'étude des niveaux d'expression des transcrits chez les mâles ainsi que pour les croisements issus d'autres lignées de *D. mojavensis* avec *D. arizonae*, conduisant à une fertilité variable chez les hybrides.

Dans un second temps, j'ai participé à la validation du logiciel KisSplice pour la détection de SNP dans des données RNA-seq sans génome de référence. Celui-ci permet de trouver différents types de variants (épissage, indels) directement dans le graphe de de Bruijn construit à partir des lectures séquencées. J'ai également participé au développement d'outils de post-traitement permettant de prédire l'impact des SNP sur les protéines.

Abstract

Next-generation high throughput sequencing technologies provide efficient, rapid, and low cost access to sequencing. Its application to transcriptomes, called RNA-seq, enables the study of both the sequence and the expression of the transcripts. Many bioinformatics methods are still developed for RNA-seq data processing, trying to get the maximum out of it. Assembly methods allow us to study non-model species (no reference genome available) as well as model species. The work presented here is mostly related to RNA-seq data on non-model species.

In the first study, to understand the initiation of hybrid incompatibility, we performed a genome-wide transcriptomic analysis on ovaries from parental lines and on hybrids from reciprocal crosses of *D. mojavensis* and *D. arizonae*. We didn't see a global deregerulation of genes or transposable element. Instead, we show that reciprocal hybrids presented specific gene categories and few transposable element families misexpressed relative to the parental lines. The analytical workflow developed for this project will be used to analyze transcriptomic data from the testis, but also to study the reciprocal crosses from other lines of *D. mojavensis* with *D. arizonae* leading to variable levels of sterility in hybrids.

A second project tacked here is the identification and quantification of SNPs from RNA-seq data without a reference genome with KisSplice. Kissplice was developed to identified several type of variants (splicing events, indels) directly from the de Bruijn graph, build from the sequenced reads. We also developed other KisSplice-tools, for downstream analyses of the SNPs, including the prediction o their impact on the protein sequence.

Table des matières

Introduction générale **10**

1 Introduction **12**

1 Transcriptomique 13

1.1 La transcription 13

1.2 Technologies pour l'analyse à large échelle du transcriptome 15

1.3 Design expérimental 20

2 L'analyse de données RNA-seq 22

2.1 Méthodes d'alignements des lectures 22

2.2 Méthodes d'assemblage (sans génome de référence) 24

2.3 Reconstruction des transcrits et épissages alternatifs 27

2.4 Identification des variants nucléotidiques 28

2.5 Accès à la quantification 29

2.6 Comparer deux (ou plus) échantillons 30

3 Les répétitions et éléments transposables 31

3.1 Analyse bio-informatique des éléments transposables en génomique 34

3.2 Analyse bio-informatique des éléments transposables en RNAseq . . 35

4 Conclusion 37

2 Expression des éléments transposables (et des gènes) chez les hybrides de *D. mojavensis* et *D. arizonae* **38**

1 Avant-propos 39

2 Article 1 : Identification of misexpressed genetic elements in hybrids between Drosophila-related species 40

3 Supplementary Information 80

3 Détection de SNP dans les données RNAseq sans génome de référence 96

1 Avant-propos 97

2 Article 2 : SNP calling from RNA-seq data without a reference genome : identification, quantification, differential analysis and impact on the protein sequence 98

3 Supplementary Information 112

4 Conclusion et Perspectives 133

1 Hybrides 134

2 Détection des variants nucléotidiques 136

A Articles annexes 151

1 Hybrides et éléments transposables 152

2 TEtools : quantification des éléments transposables et des piRNA dans des données RNA-seq 215

Liste des figures

1 Transcription, épissage alternatif et traduction 14

2 Les différentes technologies de séquençage (par génération) 17

3 Séquençage illumina 18

4 Graphe d'overlap et graphe de de Bruijn 24

5 Les répétitions dans le graphe d'overlap et graphe de de Bruijn 25

6 Exemple de graphe de de Bruijn construit à partir de 100000 lectures issues du séquençage transcriptomique de *D. mojavensis*. 26

7 Exemple d'erreur d'assemblage dans un graphe de de Bruijn. 27

8 Structure des différents types d'éléments transposables 33

Introduction générale

Introduction générale Avec les développements des technologies de séquençage à haut débit, il est maintenant possible d'étudier, avec une même expérience, le transcriptome à large échelle. Nous pouvons non seulement avoir accès à la séquence des transcrits mais aussi à la quantification de leur expression (RNAseq). Les technologies de séquençages ne nécessitent pas de connaissances au préalable des séquences de l'espèce étudiée, et peuvent ainsi être utilisées sur les espèces modèle comme sur les espèces non modèles, avec ou sans génome de référence. L'analyse de données issues d'espèces non-modèles, nécessite le développement d'outils appropriés, basés notamment sur des méthodes d'assemblage. De plus, un des grands défis des méthodes d'assemblage concerne la prise en compte des répétitions, que souvent sont un entrave à un bon assemblage.

Durant ma thèse, je me suis intéressée à l'analyse de données de RNAseq dans le cas d'espèces non modèles. Ce travail a porté d'une part sur la validation d'une méthode pour la détection de SNP (« single nucleotide polymorphism ») et d'autre part à l'analyse de données RNAseq issu de deux espèces non modèles et de leurs hybrides.

Le premier chapitre de cette thèse constitue une introduction méthodologique à l'étude du transcriptome à partir de données RNA-seq. J'y présente les différentes méthodes (et les problématiques associées) permettant l'analyse de ce type de données, pour les gènes mais également pour les éléments transposables, qui font partie de la portion répétée des génomes.

Le second chapitre est consacré à l'étude de l'impact de l'hybridation inter spécifique sur la régulation des gènes et des éléments transposables. L'étude présente l'analyse des niveaux d'expression de deux espèces de drosophile ayant divergé il y a moins d'un million d'années (D. mojavensis et D. arizonae), ainsi que des hybrides issus de leur croise-

ment réciproques. Le pipeline développé ici, et qui pourra être réutilisé pour poursuivre l'étude sur les hybrides avec d'autres lignées et d'autres tissus, a la particularité d'utiliser un co-assemblage des différents transcriptomes. Ceci a permis d'augmenter artificiellement la couverture et de pouvoir reconstituer un transcriptome de référence le exhaustive possible. Ce travail a été publié blablab bla .Cette pipeline a également été utilisé dans le cadre une collaboration avec Valèria Romero Soriano en travaillant sur un autre modèle biologique permettant d'étudier l'impact de l'hybridation inter-spécifique sur la stabilité des génomes dans les cas de deux espèces plus divergentes. Les résultats de l'étude sont présentés en annexe de ce manuscrit.

Le troisième chapitre de cette thèse est quant à lui axé sur des aspects méthodologiques. J'ai participé à la validation du logiciel KisSplice pour la détection de SNP dans des données RNA-seq sans génome de référence. KisSplice a été initialement développé au sein l'équipe Erable afin d'identifier les variants d'épissages directement dans un graphe de de Bruijn construit à partir des lectures de RNA-seq. Son développement est issue d'une collaboration de plusieurs équipes de recherche dans le cadre de l'ARN *Colib'read*, qui propose des développement de méthodes basées directement sur les lectures séquencées pour répondre à différents problèmes biologiques (détection de SNP, d'épissage, d'inversions génomiques etc.). J'ai également participé au développement d'outils de post-traitement permettant de prédire l'impact des SNP sur les protéines. Les résultats de ces développements et la validation des méthodes ont fait l'object d'une publication dans au sein du journal Nucleotide acid research.

Enfin le dernier chapitre présentera une brève conclusion de ces différents travaux ainsi que les perspectives associées.

1 Introduction

1 Transcriptomique

La technique de séquençage, arrivée dès les années 70, a donné accès à la composition en nucléotide des molécules d'ADN (Sanger and Coulson [1975]). Cette avancée technologique a eu un impact considérable sur l'acquisition de nouvelles connaissances en biologie moléculaire, en évolution, génomique environnementale, dans le domaine médical et bien d'autres, ainsi que dans le développement d'outils statistiques et informatiques adaptés à ce type de données. Depuis, les technologies de séquençage ont évolué, en particulier avec l'arrivée de la seconde génération de séquenceurs (NGS), donnant un accès massif et à moindre coût aux séquences génomiques.

Ces technologies ont également permis le renouvellement des études transcriptomiques. La transcriptomique consiste en l'étude de l'ensemble des ARN (ou transcrits), souvent plus particulièrement les ARN messagers (ARNm), qui sont utilisés comme intermédiaires pour la production de protéines. Le transcriptome étudié peut être celui d'un type cellulaire particulier ou d'un tissu spécifique. La transcriptomique constitue aujourd'hui un domaine de recherche à part entière.

Le séquençage à haut débit de l'ARN (appelé RNA-seq) est actuellement la technologie la plus employée pour identifier et quantifier, à large échelle, les transcrits extraits d'un ou plusieurs individus, tissus ou types cellulaires, dans des conditions physiologiques données. Avec la production massive de données RNA-seq, des méthodes et outils spécifiques permettant l'analyse de ce type de données ont été et sont encore développés.

1.1 La transcription

La transcription permet la copie des portions d'ADN en des molécules intermédiaires "semblables", les ARN messagers, qui peuvent ensuite être traduits en protéines.

Chez les eucaryotes, les gènes sont constitués de parties dites codantes qui peuvent être traduites en acides aminés, les exons, et de parties dites non codantes, les introns. Le transcrit issu directement de la "copie" du gène, le pré-ARNm, n'est pas directement traduit et doit d'abord subir une étape de maturation. L'épissage dit constitutif consiste en l'excision des introns des pré-ARNm. Il est cependant très fréquent que l'épissage, appelé dans ce cas épissage alternatif, aboutisse à la rétention de certains introns et/ou à l'excision de certains exons dans les transcrits (près de 90% des gènes de l'humain sont concernés par l'épissage alternatif [Barash et al. [2010]; Pan et al. [2008]]). Du fait de l'épissage

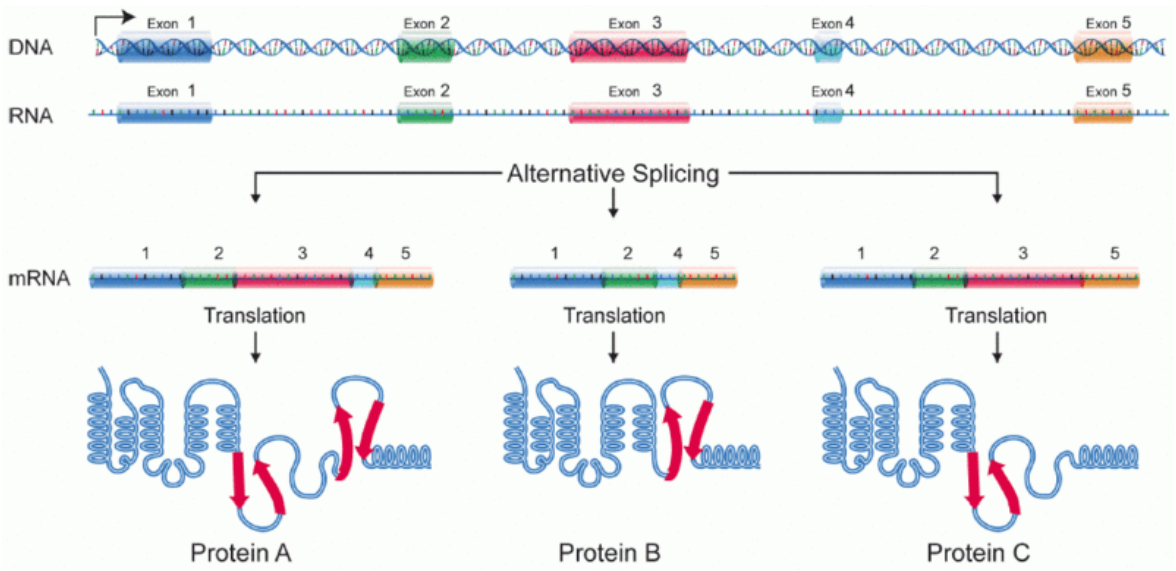


FIGURE 1 – Schéma simplifié : transcription, épissage alternatif et traduction (adaptée d'une figure du NHGRI [2014], domaine public)

alternatif, un même gène conduit couramment à la formation de transcrits constitués de différentes suites d'exons, et donc potentiellement à la synthèse de plusieurs protéines (Figure 1).

La très grande majorité des ARNm, et certains longs ARN non codants, sont également polyadénylés (addition d'une queue polyA en 3') avant l'étape d'épissage. La queue polyA joue un rôle dans stabilité des ARN et, chez les eucaryotes, permet leur transport vers le cytoplasme.

Si les ARN messagers (ARNm) sont les molécules le plus souvent visées par les analyses de transcriptome, de part leur rôle d'intermédiaire pour la synthèse des protéines, il existe d'autres types d'ARN qui sont eux non codants :

- De longs ARN non codants, qui n'entraîneront pas la synthèse d'une protéine.
- Les ARN ribosomiques (ARNr) constituent la plus grande part de l'ARN total d'une cellule (80% chez les mammifères). Ils forment, en association avec des protéines, les ribosomes chargés de la synthèse des protéines à partir des ARNm.
- Les ARN de transfert (ARNt) qui permettent la traduction d'un codon d'un ARNm en acide aminé.
- Divers petits ARN non codants, dont on sait que certains jouent un rôle dans les systèmes de régulation d'expression de différents compartiments du génomes (gènes et éléments transposables).

1.2 Technologies pour l'analyse à large échelle du transcriptome

Différentes technologies permettent l'étude à large échelle des transcriptomes. On peut séparer les puces à ADN, basées sur l'hybridation des séquences pour quantifier l'expression des transcrits, et les technologies de séquençage, en particulier le RNA-seq qui permettent à la fois l'accès à la séquence des transcrits, et, selon la profondeur de séquençage et leur niveau d'expression, un accès plus ou moins précis à la quantification de ceux-ci.

Il existe également des technologies (dites "gène à gène") permettant d'étudier et d'analyser les transcrits et/ou les gènes au cas par cas. C'est le cas des RT-PCR et des RT-PCR quantitatives, qui sont toujours utilisées, en particulier pour valider spécifiquement certains résultats obtenus après analyse des données dites “haut débit”.

1.2.1 Les puces à ADN

La puce à ADN est une technologie développée au cours des années 90 et qui est principalement utilisée afin de quantifier l'expression des gènes (ou transcrits). Il s'agit d'une petite surface (puce) sur laquelle sont fixées plusieurs milliers de molécules d'ADN (appelées sondes) dont la séquence en acide nucléique, ainsi que la position sur la puce, sont connues.

Comme pour la grande majorité des technologies permettant l'étude du transcriptome, elle nécessite au préalable de rétro-transcrire les molécules d'ARN en ADN dit complémentaire (ADNc).

Elle permet, via l'hybridation des sondes fixées sur la puce avec les brins d'ADNc présents dans l'échantillon étudié, de mesurer la concentration relative d'une séquence nucléotidique dans cet échantillon. Celle-ci est mesurée par la fluorescence émise par les brins d'ADNc, marqués avant hybridation. L'analyse des intensités mesurées permet ensuite d'identifier et de quantifier les transcrits présents dans l'échantillon et généralement de comparer plusieurs échantillons.

La principale limite des puces à ADN pour l'étude des transcriptomes vient du fait que cette technologie ne donne pas accès à la séquence des gènes ou des transcrits et qu'elle nécessite des connaissances a priori sur les gènes (ou transcrits) à étudier. Elle est donc peu adaptée pour travailler sur des espèces non modèles.

1.2.2 Le RNA-seq

Le RNA-seq est une approche relativement récente utilisant la seconde génération de séquenceur, appelés NGS (Next Generation Sequencing). Elle est actuellement la méthode la plus utilisée pour les analyses de transcriptome à large échelle et permet d'identifier et de quantifier les transcrits. Elle ne nécessite pas de connaître à priori les séquences des gènes ou des transcrits et peut donc être utilisée dans des études portant sur des espèces non modèles, c'est à dire dont le génome de référence n'est pas disponible.

Il existe plusieurs technologies différentes regroupées sous le terme "NGS". On peut citer les technologies Illumina, 454, Ion Torrent ou encore SOLiD. Celles-ci sont dites à "haut-débit", permettant de produire plus de séquences et à un prix plus bas que le séquençage Sanger, mais les séquences produites sont également plus courtes (Figure 2). Selon la machine utilisée, un *run* Illumina permet par exemple de produire plusieurs dizaines de millions de lectures (voire un peu plus d'un milliard) d'une centaine de nucléotides en moyenne.

Les NGS possèdent plusieurs étapes communes, notamment au niveau de la préparation des librairies d'ADN qui seront séquencées. En effet, même dans le cadre d'études transcriptomiques, c'est de l'ADN qui est lu par les machines. Il est donc nécessaire de rétro-transcrire l'ARN en ADNc au préalable. La préparation des librairies et la suite du séquençage sont alors identiques (Figure 3), que ce soit pour des données génomiques (DNA-seq) ou transcriptomiques (RNA-seq). L'ADN à séquencer est ensuite fragmenté et il y a généralement sélection des fragments selon leur taille. Des adaptateurs sont ensuite ajoutés aux extrémités de chaque fragment. Ces adaptateurs, spécifiques à chaque technologie, permettent l'amplification et la fixation des fragments à séquencer. Dans le cas où plusieurs échantillons différents sont séquencés, les adaptateurs peuvent également permettre l'identification de chaque échantillon (code barre).

L'étape du séquençage est propre à chaque technologie. La plus répandue actuellement est la technologie Illumina (Bentley et al. [2008]; Lister et al. [2008]; Mortazavi et al. [2008]; Nagalakshmi et al. [2008]). Celle-ci permet d'obtenir des lectures assez courtes, aujourd'hui en moyenne de 100 à 150 nt et pouvant atteindre 300 nt. Le séquençage peut concerner une extrémité (single end) ou les deux extrémités (paired-end) des fragments. Il est également possible de préparer des librairies dites brin spécifiques dans lesquelles l'information du brin d'origine des transcrits est conservée.

Le RNA-seq permet, contrairement à la puce à ADN, d'avoir accès à la séquence, au

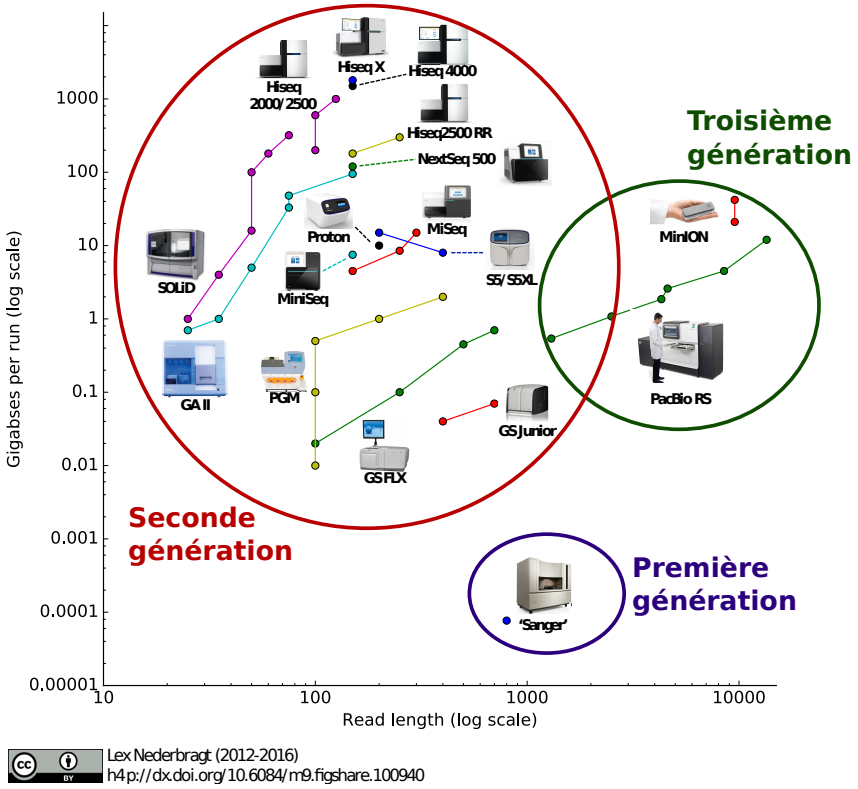


FIGURE 2 – Les différentes technologies de séquençage. Graphique représentant le nombre de lectures obtenues par *run* en fonction de la longueur des lectures pour les différentes technologies de séquençage de première, deuxième et troisième génération. (Adaptée de d'une figure de Lex Nederbragt `http://dx.doi.org/10.6084/m9.figshare.100940`)

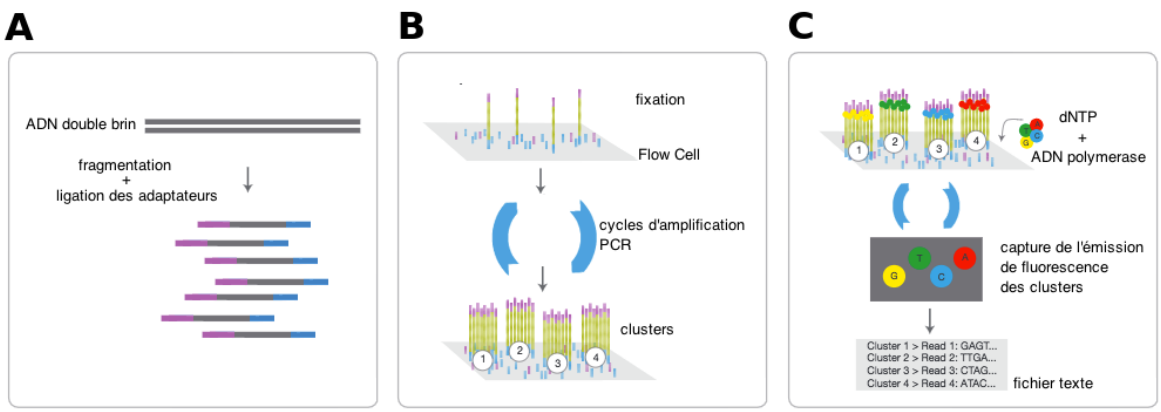


FIGURE 3 – Séquençage illumina. A) L'ADN à séquencer est fragmenté et des adaptateurs sont liées aux extrémités de chaque fragment obtenu. B) Les fragments à séquencer vont être amplifiés par PCR, après avoir été fixés sur la "flow cell" (lame de verre). On obtient des clusters correspondant chacun à un fragment initial et contenant environ 1000 copies de celui-ci. C) Dans chaque cluster, une base est incorporée dans chaque fragment, celle-ci émet une signal fluorescent qui est enregistré. Ces étapes sont répétées jusqu'à séquençage du fragment de taille souhaitée (Adaptée d'une figure de Illumina [2016])

nucléotide près, des transcrits de l'échantillon, et ne nécessite pas de connaissance a priori de ces séquences. Elle permet potentiellement, selon la profondeur de séquençage, d'avoir accès à l'ensemble des transcrits exprimés. La quantification relative des transcrits obtenue à partir de données RNA-seq est également plus précise qu'avec la puce à ADN.

1.2.3 Technologie de séquençage de troisième génération : les longs reads

Une troisième génération de séquenceurs est apparue depuis 2010, celle-ci permet de séquencer entièrement de très longs fragments d'ADN (taille moyenne entre 3kb et 10kb, avec des lectures supérieures à 100kb). Contrairement à la seconde génération de séquenceur elle ne nécessite pas d'amplification des molécules d'ADN en amont du séquençage, ce qui élimine les erreurs produites à cette étape (artefacts d'amplification conduisant à des biais de couverture selon la composition nucléotidique et le taux de GC).

Aujourd'hui il existe principalement deux technologies capables de produire ce type de séquence : Single Molecule Real Time de Pacific Bioscience (PacBio) et MinION d'Oxford Nanopore :

Technologie SMRT de PacBio (Metzker [2010]) : Une molécule d'ADN polymérase est fixée au fond de chaque puits (50 000 puits) dans lequel passe une molécule d'ADN dont le brin complémentaire est synthétisé à partir de nucléotides marqués par fluorescence (quatre couleurs pour les quatre types de nucléotides). Des capteurs intégrés dans les puits permettent de mesurer en temps réel le signal fluorescent émis par l'intégration de chaque nucléotide.

Technologie MinION (Wanunu [2012]) : Les molécules d'ADN sont liées à un premier adaptateur permettant sa prise en charge par une protéine motrice qui va permettre le passage de la molécule d'ADN dans un pore (512 pores par flow cell). Le passage des différents nucléotides dans le pore induit des changements de l'intensité du courant. Chaque nucléotide produit un signal spécifique permettant de déduire la séquence de la molécule d'ADN. Un second adaptateur relie les deux brins complémentaires d'une molécule d'ADN. Les deux brins sont séquencés successivement dans un même pore (séparés par l'adaptateur), ce qui permet d'augmenter la précision du séquençage.

Ces technologies sont déjà utilisées pour le séquençage de génome et permettent, grâce à la longueur des lectures produites, d'améliorer l'assemblage, en particulier des génomes riches en éléments répétés. En effet, la présence d'éléments répétés en forte proportion dans les génomes empêche un bon assemblage des scaffolds. Ce type de séquençage permet aussi de reconstituer des copies d'éléments transposables (ET) dans leur intégralité et d'avoir accès à leur site d'insertion dans le génome. Les séquences produites souffrent en revanche de forts taux d'erreur de séquençage (actuellement entre 4% et 10%, contre 0.1% pour Illumina) lié au séquençage en temps réel (vs séquençage pas à pas pour les NGS), ce qui constitue une de leurs principales limites aujourd'hui. On peut cependant noter que l'évolution de ces technologies est rapide, et on attend dans les années à venir une diminution significative de ces taux d'erreurs.

Quant à une utilisation de ces technologies pour l'étude des transcriptomes, toutes deux restent également limitées par leur faible profondeur de séquençage. Elle produisent en effet autour de 100000 lectures par run, contre plusieurs dizaines de millions (au minimum) pour les machines Illumina, ce qui est problématique pour la quantification des transcrits, dont la précision dépend du nombre de lectures produites, ainsi que pour la détection de variants de séquences ou d'épissages.

1.3 Design expérimental

Il est crucial, avant tout séquençage, de tenir compte des questions biologiques auxquelles on souhaite répondre grâce à l'analyse des séquences obtenues, afin de concevoir en amont un design expérimental adapté.

Il est également nécessaire d'anticiper autant que possible les méthodes utilisées pour l'analyse des données, et de tenir compte des informations déjà à disposition, en particulier l'existence ou non d'un génome de référence. Pour le choix des méthodes et logiciels, les ressources de calcul nécessaires (en temps et en utilisation mémoire) peuvent être limitants.

Parmi les principaux aspects (choix) à considérer dans cette optique, on trouve : le choix de l'ARN extrait (par exemple ARN total, cytoplasmique ou nucléaire), choix de l'ARN que l'on souhaite sélectionner (les ARNm, les petits ARN etc.), le tissu ou type cellulaire à séquencer, le nombre de réplicats biologiques nécessaires, le nombre d'individus nécessaires, la taille des lectures, la profondeur de séquençage (nombre de lectures nécessaires), le choix d'une librairie "brin spécifique", choix de lectures *single* ou *paired-end* etc. Ces différents choix doivent tenir compte à la fois des questions biologiques (*Qu'est ce qu'on veut comme ARN ?*) mais aussi méthodologiques (*Quelle taille des lectures permet un assemblage ?*).

En fonction de la profondeur de séquençage, c'est à dire le nombre de lectures obtenues après séquençage, on pourra analyser les transcrits plus ou moins exprimés : plus on investit dans la profondeur, avec un nombre important de lectures, plus les transcrits faiblement exprimés auront des chances d'être séquencés. La profondeur de séquençage a également un impact important sur la quantification des transcrits et la comparaison de deux conditions : plus la profondeur est importante, plus les différences d'expressions seront "faciles" à détecter. Le nombre de réplicats biologiques est également important dès lors qu'on souhaite comparer plusieurs conditions (population, lignées, tissus, effet d'un traitement etc.). En effet, il faut une certaine puissance statistique pour mettre en évidence une différence d'expression entre deux conditions. Cette puissance augmente avec la profondeur de séquençage et le nombre de réplicats biologiques (pour une taille de l'effet donné). Pour un nombre de lectures total fixe (budget contraint), d'après Liu et al. [2014], l'augmentation du nombre de réplicats permet de détecter plus de gènes différentiellement exprimés entre deux conditions que l'augmentation de la profondeur de séquençage par réplicat (les auteurs observent ce résultat à partir de 10 millions de

lectures par réplicat au sein de la lignée cellulaire MCF7 chez l'humain).

La taille des lectures est quant à elle importante en particulier si on s'intéresse aux répétitions, que ce soit pour leur identification ou leur quantification. En effet, plus la lecture est grande, plus il est facile de l'assigner à une position unique du génome (quand un génome de référence est disponible) ou d'assembler la répétition (quand on ne dispose pas d'un génome de référence).

Le plus couramment, le RNA-seq vise les ARNm, et on cherche à éliminer les ARNr qui constituent la grande majorité des transcrits. Le protocole "polyA+" permet de sélectionner avant séquençage les ARN possédant une queue polyA, c'est à dire la plupart des ARNm ainsi qu'une partie des longs ARN non codants. Certains ARNm sont néanmoins perdus lors de cette sélection. Le protocole Ribo-Zero permet lui d'éliminer les ARNr de l'ARN extrait. Il permet ainsi de garder les autres types d'ARN : les ARNm, les longs ARN non codants et les petits ARN. On sélectionne généralement les ARN d'une taille supérieure à 200 nt pour garder les ARNm et les longs ARN non codants. Il est également possible de sélectionner les petits ARN, on parle alors de *small RNA sequencing*. Il existe d'autres filtres/protocoles permettant de sélectionner des ARN d'intérêt, par exemple des ARN en interaction avec des protéines, comme pour les piRNA (Grentzinger and Chambeyron [2014]).

Lorsque l'on souhaite séquencer plusieurs individus, il est possible de les séquencer séparément, et d'utiliser au moment de la création de la librairie, un code-barre unique dans la séquence des adaptateurs pour différencier chaque individu. Ce bar-coding a néanmoins un coût et on peut faire le choix de ne pas conserver l'information de la provenance des séquences en y renonçant, les individus sont dits "poolés". Dans le cas des expériences de RNAseq, souvent, il est nécessaire d'extraire des RNA à partir de plusieurs individus de façon à obtenir suffisamment de matériel. Par exemple, dans le cas du séquençage des transcrits issus d'ovaires de drosophiles (analysés dans le chapitre suivant), nous avons extrait en moyenne 200 ng d'ARN par paire d'ovaires (ensuite converti en ADNc), alors que les plateformes de séquençage demandent généralement un minimum de 1 µg d'ADN ou d'ADNc (souvent plus).

Les NGS, et donc le RNA-seq, ne sont pas des technologies parfaites, sans biais. Des erreurs de séquençages sont possibles (moins de 1%, voire 0.1% chez Illumina)et leur position est souvent fonction de la composition nucléotidique (Dohm et al. [2008]; Hansen et al. [2010]). On observe également une variabilité de la profondeur de séquençage liée à

des sites de fragmentations préférentiels, à nouveau selon la composition nucléotidique (Sendler et al. [2011]). Le profil de couverture le long d'un ARNm sera donc hétérogène.

2 L'analyse de données RNA-seq

Les données RNAseq permettent l'analyse des transcriptomes, que ce soit au niveau de leur séquence ou de leur abondance. Ces données nous permettent d'avoir accès à l'identification et la quantification les gènes exprimés dans l'échantillon étudié.

Dans la plupart des cas, l'épissage alternatif produit plusieurs types de transcrits matures par gène (Figure 1). Il est possible d'approfondir l'analyse à l'échelle des transcrits, et d'identifier les variants d'épissages alternatifs appartenant au même gène. Dans la pratique, l'association des différents variants d'épissages alternatifs pour reconstruire l'ensemble des transcrits présents dans l'échantillon reste un problème complexe et ce même avec l'utilisation d'un génome de référence.

Le RNA-seq permet également d'avoir accès aux autres variations nucléotidiques (par exemple les SNP ou les indels). Ces variations peuvent avoir lieu au sein des génomes ou bien pendant/après la transcription (dans ce cas on parle de RNA editing).

Pour l'analyse de données en RNAseq, on peut globalement séparer les méthodes existantes en deux catégories : celles basées sur l'alignement des lectures sur un génome (ou un transcriptome) de référence, et celles basées sur l'assemblage *de novo* des lectures.

2.1 Méthodes d'alignements des lectures

Dans le cas où un génome de référence est disponible, les méthodes basées sur l'alignement sont les plus utilisées : on assigne une position génomique aux lectures en les alignant directement sur celui-ci. L'identification de transcrits et leur quantification, mais aussi la détection de variants dépendent fortement de la qualité de l'alignement sur le génome de référence, et donc de la qualité du génome de référence.

Une des spécificités, et difficulté, de l'alignement des données RNA-seq est que, du fait de l'épissage, certaines lectures correspondent à des jonctions de deux exons et s'alignent donc en deux blocs (ou plus) sur le génome, séparés par au moins un intron. Il existe plusieurs aligneurs dédiés aux données RNA-seq et permettant de tenir compte de cette caractéristique et d'aligner les lectures générées par des séquenceurs haut-débit en un temps raisonnable. Ceux-ci peuvent être répartis en différents groupes, correspondants à

des approches différentes. Les méthodes dites *exon-first* cherchent d'abord à aligner les lectures en un seul bloc sur le génome. Cette étape permet de définir les exons. Elles utilisent ensuite les lectures non alignées pour trouver les jonctions entre les exons. On peut notamment citer TopHat (Trapnell et al. [2009]), MapSplice (Wang et al. [2010]), SpliceMap (Au et al. [2010]), SOAPsplice (Huang et al. [2011]) , PASSion (Zhang et al. [2012]), GEM (Marco-Sola et al. [2012]) qui sont basées sur cette idée. Les méthodes *seed-and-extend*, vont elles chercher à aligner une partie de la lecture (*seed*) en un bloc, puis à étendre cet alignement. Parmi ces méthodes on peut citer GSNAP (Wu and Nacu [2010]), STAR (Dobin et al. [2013]) ou plus récemment HISAT et HISAT2 (Kim et al. [2015]). Ces méthodes permettent généralement d'identifier plus facilement de nouvelles jonctions d'épissages. Il existe également d'autres types d'approches. On peut citer CRAC (Philippe et al. [2013]) qui utilise le profil en *k-mers* (mots de taille k) des lectures pour leur assigner une position génomique.

Par ailleurs, certaines de ces méthodes d'alignement utilisent les annotations du génome comme guide, ou s'appuient sur la recherche de motifs spécifiques pouvant correspondre au début ou à la fin d'un intron. Celles-ci pourront identifier plus précisément les jonctions déjà connues. D'autres comme CRAC sont moins contraintes et sont ainsi plus performantes quant à la détection des nouveaux épissages (non annotés).

Une autre difficulté en RNA-seq (comme en DNA-seq) concerne la gestion des alignements dits "multiples", lorsqu'une lecture peut être assignées à différents endroits du génome. Certains aligneurs comme Bowtie ou Bowtie2 (sur lequel s'appuie TopHat) proposent dans ce cas plusieurs solutions :

a) recenser tous les alignements valides, c'est à dire ceux qui s'alignent selon les paramètres demandés par l'utilisateur (par exemple moins de 3 mismatchs)
b) recenser tous les alignements optimaux, c'est-à-dire parmi les alignements valides celui ou ceux qui ont le meilleur score
c) recenser les N premiers alignements parmi les alignement valides
d) choisir aléatoirement un alignement parmi tous ceux qui sont optimaux.

Par défaut, c'est généralement cette dernière qui est implémentée ou choisie. Elle correspond à l'hypothèse que toutes les copies d'une répétition ont le même niveau d'expression. Ce choix silencieux est rarement discuté. Pour l'étude des éléments répétés il est crucial de le questionner.

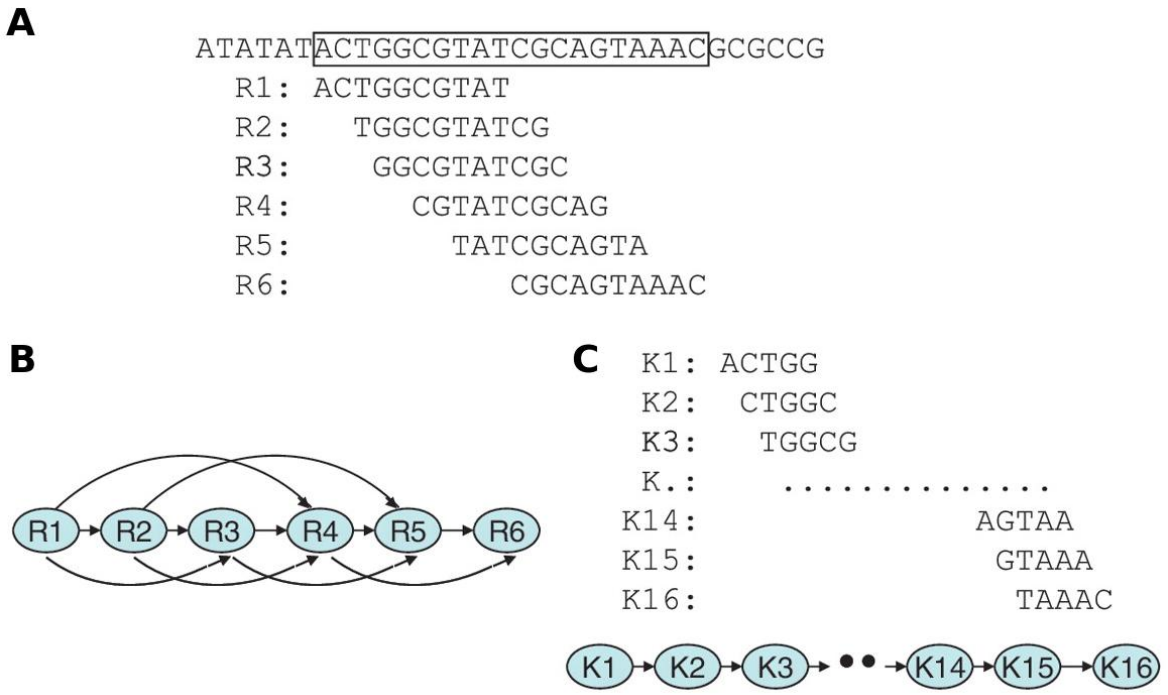


FIGURE 4 – Principe de construction d'un graphe d'overlap et d'un graphe de de Bruijn à partir d'une séquence de 20 nt et de 6 lectures générées par le séquençage de cette région (A). B) Dans le graphe d'overlap, chaque lecture (R1 à R6) est représentée par un nœud du graphe, et les nœuds sont reliés s'ils se chevauchent d'au moins 5 nt. C) Les lectures sont découpées en mot de taille 5. On obtient 16 mots différents. Chaque mot n'est représenté que par un seul nœud. Deux mots ayant un chevauchement excat de taille 4 sont reliés. (Adaptée d'une figure de Li et al. [2012])

2.2 Méthodes d'assemblage (sans génome de référence)

Lorsque aucun génome ou transcriptome de référence n'est disponible, il est possible d'assembler les lectures pour reconstruire les transcrits présents dans l'échantillon séquencé. On parle d'assemblage de novo.

Ce type de méthode permet notamment l'analyse de données RNA-seq dans le cadre d'espèce dites "non-modèles" pour lesquelles il n'y a pas de génome de référence. L'utilisation de méthodes d'assemblage est également pertinent chez les espèces modèles lorsque l'on souhaite identifier de nouveaux gènes, de nouveaux variants d'épissages ou dans certains cas particuliers, lorsque le génome de référence est trop différent de celui étudié, comme cela peut-être le cas dans des cellules cancéreuses. L'assemblage consiste en l'utilisation des chevauchements entre les lectures afin de reconstituer les séquences.

Il existe différentes méthodes d'assemblage et celles-ci sont généralement basées sur deux types de graphes : les graphes d'overlap et les graphes de de Bruijn (Figure 4).

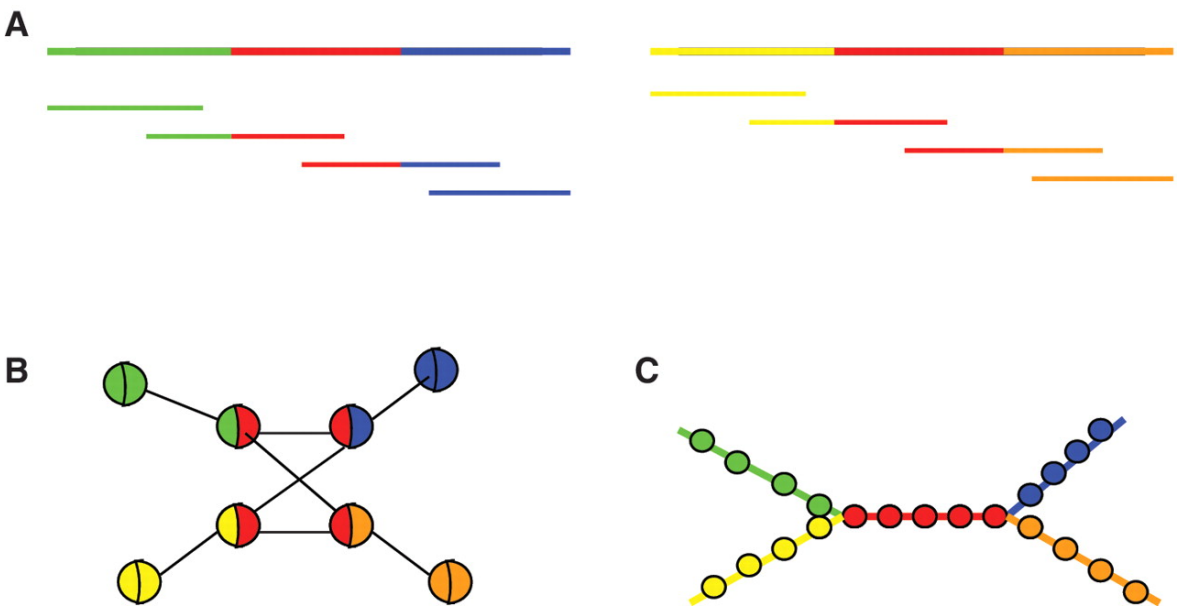


FIGURE 5 – Les répétitions dans le graphe d'overlap et graphe de de Bruijn. A) Deux transcrits (ou régions génomiques) partagent une région répétée (en rouge). B) Les lectures contenant la répétition créent une zone plus fortement connectée. C) Le graphe de de Bruijn correspondant. Les k-mers de la région répétée ne sont représentés qu'une fois. (Figure de Li et al. [2012])

Les graphes d'overlap (dit Overlap-layout-consensus) ont été les premiers utilisés pour l'assemblage des lectures issues de la première génération de séquenceurs. Parmi les assembleurs qui se basent sur ce type de graphe on peut citer Arachne (Batzoglou et al. [2002]), Celera Assembler (Myers et al. [2000]), CAP3 (Huang and Madan [1999]), PCAP (Huang and Yang [2005]), Phrap (Bastide and McCombie [2007]) et Phusion (Mullikin and Ning [2003]). La construction d'un graphe d'overlap est assez intuitive : chaque lecture obtenue par séquençage est représentée par un nœud, et deux lectures sont reliées par une arête si elles se chevauchent de plus de T nucléotides (Figure 4 A et B). La construction d'un tel graphe nécessite donc la comparaison de chaque lecture deux à deux, et il n'est donc pas adapté au traitement des données NGS. En effet, l'augmentation du nombre de lectures permise par les NGS entraîne une augmentation importante des ressources informatiques nécessaires à la construction du graphe d'overlap (temps et mémoire).

Depuis l'arrivée des NGS, les méthodes développées pour l'assemblage des lectures s'appuient donc davantage sur le graphe de de Bruijn. La construction de celui-ci est moins intuitive que le précédent. Les lectures sont d'abord découpées en mots de taille k appelés k-mers (en général compris entre 25 et 50 bp selon les méthodes). Dans un graphe de de Bruijn, chaque k-mer est représenté par un nœud du graphe. Deux nœuds

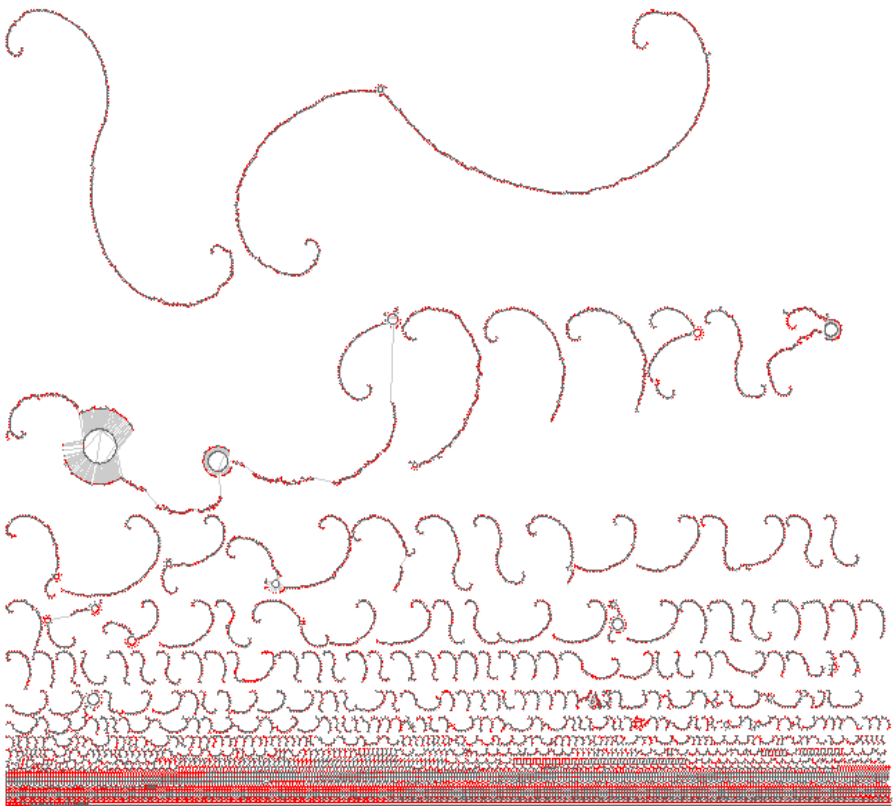


FIGURE 6 – Exemple de graphe de de Bruijn construit à partir de 100000 lectures issues du séquençage transcriptomique de *D. mojavensis*. Les transcrits fortement exprimés (donc fortement couverts) sont plus faciles à assembler, tandis que les transcrits les moins exprimés sont fragmentés.

sont reliés par une arête si les k-mers correspondant se chevauchent de k-1 nucléotides. Ce graphe a l'avantage de représenter explicitement chaque nucléotide. Plusieurs assembleurs génomiques utilisent le graphe de de Bruijn, on peut citer Euler-USR (Chaisson et al. [2009]), Velvet (Zerbino and Birney [2008]), ABySS (Simpson et al. [2009]) et SOAPdenovo (Li et al. [2010]). Il existe également des assembleurs dédiés à l'assemblage transcriptomique : Trinity (Grabherr et al. [2011]), Oases (Schulz et al. [2012]), SOAPdenovo-Trans (Xie et al. [2014]) ou encore Trans-ABySS(Robertson et al. [2010]). Dans l'idéal, lorsque l'on assemble des lectures RNA-seq, on espère qu'un chemin du graphe de de Bruijn corresponde à un transcrit (Figures 4 et 6). En réalité, du fait des erreurs de séquençage, des répétitions, du manque de couverture et de l'épissage alternatif, les transcrits assemblés ne sont pas toujours complets ou exacts (Figure 5, 6 et 7).

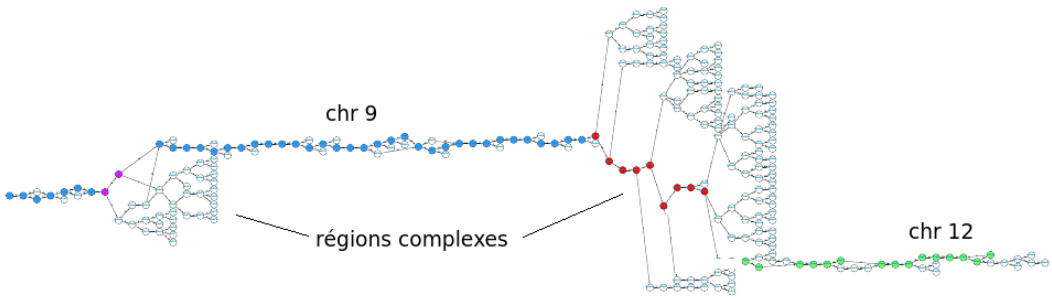


FIGURE 7 – Exemple d'erreur d'assemblage dans un graphe de de Bruijn. Du fait d'une répétition commune aux chromosomes 12 et 9 chez l'homme, l'assembleur choisi un mauvais chemin parmi les multiples possibilités.

2.3 Reconstruction des transcrits et épissages alternatifs

Les deux approches présentées précédemment sont utilisées pour tenter de reconstruire les transcrits complets à partir des lectures séquencées.

Les méthodes basées sur l'alignement des lectures comme Cufflinks (Trapnell et al. [2010]), Scripture (Guttman et al. [2010]), StringTie (Pertea et al. [2015]), FlipFlop (Bernard et al. [2014]) ou SLIDE (Li et al. [2011]) utilisent également des graphes pour la reconstruction des transcrits. Cufflinks construit un graphe d'overlap à partir des lectures qui s'alignent sur un locus du génome. Ce graphe est ensuite parcouru pour reconstruire les transcrits, en considérant le plus petit ensemble d'isoformes permettant d'expliquer les lectures. Scripture et StringTie construisent eux un graphe d'épissage : les nœuds représentent des exons, ou morceaux d'exons et les arrêtes les variations d'épissages.

Les méthodes comme Trinity ou Oases utilisent directement l'assemblage des lectures à partir d'un graphe de de Bruijn pour reconstruire les transcrits. Les principales difficultés de ces méthodes concernent les régions répétées (qui créent des régions complexes dans lesquelles il est difficile de choisir le bon chemin) et les régions faiblement couvertes (qui créent des trous aboutissant à l'assemblage partiel ou fragmenté de ces transcrits).

Que ce soit à partir de lectures alignées sur une référence ou par assemblage *de novo*, la reconstruction complète du transcriptome à partir de lectures courtes reste un problème difficile. Il est cependant moins complexe d'identifier les variants d'épissages alternatifs de manière locale, sans chercher à reconstruire les transcrits complets. C'est ce que proposent les méthodes dites locales, qu'elles soient basées sur des approches d'alignement ou d'assemblage. On peut par exemple citer Miso (Katz et al. [2010]), MATS (Shen et al. [2012]) et CRAC (Philippe et al. [2013]) basés sur l'alignement des lectures sur un

génome de référence, ou KisSplice (Sacomoto et al. [2012]), basé sur l'assemblage des lectures.

2.4 Identification des variants nucléotidiques

Les données RNA-seq permettent l'accès aux séquences au nucléotide près. Aussi il est possible de détecter des SNP et des indels présents dans le transcriptome séquencé. Les SNP (Single Nucleotide Polymorphisme), sont des variants d'un nucléotide de type substitution, ce sont les variants les plus présents dans les génomes, et représentent chez l'Homme 90% de l'ensemble des variants génétiques (Collins et al. [1998]). Leur impact est variable et dépend de leur position. Dans les régions codantes les variants peuvent ne pas avoir un impact direct sur la séquence en acide aminé des protéines, du fait de la redondance du code génétique (variant synonyme ou non-synonyme). Ils peuvent également impacter l'expression des gènes, par exemple dans des régions promotrices ou régulatrices.

Les SNP peuvent être détectés via des méthodes s'appuyant sur l'alignement des lectures contre un génome de référence, comme GATK (McKenna et al. [2010]), SAMtools mpileup (Li [2011]), SNVer (Wei et al. [2011]), MAQ (Li et al. [2008]) ou encore CRAC (Philippe et al. [2013]) . Elle détectent, sur un ensemble de lectures alignées, les nucléotides qui différent de la référence, et proposent des filtres permettant d'éliminer les différences observées trop rarement, qui ont plus de chances de correspondre à des erreurs de séquençage. Certaines méthodes comme MAQ, on été pensées pour l'analyse de données DNA-seq d'un individu diploïde. Elles ne sont pas appropriées aux données DNA-seq poolées, RNA-seq (poolées ou non) puisqu'elles s'attendent à trois génotypes différents pour la position donnée (l'individu peut être hétérozygote, homozygote comme la référence, ou homozygote différent de la référence, c'est-à-dire fréquence allélique observée : 0, 0.5 ou 1). En RNA-seq même lorsque le séquençage concerne un seul individu hétérozygote pour une position donnée, du fait de l'expression allèle spécifique la fréquence allélique exprimée peut être assez différente de 0.5.

Il est également possible de détecter les SNP via des méthodes basées sur la représentation des lectures sous forme de graphe. Les SNP produisent en effet un motif particulier dans un graphe de de Bruijn : une "bulle" de $2k-1$ nucléotides (cf Chapitre 3)

2.5 Accès à la quantification

Le RNA-seq permet également d'estimer l'abondance relative des transcrits exprimés. En effet, on peut supposer que le nombre de lectures provenant d'un transcrit est proportionnel à son expression. Si on retrouve le même génome dans toutes les cellules d'un organisme, les gènes ne s'expriment pas de la même façon selon les types cellulaires et les tissus, mais aussi des conditions physiologiques dans lesquelles elles se trouvent (âge, traitement, stress, etc.)

2.5.1 Méthodes d'alignement et comptage

Les méthodes les plus répandues sont celles basées sur l'alignement des lectures contre le génome de référence ou contre le transcriptome assemblé.

Il y a peu de difficulté particulière à quantifier l'expression des gènes ou des exons lorsqu'on aligne sur le génome de référence. Il "suffit" de compter les lectures s'alignant sur une portion de génome. C'est notamment ce que propose HTseq count (Anders et al. [2014]).

Si l'on veut avoir accès à l'expression des transcrits, la première difficulté sera en amont l'identification des transcrits issus d'un même gène. Néanmoins, certains outils comme StringTie et FlipFlop tiennent compte de l'abondance des transcrits pour réaliser leur reconstruction. La quantification et la reconstruction ont donc lieu simultanément.

Si l'on souhaite obtenir l'expression des transcrits à partir des lectures alignées sur transcriptome de référence/assemblé, la principale difficulté concerne la gestion et le comptage de lectures issues d'exons communs à plusieurs transcrits alternatifs. Des méthodes comme RSEM (Li and Dewey [2011]) ou eXpress (Roberts and Pachter [2013]) permettent de tenir compte de l'alignement multiple. Elles se basent sur les alignements effectués par des aligneurs comme Bowtie (Langmead et al. [2009]), Bowtie2 (Langmead and Salzberg [2012]), STAR (Dobin et al. [2013]) qui doivent être paramétrés de manière à reporter l'ensemble des alignements valides. Ces méthodes de quantification vont choisir à la place de l'aligneur utilisé, le "meilleur" alignement en cas d'alignement multiple.

2.5.2 Autre type de méthodes de quantification

Il existe également des méthodes de quantifications qui ne sont pas basées sur de l'alignement de séquence, mais sur l'utilisation et/ou le comptage des k-mers : Sailfish

(Patro et al. [2014]), RNA-skim (Zhang and Wang [2014]), Kallisto (Bray et al. [2016]). Sailfish, la première méthode de ce type, aligne non pas les lectures mais les k-mers sur les transcrits afin de les quantifier. Si elle s'avère bien plus rapide que les méthodes comme RSEm ou eXpress, la quantification qu'elle propose est cependant moins précise. RNA-skim propose alors d'identifier et d'utiliser certains k-mers spécifiques d'un transcrit, appelés sig-mers, pour une meilleure quantification des transcrit. Kallisto est quant à lui basé sur ce que les auteurs appellent du "pseudoalignement" : les transcrits sont utilisés pour construire un graphe de de Bruijn, et à chaque k-mer on assigne une *k-compatibility class*, c'est à dire qu'on l'associe à un ou plusieurs transcrits, et on cherche ensuite à savoir quels transcrits sont compatibles avec les lectures (découpées en k-mers). Ces deux dernières méthodes sont bien plus rapides, demandent moins de ressources de calcul et avec des résultats similaires aux méthodes d'alignement puis comptage. Kallisto et RNAskim permettent ainsi de traiter plusieurs dizaines de millions de lectures en moins d'une dizaine de minute sur un ordinateur portable.

2.6 Comparer deux (ou plus) échantillons

L'objectif d'une analyse différentielle est de tester si l'expression des gènes (ou des transcrits) est modifiée entre deux conditions. De nombreuses méthodes permettent de comparer deux (ou plus) échantillons en tenant compte de la variabilité biologique (utilisation de réplicats biologiques). DESeq (Anders and Huber [2010]) et edgeR (Robinson et al. [2010]) sont aujourd'hui les plus utilisées.

Avant de comparer deux échantillons, il faut les rendre comparables en normalisant les comptages. Cette normalisation tient compte, entre autres, du nombre de lectures (utilisées pour la quantification) par échantillon : on ne veut pas observer une différence qui serait uniquement due à un séquençage plus profond dans un échantillon, ou à un meilleur alignement des lectures d'un échantillon par rapport à un autre. DESeq propose une normalisation par les médianes : on considère que le médianes sont les mêmes pour deux (ou plus) échantillons comparés. L'hypothèse posée au départ est que, pour la plupart des gènes, l'expression ne varie pas. Cette hypothèse est également à la base des normalisations proposées par edgeR.

DESeq et edgeR utilisent une distribution binomiale négative pour modéliser les comptages et font ensuite un test d'expression différentiel qui compare l'expression pour chaque gène. Une *p-value* est associée à chaque gène testé, elle correspond à la probabilité que la

différence observée entre les conditions ne soit pas plus extrême que celle attendue sous l'hypothèse nulle de distribution identique des comptages dans les deux conditions. La taille de l'effet, c'est-à-dire le ratio des comptage entre les deux conditions, est également proposée en sortie de ces méthodes.

3 Les répétitions et éléments transposables

On sépare généralement les répétitions dans les génomes en deux grandes catégories, d'une part les répétition en tandem (incluant l'ADN satellite, minisatellites, microsatellites) et d'autre part les éléments transposables (ET). On peut également inclure les familles de gènes paralogues (issus d'une duplication).

Les éléments transposables ont été découverts par Barbara McClintock chez le maïs à la fin des années 40 (McClintock [1950]). Ce sont des séquences d'ADN répétées présentes dans presque tous les organismes, eucaryotes et procaryotes. Ils sont caractérisés par leur capacité à transposer, c'est à dire à se déplacer et se multiplier au sein du génome, et ils codent généralement pour les protéines nécessaires à leur mouvement.

Selon les organismes ils peuvent représenter une part importante du génome ; de 3% chez *S. cerevisae* (Kim et al. [1998]), en passant par 45% chez l'Homme (Makalowski [2001]), et jusqu'à plus près de 90% chez le maïs (SanMiguel et al. [1996]; Schnable et al. [2009]). De part leur caractère répété, ils contribuent à la taille des génomes. Chez les eucaryotes celle-ci est d'ailleurs bien corrélée à la proportion d'ET dans les génomes (Biémont and Vieira [2003]; Chénais et al. [2012]; Kidwell [2002]; Lynch and Conery [2003]) Qualifiés dans un premier temps d'ADN « poubelle » et longtemps considérés comme inutiles, de nombreuses études ont depuis montré l'impact des ET sur leur génome hôte.

L'activité des éléments transposables a en effet un impact considérable sur le génome hôte et son évolution (Biemont and Vieira [2006]; Casacuberta and Gonzalez [2013]; Feschotte and Pritham [2007]), leur mobilité étant cause de mutations pouvant entraîner toute une panoplie d'effets. Ceux-ci dépendent à la fois du lieu d'insertion (exon, intron, UTR, région inter-génique etc.) et de la séquence de l'ET. Il peut altérer l'expression d'un gène, créer de nouveaux exons ou introns, modifier l'épissage alternatif, donner naissance à un codon stop prématuré, etc. (Kidwell and Lisch [2000]). Par ailleurs, la présence d'ET peut altérer l'état de la chromatine et le degré de méthylation de l'ADN, ce qui peut affecter la transcription des gènes voisins et modifier ainsi leur expression. De plus, du fait

de leur caractère répété, les ET peuvent également être à l'origine de recombinaisons ectopiques (Hughes and Coffin [2005]) et provoquer ainsi des délétions (van de Lagemaat et al. [2005]), des inversions (Cáceres et al. [1999]; Sniegowski and Charlesworth [1994]), des duplications (Mishra [2008]) ou d'autres réarrangements chromosomiques (Bourque [2009]; McClintock [1950]). Si ces mutations sont souvent neutres, elles peuvent aussi avoir des effets délétères ou plus rarement avantageux pour l'organisme.

Chez l'Homme on dénombre aujourd'hui plus de 124 insertions d'ET liées au développement de certaines maladies, notamment des cancers (Hancks and Kazazian [2016]). Par exemple, des recombinaisons non homologues d'ET (principalement de type Alu et L1) et la perte de séquences génomiques contribuent à l'apparition de cas de leucémie, sarcome, hépatome, cancer du sein, ainsi que des maladies génétiques (Callinan and Batzer [2006]; Chen et al. [2005]; Chénais [2015]). Par ailleurs, ces mutations peuvent avoir lieu dans les cellules germinales (ou dans les première étapes du développement) affectant ainsi la génération suivante mais également dans les cellules somatiques (cancers).

Les exemples de mutations avantageuses ou de domestication moléculaire d'ET, bien que plus rares, ne sont plus anecdotiques. La présence et l'activité des ET peuvent alors être vues comme une source de variabilité génétique qui pourra être travaillée par la sélection naturelle (Biemont and Vieira [2006]). Plusieurs exemples, chez différentes espèces animales et végétales, montrent que leur domestication peut-être à l'origine de phénotypes adaptatifs (Casacuberta and Gonzalez [2013]; Lisch [2013]). Ainsi, chez la drosophile on observe plusieurs cas d'insertion d'ET conférant un avantage évolutif à l'hôte : adaptation au climat tempéré (González et al. [2008]), résistance aux pesticides (Mateo et al. [2014]), résistance au stress oxydatif (Guio et al. [2014]) etc. Un autre exemple de domestication est celui des gènes codant pour des protéines appelées syncytines. Ces dernières proviennent de gènes *env* de rétrovirus endogènes et sont indispensables au développement du placenta chez l'Homme et de façon plus générale chez les mammifères (Dupressoir et al. [2009]; Mi et al. [2000]; Villesen et al. [2004]).

On distingue deux grandes classes d'éléments transposables selon la nature de leur intermédiaire de transposition. Les éléments transposables qui transposent via un intermédiaire à ARN (rétrotransposons) constituent la classe I. L'ARNm de l'élément est ensuite réverse-transcrit en ADN grâce à une reverse transcriptase et inséré dans un nouveau locus sur le génome. Ces éléments fonctionnent sur le principe du « copier-coller » et peuvent ainsi être présents en un grand nombre de copies dans le génome hôte. Ils

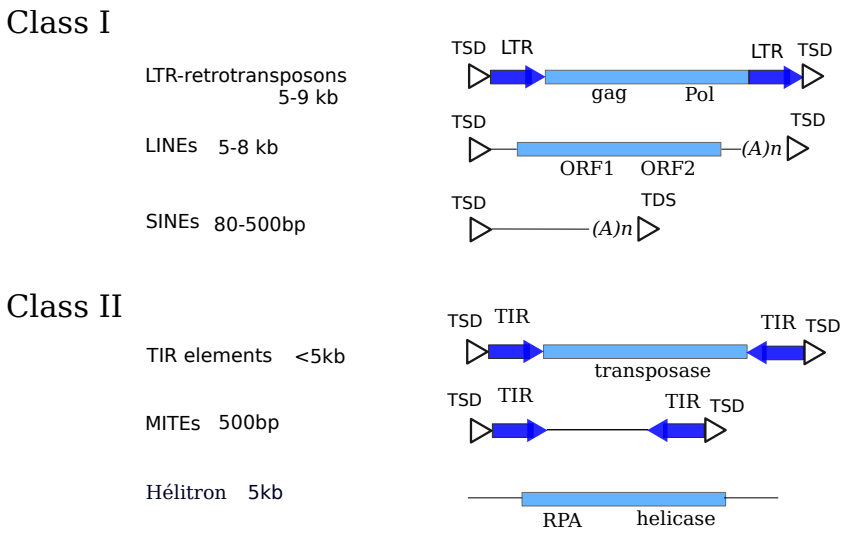


FIGURE 8 – Structure des différents types d'éléments transposables selon leur mode de transposition. La plupart des ET possèdent une duplication du site cible, *Target Site Duplication* (TSD) parfois caractéristique du type d'élément. Le gène *pol* contient plusieurs domaines protéiques codant pour une protéase, une intégrase, une réverse-transcriptase et une RNaseH, permettant à l'élément de transposer. Les éléments TIR et MITE possèdent des répétitions terminales inversées (TIR), mais seuls les éléments MITE ont perdu la capacité de transposer de manière autonome. Les Hélitrons produisent des protéines *RAP* qui leur permettent de se lier à de l'ADN simple brin, ainsi qu'une hélicase. (Adaptée de Casacuberta and Gonzalez [2013])

peuvent être répartis en deux sous-classes : les éléments dits « à LTR » (pour Long Terminal Repeats) qui sont encadrés par de longues répétitions non inversées, et les éléments qui n'en possèdent pas, les LINEs et SINEs. Les ET qui transposent via un intermédiaire à ADN, selon un processus de « couper-coller », constituent la classe II. C'est le cas des éléments de type TIR, MITE ou Hélitron (Figure 8).

Comme les gènes classiques, les ET sont soumis à des régulations diverses. Celles-ci permettent de restreindre le nombre de copies dans le génome, et de contrebalancer l'augmentation du nombre de copies liées à la transposition, ainsi que les effets délétères de la transposition. Les ET sont notamment la cible de régulations épigénétiques transcriptionnelles telles que les méthylations de l'ADN ou les modifications d'histone, mais aussi de régulations post-transcriptonnelles par des petits ARN. Les *Piwi-interacting RNA*, ou piRNA, sont des petits ARN non codants (24 à 29 nt) qui interfèrent directement avec les transcrits des ET et aboutissent à leur dégradation (Saito and Siomi [2010]; Siomi et al. [2011]).

L'étude des éléments transposables a été facilitée par l'arrivée des NGS. Cependant,

si ces techniques de séquençage ouvrent une nouvelle voie pour l'étude des génomes et notamment des ET, leur identification reste une tâche non triviale nécessitant le développement de nouvelles méthodes informatiques capables de tenir compte des spécificités liées aux données NGS.

Les problèmes méthodologiques soulevés ci-après concernent les ET mais sont également valables pour les autres types de répétitions transcrites, ainsi que pour les familles de gènes paralogues.

Les répétions sont aujourd'hui (encore) souvent exclues des analyses bio-informatiques de séquençage, que ce soit en DNA-seq (les répétitions sont "masquées") ou en RNA-seq. Il existe cependant un nombre considérable de programmes dédiés à l'étude des ET (Lerat [2010]).

3.1 Analyse bio-informatique des éléments transposables en génomique

Il existe différents types de méthodes permettant d'identifier les ET dans les génomes déjà assemblés. Plusieurs méthodes permettent l'identification des ET par similarité de séquence avec ceux déjà connus ou via la recherche de caractéristiques spécifiques à certains types d'ET (recherche des domaines protéiques gag et pol ou motifs particuliers). Le programme le plus utilisé est RepeatMasker ; il est lui même assez souvent intégré dans d'autres programmes ou pipelines. On peut également citer LTR HARVEST (Ellinghaus et al. [2008]) qui permet de détecter des rétrotransposons à LTR, ou encore MITE-Hunter (Han and Wessler [2010]) pour détecter des éléments transposables de type MITE (Miniature Inverted-repeat Transposable Elements). Ce type de méthodes est néanmoins limité par nos connaissances des ET recherchés et la présence de caractéristiques stables chez ceux-ci. Aussi, ces programmes ne sont pas adaptés à la recherche de nouveaux types d'ET (Lerat [2010]).

Il existe également des méthodes dites *de novo* utilisant la propriété répétée des ET pour les détecter dans les génomes ou données génomiques. Certaines peuvent s'appliquer à un génome assemblé (Piler [Edgar and Myers [2005]] ou Recon [Bao and Eddy [2002]]) et leur sensibilité sera dépendante de la qualité de l'assemblage du génome. D'autres cherchent les répétitions directement dans les données brutes de séquençage (non assemblées) et utilisent les graphes d'overlap comme AAARF (DeBarry et al. [2008]) et RepeatExplorer (Novak et al. [2010]) ou les graphes de de Bruijn comme ReAS (Li et al. [2005]), DNAPipeTE (Goubert et al. [2015]), Tedna (Zytnicki et al. [2014]), RepARK (Koch et al.

[2014])

3.2 Analyse bio-informatique des éléments transposables en RNAseq

La transcription des ET étant une des étapes du cycle de transposition, le taux de transcription des ET est un bon indicateur de leur activité, bien qu'il ne soit pas directement lié à la transposition. On sait en effet que la transcription d'un ET n'est pas suffisante pour sa transposition car l'étape d'insertion n'a pas toujours lieu mais aussi du fait de régulations post-transcriptionnelles par des petits ARN (Brennecke et al. [2007]).

Les ET étant généralement sous contrôle dans un génome hôte, leur niveau d'expression est faible. Il est donc nécessaire d'avoir un nombre de lecture suffisante pour identifier les ET peu exprimés.

Tout comme pour les gènes classiques, deux stratégies sont possibles pour détecter les ET en RNA-seq :

(1) Si un génome de référence fiable, correctement assemblé et annoté est disponible, les lectures séquencées peuvent être alignées contre celui-ci.
(2) Dans le cas contraire, on procède à l'assemblage *de novo* des lectures (grâce à des assembleurs dédiés aux données RNA-seq) puis, à partir des résultats de l'assemblage, on peut identifier les ET avec des méthodes basées sur la similarité de séquence ou sur la recherche de domaines ou motifs conservés. Il n'existe pas de méthode spécifique à la détection de novo d'ET pour les données RNA-seq et celles développées pour des données DNA-seq, basées sur leur propriété répétée, ne peuvent pas être utilisées.

Dans les deux cas, du fait de la faible taille des lectures et de la similarité des copies, l'étude des ET (identification et quantification) peut difficilement se faire par copie (c'est à dire pour chaque insertion) mais plutôt par famille. Les copies proches (peu divergentes), issues d'une même famille d'ET seront quantifiées ensemble.

Il est aussi possible d'étudier les variants nucléotidiques entre copies d'une même famille, de la même manière que pour les gènes. Il est en revanche difficile (voire impossible) de préciser si les variations observées proviennent d'une différence de deux copies d'une même famille, ou s'il s'agit un variant polymorphe. Dans le cas de données "poolées" cette difficulté est encore plus importante.

3.2.1 Espèces modèles et méthodes d'alignement

Si l'on dispose d'un génome de référence pour l'espèce étudiée, il est possible d'analyser les ET grâce à des méthodes basées sur l'alignement. Néanmoins, du fait du caractère répété des ET, les problèmes de lectures s'alignant à plusieurs positions génomiques seront plus fréquents et donc plus problématiques pour l'étude des ET.

Selon la divergence entre copies, les lectures provenant d'une certaine copie d'une famille d'ET peuvent s'aligner de manière optimale sur plusieurs copies de cette famille, voire sur plusieurs familles différentes. Il est plus simple d'étudier les ET à l'échelle des familles, en regroupant toutes les lectures qui s'alignent sur un ensemble de copies de la même famille. TEtools (cf. Annexes, section 2) et TEtranscripts (Jin et al. [2015]) proposent par exemple de quantifier les ET à l'échelle des familles.

TEtranscripts, comme certaines méthodes de quantification évoquées précédemment (RSEM et eXpress), se basent sur les alignements fournis par un aligneur (les auteurs proposent l'utilisation de STAR) en demandant à garder l'ensemble des alignements valides pour chaque lecture. Il est nécessaire de fournir un fichier d'annotation de type GTF pour les gènes et un autre pour les ET, et TEtranscripts propose en sortie une quantification des gènes et des familles d'ET.

TEtools propose d'aligner les lectures uniquement sur les copies d'ET. Le module TEcount de TEtools utilise un fichier appelé *rosette* qui fait le lien entre les différentes copies d'une même famille d'ET et permet donc d'obtenir des comptage au niveau des familles. Si une lecture s'aligne sur plusieurs copies, l'assignation à l'une d'elle se fait de manière aléatoire. La quantification de l'expression des ET se fait ici aussi par famille et non pas par copie.

3.2.2 Espèces non modèles

Selon les espèces, les élément transposables peuvent poser plus ou moins problème pour l'assemblage du transcriptome. Par exemple, chez *Drosophila melanogaster*, il y a peu d'ET insérés dans des gènes. Les ET sont soit actifs et produisent des transcrits, soient inactifs et ne sont pas transcrits. Ainsi, au moment d'assembler les lectures, les copies provenant d'une même famille d'ET seront généralement assemblées ensemble pour former une séquence consensus (ce qui est aussi valable pour les familles de gènes). On pourra, si on identifie correctement le transcrit assemblé, étudier ensuite la famille d'ET assemblée

(quantification, variants nucléotidiques)

Chez l'Homme en revanche, on trouve plus de 2 millions de copies d'ET (principalement de type Alu) insérées dans des gènes, le plus souvent dans les introns, et plus rarement exonisées (1824 cas d'après Sela et al. [2007]). De plus, on retrouve généralement autour de 5% de d'ARN pré-messagers dans une extraction d'ARN total avec sélection des ARN polyadénylés (protocole polA+) Tilgner et al. [2012]. Du fait de leur caractère répétés (répétition inexactes) ces ET créent dans le graphe de de Bruijn, des régions complexes et peuvent aboutir à des erreurs d'assemblage (Figure 7). Dans ce cas les ET sont également un obstacle à l'analyse des gènes.

4 Conclusion

Si de nombreuses méthodes bio-informatiques existent aujourd'hui pour permettre de tirer le maximum d'information des données de séquençage, notamment le RNA-seq, il n'existe pas de "pipeline" optimald pour l'ensemble des applications et scénarios d'analyse de données RNA-seq.

Au cours de cette thèse je me suis particulièrement intéressée à l'analyse de données RNA-seq, principalement chez des espèces non-modèles, et donc avec majoritairement des approches d'assemblage. J'ai ainsi été fortement impliquée dans deux projets.

Un projet d'analyse de données RNA-seq, avec un intérêt particulier pour l'identification et la quantification les éléments transposables (Chapitre 2 et les deux articles en annexes).

Un second projet, visant à l'identification des variants nucléotidiques directement le graphe de de Bruijn construit à partir des lectures séquencées (Chapitre 3). L'étude menée a permis de clarifier les points forts et les limites de cette approche sur des données réelles, en la comparant à des méthodes basées sur l'alignement des lectures sur un génome de référence ou sur un transcriptome assemblé.

2

Expression des éléments transposables (et des gènes) chez les hybrides de D. mojavensis et D. arizonae

Sommaire

1 Transcriptomique 13

1.1 La transcription 13

1.2 Technologies pour l'analyse à large échelle du transcriptome 15

1.3 Design expérimental 20

2 L'analyse de données RNA-seq 22

2.1 Méthodes d'alignements des lectures 22

2.2 Méthodes d'assemblage (sans génome de référence) 24

2.3 Reconstruction des transcrits et épissages alternatifs 27

2.4 Identification des variants nucléotidiques 28

2.5 Accès à la quantification 29

2.6 Comparer deux (ou plus) échantillons 30

3 Les répétitions et éléments transposables 31

3.1 Analyse bio-informatique des éléments transposables en génomique 34

3.2 Analyse bio-informatique des éléments transposables en RNAseq . 35

4 Conclusion 37

1 Avant-propos

Ce projet est issu d'une collaboration entre l'équipe *Élément Transposables, Évolution, Population* (LBBE) et l'équipe de Claudia Carareto (UNESP, Brésil).

L'hybridation entre différentes espèces, lorsqu'elle est possible, peu constituer un stress génomique et aboutir des changements du génome hybride avec des conséquences sur la viabilité de ces hybrides. Entre autres, des *bursts* de transposition ont pu être observés chez les hybrides interspécifiques de différents organismes : chez des plantes, chez des wallabys, ainsi que chez des drosophiles (Baack et al. [2005]; Labrador et al. [1999]; Metcalfe et al. [2007]). La plupart de ces études restent néanmoins élément spécifiques et ne s'intéressent pas à l'ensemble des ET des espèces étudiées. Quelques études récentes chez la drosophile ont été faites sur l'ensemble du génome et montrent une réactivation des éléments transposables chez des hybrides (Kelleher et al. [2012]; Vela et al. [2014]).

L'objectif de ce travail été de regarder l'impact de l'hybridation sur la stabilité des génomes en utilisant un modèle avec un faible temps de divergence. Le but était de se mettre dans des conditions dans lesquelles l'hybridation est relativement facile, mais les hybrides ont un niveau de fertilité réduit. Nous nous intéressons ici à l'activité des ET de manière globale, à l'échelle du transcriptome, chez les hybrides de deux drosophiles phylogénétiquement proches. *Drosophila mojavensis* et *Drosophila arizonae* sont deux espèces endémiques du sud-ouest des États-Unis et du Mexique ayant divergé très récemment (moins d'un million d'années). Nous avons établi les croisements réciproques entre ces deux espèces, puis séquencé (technologie Illumina) les transcriptomes du tissu germinal femelle pour 30 individus de chaque lignée parentale et chacune des deux lignées hybrides. Nous avons également séquencé les piRNA issus des lignées hybrides.

Afin d'identifier et quantifier les ET et les gènes, nous avons ici choisi de produire un transcriptome de référence en co-assemblant l'ensemble des lectures issues du séquençage des lignées parentales et hybrides. Ceci nous a permis de profiter d'une profondeur de séquençage artificiellement importante pour l'assemblage des ET et des gènes faiblement exprimés dans au moins une des quatre lignées.

Nos résultats montrent une différence d'expression des ET chez les lignées parentales, suggérant ainsi une différence du nombre de copies actives de ces éléments et/ou une différence de régulation des ces éléments. Chez les hybrides l'expression des éléments transposables reste proche de celle des lignées parentales et seuls deux ET montrent une

activation importante.

L'élément Copia1 est largement sur-exprimé chez les hybrides issus d'une mère *D. mojavensis*. Un élément de la famille des gypsys est lui très fortement exprimé chez les hybrides issus d'une mère *D. arizonae*.

Nos résultats ne montrent pas d'activation globale des ET chez les hybrides de *D. mojavensis* et *D. arizonae*, mais une forte dérégulation de quelques éléments en particulier. L'analyse des données de séquençage des piRNA chez lignées hybrides semblent montrer que la dérégulation des deux ET est liée à une diminution des piRNA secondaires pour ces éléments.

2 Article 1 : Identification of misexpressed genetic elements in hybrids between Drosophila-related species

Cet article a été accepté pour publication dans *Scientic Reports* le 9/12/2016.

Title: **Identification of misexpressed genetic elements in hybrids between Drosophila-related species**

Hélène Lopez-Maestre[1,2], Elias A. G. Carnelossi[3], Vincent Lacroix[1,2], Bruno Mugat[4], Séverine Chambeyron[4], Claudia M. A. Carareto[3], Cristina Vieira[1]*

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

[2]ERABLE-team, INRIA Grenoble Rhône-Alpes

[3]Department of Biology, UNESP - São Paulo State University, São José do Rio Preto, São Paulo, Brazil

[4]Institut de Génétique Humaine, Centre National de la Recherche Scientifique, UPR1142, Montpellier, France

Corresponding author*:

E-mail: cristina.vieira@univ-lyon1.fr

Abstract

Crosses between close species can lead to genomic disorders, often considered to be the cause of hybrid incompatibility, one of the initial steps in the speciation process. The way these incompatibilities are established and their causes are still unclear. To understand the initiation of hybrid incompatibility, we performed reciprocal crosses between two species of Drosophila (*D. mojavensis* and *D. arizonae*) that diverged less than 1 Myr. We performed a genome wide transcriptomic analysis on female germline tissues from parental lines and hybrids from reciprocal crosses. Using an innovative procedure of co-assembling transcriptomes, we show that parental lines differ in their gene and transposable element expression. Reciprocal hybrids presented specific gene categories and several transposable element families misexpressed relative to the parental lines. Because TEs are mainly silenced by piwi-interacting RNA (piRNA), we hypothesize that in hybrids the deregulation of specific TE families is due to the absence of such small RNAs. Small RNA sequencing confirm our hypothesis and therefore we propose that TEs can indeed be major players on genome differentiation and be implicated in the first steps of genomic incompatibilities through small RNA regulation.

Introduction

Interspecific hybridization can be considered as a stress condition with multiple consequences for the hybrid genome. It may cause chromosomal rearrangements, inversions, deletions, changes in gene expression, changes in DNA methylation, among other effects [1,2]. Global activation of transposable elements (TEs), which induces profound changes in the hybrid genome, has also been described. Such changes generate new phenotypes and the formation of reproductively isolated populations because the accumulation of structural and functional genomic changes acts as a pressure leading to speciation [3–5]. For example, hybrid *Helianthus*, derived from crosses of the same parental species with other hybrids, have 50% more nuclear DNA than the parental, mainly due to bursts of transposition [6]. Interspecific hybrids of kangaroos from the Macropodidae family also showed variation in amplification of satellite repeats and kerV-1 element, changes in chromatin structure and rearrangements of whole chromosome arms [7], which demonstrates that during hybridization, increased transposition is observed, inducing significant changes in karyotype [3,8].

In *Drosophila*, studies of intraspecific crosses revealed asymmetric sterility of the offspring. This phenomenon was named hybrid dysgenesis and was first described in the 1960s in *D. melanogaster* with the I/R system [9] and then the P/M system [10]. Hybrid dysgenesis corresponds to aberrant phenotypic traits observed in the F1 of crosses between particular strains or natural populations and was proposed as an important driver of speciation. Hybrid dysgenesis was attributed to differences in TE contents between parental lines. We now know that TEs are major components of the genome architecture because they may

encompass a large fraction of the genome size and may trigger recombination. However, we also know that most of the TEs in the genomes are inactive. The last decade shed light on TE epigenetic control. In *Drosophila*, most TEs are post-transcriptionally silenced via a particular class of small RNAs, called piRNAs (piwi-interacting RNAs) [11–13]. Subsequently, transcriptional silencing is also caused by chemical histone modifications, which change the chromatin structure [14,15]. When the efficiency of the effectors of these pathways is no longer maintained, TEs burst into genomes, which leads to significant fitness decrease up to lethality [16–18]. Due to the recent development of our knowledge in epigenetics, we know that hybrid dysgenesis is caused by differences in the piRNA contents between the parental lines. When two strains display different TE contents, and therefore different associated piRNA contents, a cross between a male with an active TE family and a female devoid of the corresponding piRNAs leads to a major increase in TE expression, disrupting the genome stability, which could result in sterility or lethality [19,20]. Hybrid dysgenesis also occurs in *D. virilis* and is due to the death of germ cells during embryogenesis related to the initiation of transcription of the retrotransposon *Penelope* [21]. In artificially interspecific hybrids between *D. melanogaster* and *D. simulans*, TEs are derepressed due to adaptive divergence in the piRNA genes of both species rather than differences in TE contents [22]. Other studies with crosses between *D. buzatti* and *D. kopferae* have shown that 70% of the genomic rearrangements observed in hybrids was due to TE insertions [23].

To understand the first steps in hybrid incompatibility, we propose the use of related species that diverged recently (less than 1 Mya). *D. arizonae* and *D. mojavensis* are endemic species of the arid southwestern United States and Mexico (Figure 1A). *D. arizonae* occurs in the cape region in Baja California, southeastern Arizona, southeastern New Mexico, the

southeastern Sonoran Desert, eastern Mexico and Guatemala. *D. mojavensis* occurs in the Mojave and Sonoran Deserts, southern California and Baja California (USA) and along the west coast of Sonora and Sinaloa (Mexico), where it is sympatric with *D. arizonae* [24–26]. The two species diverged recently (between 0.6 and 1 Mya) [27–29] and the degree of pre-zygotic isolation between them is strong, but it is incomplete and variable, depending on the geographic origin of the populations. The pre-zygotic isolation is higher between the sympatric than allopatric populations [24,30,31]. Hybridization between the two species does not occur in nature or is extremely rare [24,26], but in the laboratory, crosses between *D. mojavensis* and *D. arizonae* are possible and present variation in the degree of sterility of the males [32,33]. Most of the studies performed up to now in this system consider the pre-zygotic mechanisms of isolation [34], and to our knowledge, no data are available after the breakdown of the pre- to post-zygotic isolations. We chose to cross two allopatric strains for which we can obtain hybrids in the laboratory and analyzed the transcriptomes from the female ovaries of both parental and reciprocal hybrids (Figure 1B).

We showed that reciprocal hybrids presented average levels of gene expression compared to the parental lines, with some specific gene categories being misexpressed such as genes related to embryo development. As for TEs, we identified several families that were highly expressed in hybrid crosses, relative to the parental lines. Because TEs are mainly silenced by small RNAs from the piwi small RNA class (piRNA), we hypothesize that in hybrids the deregulation of specific TE families is due to the absence of such small RNAs. Indeed, small RNA sequencing confirm our hypothesis and therefore we propose that TEs can indeed be major players on genome differentiation and be implicated in the first steps of genomic incompatibilities through small RNA regulation.

Results

Co-assembling - Quantification - Genes and TE identification

We sequenced the ovarian transcriptomes of two parental allopatric strains (*D. mojavensis* and *D. arizonae*) (Figure 1A) and of reciprocal hybrid crosses (named hereafter as crosses Hybrid A and B, see Figure 1B). We obtained a total of 456 million paired-end reads, corresponding to 55 to 60 million reads for each of the parental and hybrid libraries (2 replicates for each condition). The reads were trimmed according to their quality [35]. To produce a reference transcriptome, we co-assembled all reads using the Trinity assembler [36]. Our choice to co-assemble all reads was motivated by the following reasons: 1) no reference genome was available for *D. arizonae*; hence, mapping all reads to the *D. mojavensis* genome would have biased the results towards *D. mojavensis* genes and 2) assembling each dataset separately results in a poor resolution for genes that are moderately or lowly expressed. To control the efficiency of the co-assembly, we verified that the number of contigs obtained was higher (21000 vs 15000), as was their total length (24 Gb vs 19 Gb), when compared to the individual assemblies of each dataset. One risk of co-assembling is the increased possibility of generating chimeric contigs. We therefore checked for chimeric contigs and found that the number of contigs not mapping to the *D. mojavensis* genome was similar when co-assembling compared to using single assemblies (815 vs 728 and 1227). These results are summarized in SM Table 1.

This reference transcriptome contains 36,459 transcripts grouped in 21,889 loci. We quantified each transcript using Bowtie and RSEM (see Materials and Methods) and assigned

a measure of expression to each one. The distribution of the expression levels is reported in SM Figure 1. There are two modes in this distribution, suggesting that half of the loci are highly expressed, whereas the other half are lowly expressed and could be interpreted as transcription noise, which has been previously reported with transcriptome data [37].

We further attempted to identify all loci by aligning them against the *D. mojavensis* genome (see Materials and Methods). From the initial 21,889 loci, 11,155 were unambiguously assigned to a single protein coding gene, 2,109 matched several protein coding genes, 7,610 corresponded to intergenic regions, 219 corresponded to TEs and 795 did not align to the reference genome. The assembler may produce several loci that correspond to the same gene; for instance, when a gene has a low expression level, some of the genes can be low-covered or not covered at all by the reads, and the assembler will fail in the reconstruction of the complete gene but may assemble some part of it. Therefore, the 11,155 loci that mapped to unique genes were then clustered into 5,450 genes, for which we have a gene annotation. The 219 loci that corresponded to transposable elements were clustered into 72 TE families. The analysis was then performed for 72 TEs and a total of 15,964 loci that corresponded to 5,450 predicted/annotated genes, 2,109 contigs matching several protein coding genes, 7,610 intergenic RNAs and 795 other loci.

Expression divergence of the parental transcriptomes

The identified loci for each species and hybrids were classified according to the GO terms. As seen in Figure 2, the distribution of the GO terms was homogeneous between species and hybrids, which indicates that the same genes were found in the four transcriptomes. Most of the transcribed genes belong to biological regulation, cellular component and cellular process

GO terms. From the 15,954 loci, 19% (3,202) were differentially expressed between *D. mojavensis* and *D. arizonae,* with a maximum fold-change of 2,131. Of the 3,202 differentially expressed loci between *D. mojavensis* and *D. arizonae*, 1,791 (56%) corresponded to protein coding regions. SM Table 2 shows the top 30 differentially expressed genes. Most of these genes have unknown functions based on their orthologs from *D. melanogaster* (21/30). As seen in Figure 3A and SM Table 3, the distribution of the fold changes is symmetric, which indicates that a similar number of loci are under- (55%) or over- (45%) expressed in each species.

From the 72 TE families identified in our data, 29 were differentially expressed between the two parental lines (40%) that belong to the different classes of TEs: eight DNA-transposons (Class II), 19 LTR retrotransposons (Class I) and one non-LTR retrotransposon (Class I) (Figure 3B, SM Table 4). As for genes, no asymmetry was detected in the distribution of the fold changes for TEs.

Transcriptome of the hybrids

Hybrids were obtained in a reciprocal manner, which allowed us to search for parental effects. We found that 840 loci (5.3% of all identified loci from the co-assembling procedure) were differentially expressed between the two hybrid lines (SM Table 5, Figure 3C) with a maximum fold-change of 595 (SM Table 6). Of these 840 loci, 597 (71%) were annotated as genes and 64% were included in those that were differentially expressed between the parental lines.

In contrast to the fold changes observed between the parental lines, Figure 3C and SM Table 6 show that there is an important asymmetry in the distribution of the fold changes

between the hybrids. Indeed, 721 loci are over-expressed in hybrid A, whereas 119 are over-expressed in hybrid B (respectively, 86% and 14%). This asymmetry is also true if we restrict the results to loci identified as protein coding genes: 529 (88%) are up-regulated in hybrid A, whereas only 68 (12%) are up-regulated in hybrid B.

Moreover, if we look at the number of genes differentially expressed between the hybrids and each parental line, hybrids are more similar to the females of the maternal line than to the females of the paternal line. Hybrid A has 1,207 genes that are differentially expressed with its maternal line, *D. mojavensis*, and 1,422 genes that are differentially expressed with its paternal line, *D. arizonae*. Hybrid B has 954 genes that are differentially expressed with his maternal line, *D. arizonae*, and 1,752 genes that are differentially expressed with its paternal line, *D. mojavensis*.

For the TE families, eight (12%) are differentially expressed between the two hybrids (Figure 3D, SM Table 7), from which seven were already detected as differentially expressed between the parental lines. *Copia1* and GTWIN, two LTR retrotransposons, showed the greatest difference (Figure 4 A and B), with a total of 473,178 reads in hybrid B (0.4% of the total reads) corresponding to GTWIN. These results were confirmed by RTqPCR experiments (SM Figure 2).

Expression Inheritance

We determined the mode of expression inheritance for the loci and the TEs by comparing the expression levels between one hybrid and each of the parental lines. The expression inheritance was analyzed according to [38] (Figure 5A).

Gene expression in hybrids is highly conserved

For all genes, the "conserved" category (in which hybrids have the same levels of expression as the parental lines and there is no difference between parental lines) is the most common for both hybrid lines, including 9,127 loci in hybrid A and 9,138 in hybrid B (>71%). The conserved genes in hybrid A and hybrid B are mostly the same (98%) (Figure 5 B), which indicates that the loci that are not differentially expressed between the parental lines have the same expression in the hybrid lines. Thirteen percent of the loci (1,793 loci in hybrid A and 1635 in hybrid B) follow the additive model, which means hybrid expression is intermediate between both parental lines. Twelve percent of the loci in hybrid A and 13% in hybrid B follow a dominant model, with hybrid A having more *D. mojavensis*-dominant loci and hybrid B more *D. arizonae*-dominant loci.

We found no massive misexpression of the loci in hybrids. Few loci were classified as over-dominant (148 in hybrid A, 70 in hybrid B) or under-dominant (23 in hybrid A, 105 in hybrid B), of which 74% were identified as protein coding genes (Figure 5B). Very few misexpressed loci were common between both hybrids. There was a total of 43 common over-dominant loci (Table 1), most of which were involved in metabolic processes and/or had catabolic activity, and a total of 7 under-dominant loci (Table2), all of which were involved in embryo development.

TEs are under control in hybrids

From the 43 TEs not differentially expressed between the parental lines, 37 were also not differentially expressed in the hybrids and belonged to the conserved category (Figure 5). Fourteen elements in hybrid A and nine in hybrid B followed the additive model; 14 elements

in hybrid A and 25 in hybrid B were either *D. mojavensis*-dominant or *D. arizonae*-dominant. Only one element (the *I* element) in hybrid A was in the under-dominant category. Four TEs in hybrid A (*Gypsy7*-Dmoj, Homo7, FROGGER and *Copia1*-Dmoj) and only one in hybrid B (GTWIN) belonged to the over-dominant category. For two of them, Copia1 in hybrid A and GTWIN in hybrid B, the over-expression was especially high (Figure 4 A and B), with fold-changes higher than 10 comparing to the parental line with the highest expression.

GTWIN and Copia1 element

We determined the copy number and structure of these two TE families in the *D. mojavensis* sequenced genome. GTWIN (which belongs to the *gypsy*-like family) is highly expressed in hybrid B and is present as eight copies in the *D. mojavensis* genome. The average identity between copies (pairwise) was 99%, which indicates that GTWIN insertions are recent in the sequenced genome and may correspond to still active copies. For this element, no SNPs were found along the sequence in the reads of hybrid B or hybrid A, which indicates that only one type of insertion is being transcribed.

The *Copia1* element, which was significantly more highly expressed in Hybrid A, is present as approximately 40 copies in the *D. mojavensis* genome, with an average identity up to 70%, which indicates that the elements were probably active at a more distant time and that the transcripts are from the most intact copies. For *Copia1* element, only two SNPs were identified along the sequence in hybrid A, which indicates that only one type of insertion is being transcribed.

piRNAs are a class of small, non-coding RNA (23 to 29 nucleotides) that play a role in the silencing of TEs. piRNAs can be produced in two different pathways: primary piRNAs come

from piRNAs clusters distributed throughout the genome and are produced in somatic and germline cells, whereas secondary piRNAs are derived from the product of cleavage of functional TE transcripts and are maternally transmitted to embryos. Secondary piRNA production, also called the "ping-pong" pathway, is characterized by piRNA sequences that present complementarity with exactly 10 nucleotides of the primary piRNA.

To better understand the expression increase of these TEs in hybrids, we sequenced piRNA from Hybrid A and B and searched for ping-pong signatures for GTWIN and *Copia1* (Figure 4 C and D) [39,40].

In hybrid B, the GTWIN element was 32 times more expressed than in hybrid A. This high level of mRNA is accompanied by a weak ping pong signature in the piRNA pool (Figure 4 A –D), which is compatible with the hypothesis that no secondary piRNA were maternally transmitted to silence the element in the germline. However, there was a significant amount of total piRNA in hybrid B (SM Figure 3), mainly primary piRNA, showing that these sequences do not contribute to the silencing of GTWIN.

For *Copia1,* we found a high ping-pong signature in hybrid B and a lower ping-pong signature in hybrid A, where the element is highly expressed. There is a positive relation between the amount of mRNA and the abundance of *Copia1* piRNA: hybrid A had 98-times higher expression than hybrid B, and the abundance of piRNA was 2.2-times higher in hybrid A (Figure 4 A to D, SM Figure 3)).

Discussion

Twenty percent of genes were differentially expressed between the two parental lines, *D. mojavensis* and *D. arizonae*, which diverged between 0.6 and 1 Mya [27,30,31,41]. This was consistent with data obtained by Matzkin and Markow (2013) [42], who found that up to 17% of genes were differentially expressed between *D. mojavensis* subspecies. Additionally, studies comparing more distant species, such as *D. melanogaster* and *D. sechellia*, which diverged approximately 1.2 Mya ([43]), showed up to 78% of genes with differences in expression [38]. In other studies comparing *D. melanogaster*, *D. simulans* and *D. yakuba* [44,45], at least 27% of genes were differentially expressed between species or strains. Genes that were differentially expressed between the parental lines were essentially genes related with development.

We performed reciprocal crosses to check for parental effects on hybrids between *D. mojavensis* and *D. arizonae*. In general, gene expression was fairly similar between hybrids, with fewer genes differentially expressed than between the parental lines. Moreover, for the 5% of genes that differed between the hybrids, most were up-regulated in hybrid A. This indicates that for some genes, there is an effect of the parental line. Despite the studies conducted on hybrid dysgenesis, we have no other *Drosophila* data with reciprocal crosses to compare with because most previous studies were perform in one cross direction [46].

In hybrids between *D. melanogaster*/*D. sechellia* and *D. melanogaster*/*D. simulans*, most of the genes were either *sechellia/simulans*-dominant or under-expressed [38,44]. In our study, the comparison between the hybrids and the parental lines showed that most of the genes had expression that was conserved or additive due to the low divergence between the parental species. A few genes had an expression level closer to the maternal line, which was

either *mojavensis*-dominant (hybrid A) or *arizonae*-dominant (hybrid B). Few genes were up- or down-regulated. The detailed analysis of these unregulated categories shows that the genes that are in common in both hybrids are related to metabolic and embryo development. In a previous study, different life history traits and viability were measured in hybrids of *D. mojavensis* and *D. arizonae* and were compared to their parents [31]. Female hybrids (from both crosses) had performances equal to their mothers. This is consistent with our observation because the vast majority of genes had a conserved pattern between the hybrids and parents. Moreover, genes that are up-regulated in hybrids are implicated in the good performance of the hybrids. In contrast, down-regulated genes are related to embryonic development and could preclude sterility problems in the hybrids. We followed the allele specific expression to investigate differences in the regulatory systems. For the vast majority, there was no significant evidence of regulatory divergence, contrary to what had been described for *D. melanogaster*/*D. sechelia* hybrids, but which is in agreement with the expression inheritance data from this study.

The comparison of expression between *D. mojavensis* and *D. arizonae* showed that of the 72 TEs that were identified in the transcriptome, 40% were differentially expressed. This emphasizes the fact that closely related species may have very different amounts and expression levels of TEs [47–50] and that these differences may also exist between strains [5,50]. Again, when comparing both hybrids, very few elements were differentially expressed, indicating that species-specific regulatory systems are operating in the hybrids. This has not been observed in hybrids between more distantly related species. In crosses between *D. melanogaster* and *D. simulans*, which were performed with specific mutant strains of *D.*

simulans that "allow" the development of the F1 hybrids, a massive increase of transposition was observed for most of the elements. The authors claimed that time allowed divergence in the regulation system, namely, the implication of the proteins of the piRNA biogenesis that have diverged [22]. In another system, with hybrids between *D. buzatti* and *D. koepferae*, the authors showed, in a genome-wide manner, massive rearrangement in the F1 hybrids [23]. A wide variety of TEs were responsible for most of the genomic instability in the hybrids.

In our analysis, we identified eight TEs (SM Table 7) belonging to different classes of TEs that were differentially expressed between hybrids, but only two were highly up-regulated compared to the parents. GTWIN is highly expressed in hybrid B, and *Copia1* is highly expressed in hybrid A. The specific analysis of RNA sequences from these elements allows us to propose a scenario that is consistent with the idea of clusters producing piRNA that are not equally present in the parental lines. GTWIN insertion could be present in the paternal line of hybrid B, *D. mojavensis*, but not in the maternal line because the expression of GWTIN is low in *D. arizonae*; therefore, the secondary piRNA corresponding to the element could not be transmitted by the maternal line and did not lead to a ping-pong amplification cycle in hybrid B. The same scenario can be proposed for *Copia1*. The *Copia1* insertion could be present in the paternal line of hybrid A, *D. arizonae*, but not in the maternal line because the expression of *Copia1* is low in *D. mojavensis*. Therefore, the secondary piRNA corresponding to the element could not be transmitted by the maternal line and did not lead to a ping-pong amplification cycle in hybrid A. This scenario corresponds to what is observed when crossing different strains of *D. melanogaster*, *D. simulans* and *D. virilis* harboring different TE amounts and activities, which results in the derepression of TE [10,15,19,51].

Crosses between closely related species often result in male sterility, which is one of the

expected steps of speciation and is known as the Haldane's rule [52]. In crosses between *D. melanogaster* that induce hybrid dysgenesis, strong advances have been made that show that the absence of maternally transmitted piRNAs from specific TEs is responsible for the female phenotype that can be visible in the first generation by gonadic atrophy or by female sterility. What is happening in the male germline is much less understood.

Reproductive isolation between *D. mojavensis* and *D. arizonae* has been studied extensively, with both pre-zygotic and post-zygotic barriers contributing to isolation [31,53,54]. Sexual isolation between these species varies according to the strains used [24,53], and with respect to post-zygotic isolation, there is an asymmetry in the production of sterile hybrid males. When *D. arizonae* mothers are used, the hybrid sons are sterile, but in the reciprocal cross, hybrid males are only sterile when the *D. mojavensis* populations are from certain geographic host races [32].

We have already shown, in accordance with this variation and asymmetry of the post-zygotic isolation, that in *D. mojavensis*/*D. arizonae* hybrids, some TEs were specifically derepressed in the male germline, such as the non-LTR retrotransposon *I* and *Helena* elements, depending on the source population of males and females and on the direction of the crosses [55], unpublished). Because maternally transmitted piRNAs are an important way of controlling TEs across generations, we can speculate that such small RNAs do not contribute to the male germline regulation, which could explain why it is usually the male that is sterile. The sterility could be associated with the mobilization of TEs. Our results also suggest that the female germline is successfully protected (even if some specific elements escape this control) against transposition by the maternally transmitted secondary piRNAs.

Although sterility of the heterogametic sex is one of the most common and presumably

earliest manifestations of postzygotic reproductive isolation it appears to be a complex trait, and consequently the genetic basis for its appearance is not yet completely understood. Our findings on TE expression variation in female germ line, depending on the parental lines and reciprocal crosses, point out for the necessity of further population studies in order to investigate a role of these mobile elements in the post-zygotic reproductive isolation of these pair of species. The study of the male germ lines is also fundamental because it could explain why TEs, despite a strong negative selection against deleterious effects of transposition, are successful to stay active in the male line, and transmitted across generation. Population studies on TEs in such a system can give insights into how reproductive isolation evolves.

We show that *D. mojavensis* and *D. arizonae* parental lines differ in their gene expression (~20% genes differentially expressed) and in their TE expression (~40% TE differentially expressed). Reciprocal hybrids presented average levels of gene expression compared to the parental lines, with some specific gene categories being misexpressed such as genes related to embryo development. As for TEs, we identified several families that were strongly expressed in hybrid crosses, relative to the parental lines. Moreover, piRNA sequencing confirms that in hybrids the deregulation of specific TE families is due to the absence of such small RNAs. We therefore propose that TEs can indeed be major players on genome differentiation and be implicated in the first steps of genomic incompatibilities through small RNA regulation.

Methods

Drosophila strains and RNA sequencing

We sequenced RNA-poly (A) from the ovaries of flies. The sequenced strains were *D. mojavensis*, from the Anza Borrego Desert, CA (stock number: 15081-1352.01) and *D. arizonae*, from Metztitlan-Hidalgo, Mexico (stock number: 15081-1271.17), both obtained from the US San Diego Drosophila Stock Center. These are two allopatric species with which we can perform reciprocal crosses in laboratory conditions to provide sufficient F1 hybrid individuals to obtain enough RNA for sequencing. Parental individuals were separated to collect virgins one day after hatching. Crosses were performed with 3-day-old flies; ten males and eight females were placed in 2.3 x 9.5 cm tubes containing culture medium under the same temperature and humidity conditions. Virgin female parental flies and F1 female hybrids were collected after hatching, at one day of age and were isolated until they reached ten days. The RNA was extracted from the ovaries of 10-day-old flies (i.e., *D. mojavensis*, *D. arizonae* and hybrids from reciprocal crosses). The extractions were performed using the RNeasy kit (Qiagen), and the samples were then treated with DNase (DNA-free Kit, Ambion) and stored at -80°C. The samples were quantified by fluorescence in a Bioanalyzer 2100 (Agilent).. For each line, the extracted RNA was divided into two parts to generate two cDNA libraries (two replicates per condition). RNA was sequenced by Illumina Technology in an Illumina HiSeq 2000.We sequenced 2x51 bp reads and the medium size of the inserts was 300 bp. We used UrQt [35] with the default parameters to remove the low quality bases and the polyA tail from the data set.

Assembly of the transcriptome

The reads were co-assembled, i.e., we use the reads from all (parental and hybrid) lines that passed purity filtering to construct a de novo reference transcriptome. We ran Trinity [36]

version r2013_08_14 with the default parameters and a group_pairs_distance of 600. Thus, these transcripts are consensus transcripts.

This approach is possible because the two parental lines diverged recently, so we assumed that the transcripts of the species and the hybrids are similar enough to be assembled together. This method has the effect of increasing the sequencing depth and allows us to better assemble transcripts that are too low-expressed in one or more species and that could not be assembled otherwise, which can be the case for TEs, which can be low-expressed in parental lines. Additionally, unlike the mapping method, this approach has no bias in favor of *D. mojavensis*.

Quantification of expression

The quantification of the contigs expression of each replicate of each line was performed with Bowtie [56] and RSEM [57]. Bowtie (with default parameters) was used to map the reads to the contigs of the reference transcriptome we assembled. The number of reads aligning against each sequence was then counted by RSEM, which provided access to the expression of the transcripts and the genes (in FPKM). RSEM also addresses multiple mapping and assigns the read to its most likely location.

Gene and TE identification

To identify genes among the contigs assembled by Trinity, we downloaded the 15,179 sequences of annotated and predicted genes from *D. mojavensis* (version r1.3 from http://flybase.org/) and aligned our contigs with BLAT [58] with at least 80% identity and with a minimum query coverage of 80%. We also aligned all of the contigs with BLAT to the

reference genome of *D. mojavensis* (version r1.3 from http://flybase.org/) with at least 80% identity and with a minimum query coverage of 80% to search for transcripts originating from the intergenic region.

To the genes predicted in *D. mojavensis*, we assigned the GOterm of the orthologous genes in *D. melanogaster* using the orthologous tables downloaded from http://flybase.org/. We also ran Blast2GO[59] on the assembled transcripts and obtained the GO term for the transcripts. We kept all of the GO terms provided by at least one of the methods.

For TE identification, we used BLAT to align our sequences against consensus TEs from RepbaseDrosophila[60] (2,296TEs) and against a homemade database (4575 TEs). The homemade database was generated by running Repeatmasker[61] (http://www.repeatmasker.org/) on the *D. mojavensis* reference genome. We kept the alignments with an identity percentage higher than 70%, and with a minimum query coverage of 80%. Fourteen of the 72 TEs are lowly expressed in all species and hybrids (<10 reads), as are another 3,322 loci of the total 15,964. These loci were included in the analyses but were not tested for differential expression and therefore were not considered in the analyses of expression inheritance. Eight other loci were identified as mitochondrial genes (4-5 million reads per replicate) and were not included in our analyses.

Differential Expression with DESeq

We used DESeq[62], an R package, to identify loci and TEs that were differentially expressed between two lines. DESeq estimates the means and variances of raw read counts and tests for differential expression based on a model using the negative binomial distribution. Loci and

TEs are classified as significantly differentially expressed if 1) the p-value, after correction for multiple tests with the False Discovery Rate (FDR), is below 0.001 and 2) if the fold-change (expression ratio between the compared conditions) is above 1.5. Loci and TEs were considered to be too lowly expressed in all conditions when the counts for each line did not exceed 10. These loci and TEs were excluded from the inheritance expression analyses.

RT-qPCR proof of expression

The levels of expression of *Copia-1* and GTWIN were validated by RTq-PCR. Primers were designed from the consensus obtained after the transcriptome assembly and were specific to our strains. One microgram of sequenced RNA was treated with DNase (DNA-free Kit, Ambion) and was converted to cDNA using a Thermoscript Invitrogen kit. The cDNA was diluted 50 times, and the relative mRNA level was quantified using SYBR green qPCR in a LightCycler 480 instrument (Roche Diagnostics). The RT–qPCR experiments were performed with technical triplicates. Only RT–qPCR experiments with efficiencies greater than 1.9 were retained. The following primers were used: GTWIN forward 5'- CGC TGA CGG CAA TAA TGA AAG C – 3' and GTWIN reverse 5' – ATC TTC CGA TGC CAA GAT A -3'; Copia1 forward 5' - GTG GAC CTA TAA GGC AAG TAT C – 3' and Copia1 reverse 5'- AGA CCT TTC TGA CGC TCT A - 3'. The elements' relative expression levels were measured with the constitutive expression of the endogenous ribosomal gene 49 (rp49), also known as asnrpL32 [63].

Small RNA extraction and sequencing

Small RNAs from hybrid A and hybrid B ovaries were manually isolated on HiTrap Q HP anion exchange columns (GE Healthcare) as described in [64]. Library construction and 50 nt read sequencing were performed by Fasteris SA (Switzerland) on an Illumina HiSeq 2500 instrument.

Analyses of piRNA, ping-pong signatures and identification of ping-pong partners

We considered as piRNA the sequences of small RNAs of length 23 to 29nt that could be aligned against TEs from our assembled transcriptome or against TEs found in the genome of *D.mojavensis* (see TE annotation above). The alignments were performed with Bowtie using the --very-sensitive option. We then used the "Mississippi Tools" [65], which search for ping-pong signatures by counting the number of pairs of piRNA overlapping for 1 to 26 nucleotides.

Availability of supporting data

The RNAseq libraries generated in this study are available through the NCBI Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) under accession no. SRX1272419, SRX1277353, SRX1277354, SRX1277355, SRX1284317 and SRX1284318.

References

1. Fontdevila, A. Hybrid genome evolution by transposition. *Cytogenet. Genome Res.* **110,**

49–55 (2005).

2. Arkhipova, I. R. & Rodriguez, F. Genetic and epigenetic changes involving (retro)transposons in animal hybrids and polyploids. *Cytogenet. Genome Res.* **140,** 295–311 (2013).
3. Hedges, D. J. & Deininger, P. L. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutat. Res.* **616,** 46–59 (2007).
4. Oliver, K. R. & Greene, W. K. Transposable elements: powerful facilitators of evolution. *BioEssays News Rev. Mol. Cell. Dev. Biol.* **31,** 703–714 (2009).
5. Rebollo, R., Horard, B., Hubert, B. & Vieira, C. Jumping genes and epigenetics: Towards new species. *Gene* **454,** 1–7 (2010).
6. Baack, E. J., Whitney, K. D. & Rieseberg, L. H. Hybridization and genome size evolution: timing and magnitude of nuclear DNA content increases in Helianthus homoploid hybrid species. *New Phytol.* **167,** 623–630 (2005).
7. Metcalfe, C. J. *et al.* Genomic instability within centromeres of interspecific marsupial hybrids. *Genetics* **177,** 2507–2517 (2007).
8. Weil, C. F. Too many ends: aberrant transposition. *Genes Dev.* **23,** 1032–1036 (2009).
9. Picard, G. Non-mendelian female sterility in Drosophila melanogaster: hereditary transmission of I factor. *Genetics* **83,** 107–123 (1976).
10. Kidwell, M. G., Kidwell, J. F. & Sved, J. A. Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: A Syndrome of Aberrant Traits Including Mutation, Sterility and Male Recombination. *Genetics* **86,** 813–833 (1977).
11. Siomi, M. C., Sato, K., Pezic, D. & Aravin, A. a. PIWI-interacting small RNAs: the vanguard of genome defence. *Nat. Rev. Mol. Cell Biol.* **12,** 246–58 (2011).

12. Senti, K.-A. & Brennecke, J. The piRNA pathway: a fly's perspective on the guardian of the genome. *Trends Genet. TIG* **26,** 499–509 (2010).
13. Saito, K. & Siomi, M. C. Small RNA-mediated quiescence of transposable elements in animals. *Dev. Cell* **19,** 687–697 (2010).
14. Sienski, G., Dönertas, D. & Brennecke, J. Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell* **151,** 964–980 (2012).
15. Akkouche, A. *et al.* Maternally deposited germline piRNAs silence the tirant retrotransposon in somatic cells. *EMBO Rep.* **14,** 458–464 (2013).
16. Malone, C. D. *et al.* Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. *Cell* **137,** 522–535 (2009).
17. Li, C. *et al.* Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. *Cell* **137,** 509–521 (2009).
18. Vagin, V. V. *et al.* The RNA interference proteins and vasa locus are involved in the silencing of retrotransposons in the female germline of Drosophila melanogaster. *RNA Biol.* **1,** 54–58 (2004).
19. Chambeyron, S. *et al.* piRNA-mediated nuclear accumulation of retrotransposon transcripts in the Drosophila female germline. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 14964–14969 (2008).
20. Kelleher, E. S. & Barbash, D. a. Analysis of piRNA-mediated silencing of active TEs in Drosophila melanogaster suggests limits on the evolution of host genome defense. *Mol. Biol. Evol.* **30,** 1816–29 (2013).
21. Sokolova, M. I., Zelentsova, E. S., Shostak, N. G., Rozhkov, N. V. & Evgen'ev, M. B.

Ontogenetic consequences of dysgenic crosses in Drosophila virilis. *Int. J. Dev. Biol.* **57,** 731–739 (2013).

22. Kelleher, E. S., Edelman, N. B. & Barbash, D. a. Drosophila interspecific hybrids phenocopy piRNA-pathway mutants. *PLoS Biol.* **10,** e1001428 (2012).
23. Vela, D., Fontdevila, A., Vieira, C. & García Guerreiro, M. P. A genome-wide survey of genetic instability by transposition in Drosophila hybrids. *PloS One* **9,** e88992 (2014).
24. Wasserman, M. & Koepfer, H. R. Character Displacement for Sexual Isolation Between Drosophila mojavensis and Drosophila arizonensis. *Evolution* **31,** 812 (1977).
25. Koepfer, H. R. Selection for sexual isolation between geographic forms of Drosophila mojavensis. *Evolution* **41,** 37–48 (1987).
26. Ruiz, A., Heed, W. B. & Wasserman, M. Evolution of the mojavensis cluster of cactophilic Drosophila with descriptions of two new species. *J. Hered.* **81,** 30–42 (1990).
27. Reed, L. K., Nyboer, M. & Markow, T. A. Evolutionary relationships of Drosophila mojavensis geographic host races and their sister species Drosophila arizonae. *Mol. Ecol.* **16,** 1007–1022 (2007).
28. Matzkin, L. M. & Eanes, W. F. Sequence variation of alcohol dehydrogenase (Adh) paralogs in cactophilic Drosophila. *Genetics* **163,** 181–94 (2003).
29. Matzkin, L. M. Population genetics and geographic variation of alcohol dehydrogenase (Adh) paralogs and glucose-6-phosphate dehydrogenase (G6pd) in Drosophila mojavensis. *Mol. Biol. Evol.* **21,** 276–85 (2004).
30. Reed, L. K., LaFlamme, B. A. & Markow, T. A. Genetic architecture of hybrid male sterility in Drosophila: analysis of intraspecies variation for interspecies isolation. *PloS One* **3,** e3076 (2008).

31. Bono, J. M. & Markow, T. a. Post-zygotic isolation in cactophilic Drosophila: larval viability and adult life-history traits of D. mojavensis/D. arizonae hybrids. *J. Evol. Biol.* **22,** 1387–95 (2009).
32. Reed, L. K. & Markow, T. A. Early events in speciation: polymorphism for hybrid male sterility in Drosophila. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 9009–9012 (2004).
33. Machado, C. A., Matzkin, L. M., Reed, L. K. & Markow, T. A. Multilocus nuclear sequences reveal intra- and interspecific relationships among chromosomally polymorphic species of cactophilic Drosophila. *Mol. Ecol.* **16,** 3009–3024 (2007).
34. Bono, J. M., Matzkin, L. M., Kelleher, E. S. & Markow, T. A. Postmating transcriptional changes in reproductive tracts of con- and heterospecifically mated Drosophila mojavensis females. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 7878–7883 (2011).
35. Modolo, L. & Lerat, E. UrQt: an efficient software for the Unsupervised Quality trimming of NGS data. *BMC Bioinformatics* **16,** 137 (2015).
36. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–52 (2011).
37. Hebenstreit, D. *et al.* RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol. Syst. Biol.* **7,** 497–497 (2014).
38. McManus, C. J. *et al.* Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res.* **20,** 816–25 (2010).
39. Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128,** 1089–103 (2007).
40. Yin, H. & Lin, H. An epigenetic activation role of Piwi and a Piwi-associated piRNA in Drosophila melanogaster. *Nature* **450,** 304–8 (2007).

41. Matzkin, L. M. The molecular basis of host adaptation in cactophilic Drosophila: molecular evolution of a glutathione S-transferase gene (GstD1) in Drosophila mojavensis. *Genetics* **178,** 1073–83 (2008).

42. Matzkin, L. M. & Markow, T. A. Transcriptional Differentiation Across the Four Subspecies of Drosophila mojavensis. *Speciation: natural processes, genetics and biodiversity. New York: Nova Scientific Publishers* (2013). at <http://labs.biology.ucsd.edu/markow/documents/MatzkinandMarkow2013.pdf>

43. Cutter, A. D. Divergence times in Caenorhabditis and Drosophila inferred from direct estimates of the neutral mutation rate. *Mol. Biol. Evol.* **25,** 778–86 (2008).

44. Ranz, J. M., Namgyal, K., Gibson, G. & Hartl, D. L. Anomalies in the expression profile of interspecific hybrids of Drosophila melanogaster and Drosophila simulans. *Genome Res.* **14,** 373–9 (2004).

45. Rifkin, S. A., Kim, J. & White, K. P. Evolution of gene expression in the Drosophila melanogaster subgroup. *Nat. Genet.* **33,** 138–144 (2003).

46. Ranz, J. M., Yeh, S.-D., Nyberg, K. G. & Machado, C. A. Transcriptome profiling of Drosophila interspecific hybrids: insights into mechanisms of regulatory divergence and hybrid dysfunction. *Polyploid Hybrid Genomics* 15–35 (2013).

47. Vieira, C. *et al.* A comparative analysis of the amounts and dynamics of transposable elements in natural populations of Drosophila melanogaster and Drosophila simulans. *J. Environ. Radioact.* **113,** 83–6 (2012).

48. Biémont, C., Vieira, C., Borie, N. & Lepetit, D. Transposable elements and genome evolution: the case of Drosophila simulans. *Genetica* **107,** 113–120 (1999).

49. Fablet, M., McDonald, J. F., Biémont, C. & Vieira, C. Ongoing loss of the tirant

transposable element in natural populations of Drosophila simulans. *Gene* **375,** 54–62 (2006).

50. Rebollo, R. *et al.* A snapshot of histone modifications within transposable elements in Drosophila wild type strains. *PloS One* **7,** (2012).
51. Lozovskaya, E. R., Scheinker, V. S. & Evgen'ev, M. B. A hybrid dysgenesis syndrome in Drosophila virilis. *Genetics* **126,** 619–623 (1990).
52. Haldane, J. B. S. Sex ratio and unisexual sterility in hybrid animals. *J. Genet.* **12,** 101–109 (1922).
53. Markow, T. A., Fogleman, J. C. & Heed, W. B. Reproductive isolation in Sonoran desert Drosophila. *Evolution* 649–652 (1983).
54. Kelleher, E. S. & Markow, T. A. Reproductive tract interactions contribute to isolation in Drosophila. *Fly (Austin)* **1,** 33–37 (2007).
55. Carnelossi, E. A. G. *et al.* Specific activation of an I-like element in Drosophila interspecific hybrids. *Genome Biol. Evol.* **6,** 1806–1817 (2014).
56. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
57. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).
58. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12,** 656–64 (2002).
59. Conesa, A. & Götz, S. Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. *Int. J. Plant Genomics* **2008,** (2008).
60. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110,** 462–7 (2005).

61. Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-4.0*. (2013). at <http://www.repeatmasker.org>

62. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

63. Granzotto, A., Lopes, F. R., Lerat, E., Vieira, C. & Carareto, C. M. A. The evolutionary dynamics of the Helena retrotransposon revealed by sequenced Drosophila genomes. *BMC Evol. Biol.* **9,** 174 (2009).

64. Grentzinger, T. & Chambeyron, S. in *PIWI-Interacting RNAs* 171–182 (Springer, 2014).

65. Antoniewski, C. Computing siRNA and piRNA overlap signatures. *Methods Mol. Biol. Clifton NJ* **1173,** 135–46 (2014).

Acknowledgments

This work was supported by the ANR (grant Exhyb ANR-14-CE19-0016-01 to CV), the CNRS, the Institut Universitaire de France (grant to CV), the São Paulo Research Foundation-FAPESP/Brazil (grant 2010/10731-4 to C.M.A.C.) and the National Council for Scientific and Technological Development-CNPq/Brazil (CNPq fellowship 306493/2013-6 to C.M.A.C.). We would like to thank Profilxepert, DTAMB and the Centre de Calcul LBBE/PRABI for the technical facilities, M. Fablet, E. Lerat and R. Rebollo for useful discussion, and N. Burlet, S. Martinez, J. Kielbassa and G. Sacamoto for technical assistance.

Authors contributions

CV and CMAC designed the study. EC and BM performed the experiments. HLM, VL and SC analyzed the data. CV, CMAC, HLM and VL wrote the manuscript. All authors read and approved the final manuscript.

Additional Information

The authors declare no competing financial interests.

Figures Legends

Figure 1. A. **Geographic distribution of *D. mojavensis* and *D. arizonae.*** The two species occupy the south USA and Mexico with strains in sympatry and allopatry. The two strains used in this study come from allopatric regions (http://www.d-maps.com/). B. **Crosses between *D. mojavensis* and *D. arizonae.*** Reciprocal crosses were performed between the species with allopatric strains (see Materials and Methods). We named crosses made with *D. mojavensis* females hybrid A and crosses made with *D. arizonae* females hybrid B. C. **Co-assembly of the transcriptomes of the four conditions**. Co-assembly of the total number of reads allowed us to reconstruct a reference transcriptome that was non-biased to the sequenced genome of *D. mojavensis* and to identify low expressed elements.

Figure 2: **Distribution of the GOterm : Biological Process (level 2).** The genes predicted in

D. mojavensis, were assigned the GOterm of the orthologous genes in *D. melanogaster*.

Figure 3: **Distribution of the fold change** measured in loci (A) and the TE fold change between *D. mojavensis* and *D. arizonae* (B). Distribution of the fold change measured in loci (C) and the TE fold change between hybrid A and hybrid B (D).

Figure 4: **Description of *GTWIN* (left) and *Copia1* (right).** The expression (A) and coverage (B) of the TEs for each parental line and hybrid. C) The overlapping frequency of piRNA for both hybrids. A peak in the frequency for an overlapping size of 10 nucleotides is characteristic of a ping-pong amplification cycle. The height of the peak indicates the proportion of piRNA implicated in the ping-pong cycle. D) piRNA coverage of the TEs for both hybrid lines.

Figure 5: **Expression inheritance of genes and TEs**. A) Illustration of six patterns of expression inheritance. Loci are considered to be having a conserved expression when the expression is not different between the two parental lines and the expression in the hybrid is not different compared to each parental line. Loci and TEs are classified as additive when the expression is different between the two parental lines and the expression in the hybrid is intermediate. Loci and TEs for which the expression is similar to only one parental line, D. mojavensis or D. arizonae, are classified as D. mojavensis-dominant or D. arizonae-dominant. Loci and TEs are classified as over-dominant when the expression in the hybrid line is significantly higher than both parental lines and as under-dominant if the expression is significantly lower than both parental lines (adapted from MacManus et al. 2010). B) Expression inheritance of genes. D) Expression inheritance of TEs.

Tables

Table 1: **List of genes under-expressed in both hybrid A and hybrid B**

Gene ID (Flybase) or Component ID	Function
FBgn0138703	embryo development - neurogenesis - sex differentiation - vitellogenesis - lipid metabolic process
FBgn0141780	multicellular organism reproduction - neurogenesis
comp19727_c2	egg activation - chorion-containing eggshell formation - vitelline membrane formation - structural constituent of vitelline membrane
comp20848_c0	vitelline membrane formation involved in chorion-containing eggshell formation (conserved domain)
FBgn0135964	maternal specification of dorsal/ventral axis, oocyte - proteolysis
FBgn0140278	egg activation - chorion-containing eggshell formation - vitelline membrane formation - structural constituent of vitelline membrane
comp22809_c0	domain found : vitelline membrane formation

Table 2: **List of genes over-expressed in both hybrid A and hybrid B**

Gene ID (Flybase) or Component ID	Function
FBgn0137790	-
FBgn0138471	-
FBgn0135217	proteolysis
FBgn0135361	-
FBgn0139457	spermatogenesis
FBgn0136207	mannose metabolic process
FBgn0136788	synaptic vesicle exocytosis ; synaptic transmission, glutamatergic Calcium activated protein for secretion
FBgn0138627	proteolysis

FBgn0139424	proteolysis
FBgn0139425	proteolysis
FBgn0139428	proteolysis
FBgn0139429	proteolysis
FBgn0139449	proteolysis
FBgn0140182	serine-type endopeptidase activity
FBgn0141435	carbohydrate metabolic process - Maltase A4
FBgn0143612	lipid metabolic process
FBgn0143632	carbohydrate metabolic process
FBgn0143673	lateral inhibition ; Immunoglobulin-like domain
FBgn0140146	-
FBgn0146016	lipid metabolic process
FBgn0146332	proteolysis
FBgn0134493	chitin metabolic process
FBgn0147011	proteolysis
FBgn0147012	metallopeptidase activity; zinc ion binding
FBgn0147016	metallopeptidase activity; zinc ion binding
FBgn0142015	cold acclimation
FBgn0143635	carbohydrate metabolic process
comp20918_c0	-
comp22028_c0	-
FBgn0145976	lipid metabolic process
FBgn0146016	lipid metabolic process
FBgn0146018	lipid metabolic process
comp23342_c9	-
comp24075_c7	-
FBgn0135350	-
comp19798_c0	-
comp19850_c0	-
comp20845_c0	-
comp24075_c0	-
comp18308_c0	-

Figure 1

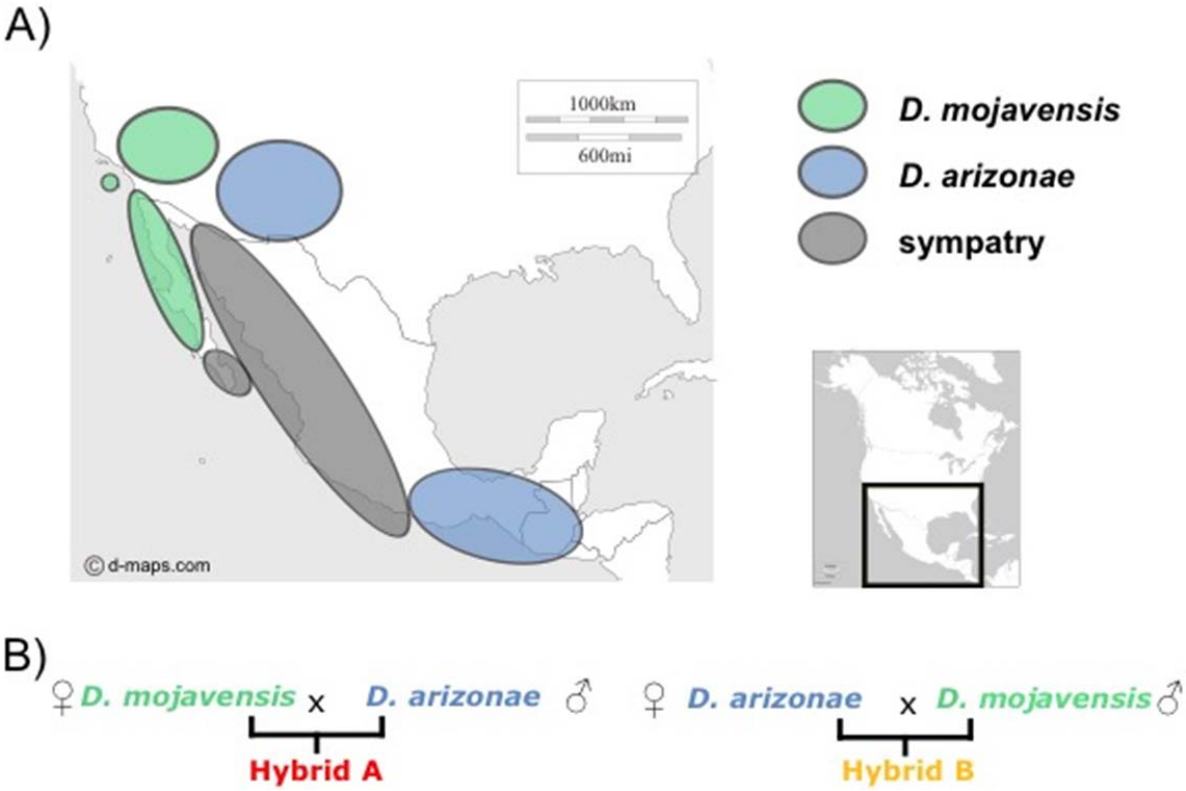

A)
1000km
600mi
© d-maps.com
D. mojavensis
D. arizonae
sympatry
B)
♀ D. mojavensis x D. arizonae ♂
Hybrid A
♀ D. arizonae x D. mojavensis ♂
Hybrid B

Figure 2

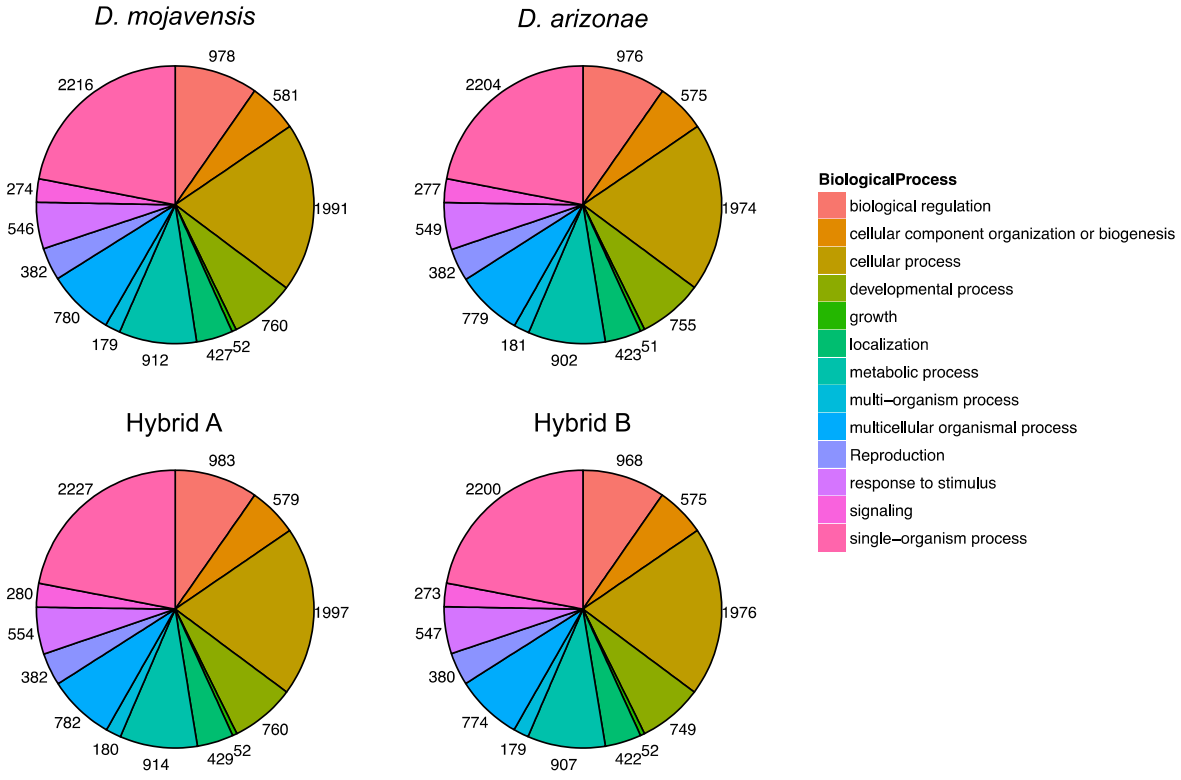

D. mojavensis
978
581
1991
760
52
427
912
179
780
382
546
274
2216
D. arizonae
976
575
1974
755
51
423
902
181
779
382
549
277
2204
Hybrid A
983
579
1997
760
52
429
914
180
782
382
554
280
2227
Hybrid B
968
575
1976
749
52
422
907
179
774
380
547
273
2200
BiologicalProcess
biological regulation
cellular component organization or biogenesis
cellular process
developmental process
growth
localization
metabolic process
multi–organism process
multicellular organismal process
Reproduction
response to stimulus
signaling
single–organism process

Figure 3

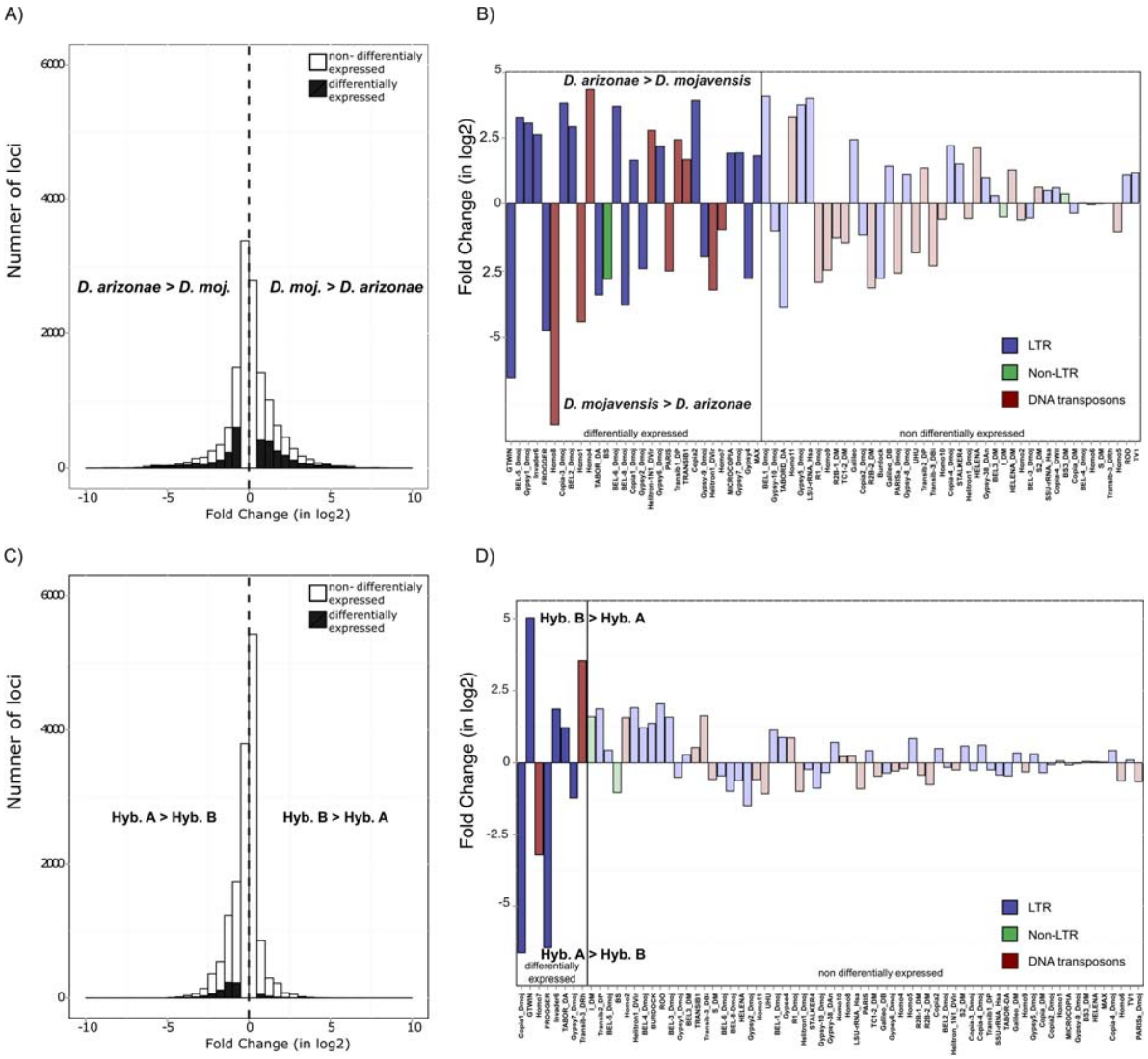

A)
B)
C)
D)
non- differentialy expressed
differentially expressed
Numner of loci
Fold Change (in log2)
D. arizonae > D. moj.
D. moj. > D. arizonae
D. arizonae > D. mojavensis
D. mojavensis > D. arizonae
Hyb. A > Hyb. B
Hyb. B > Hyb. A
LTR
Non-LTR
DNA transposons
differentially expressed
non differentially expressed

Figure 4

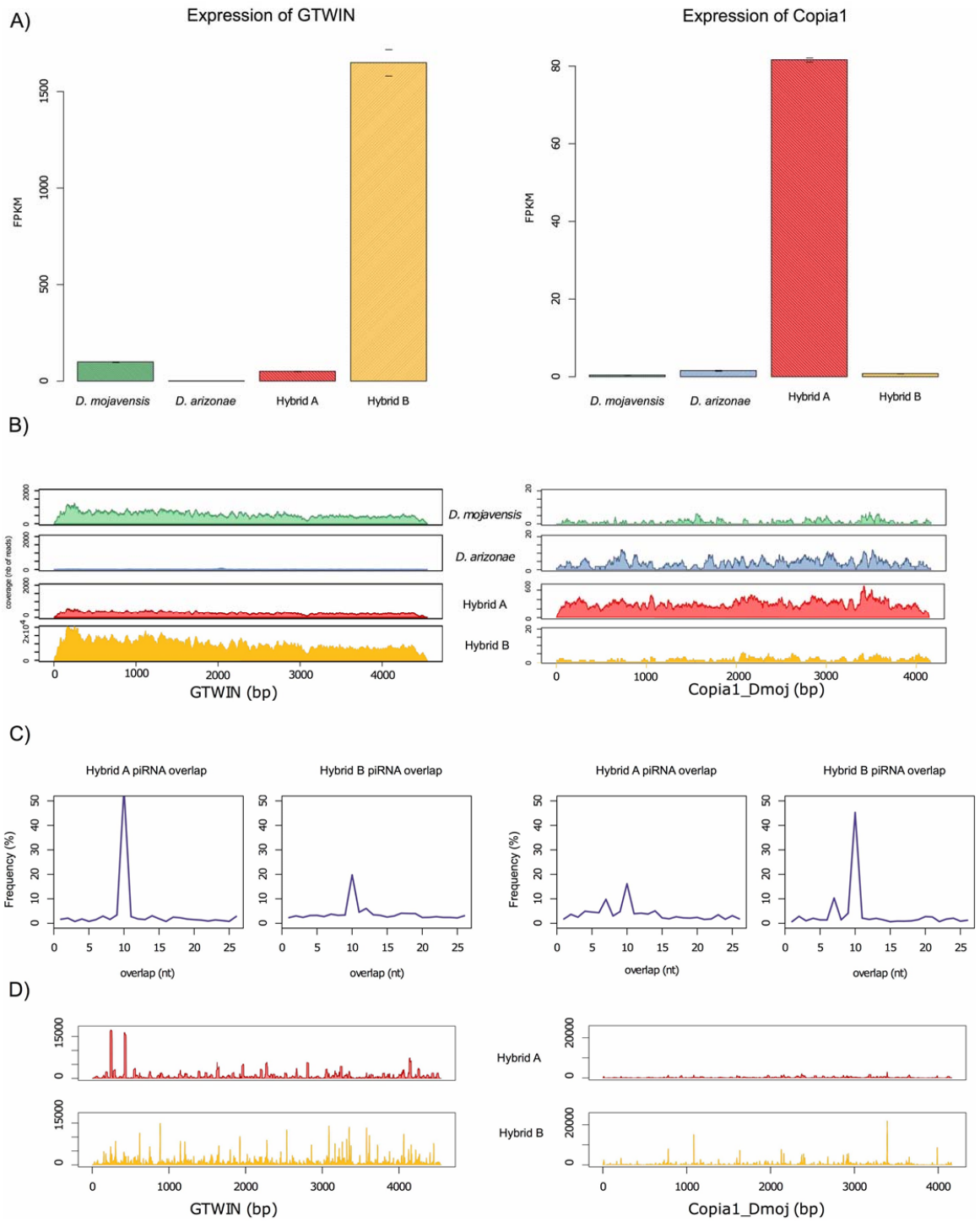
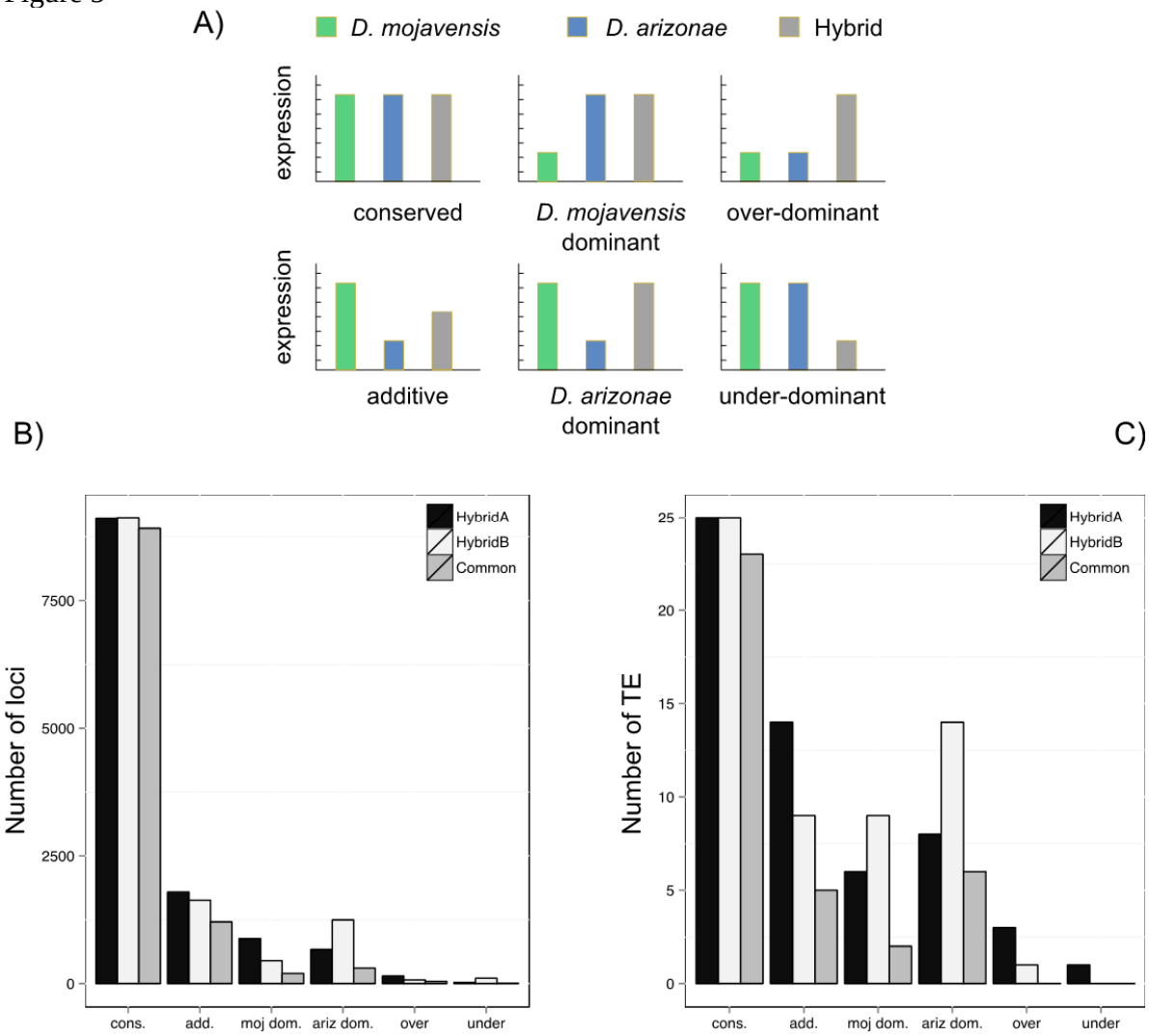

A)
Expression of GTWIN
FPKM
D. mojavensis
D. arizonae
Hybrid A
Hybrid B
Expression of Copia1
B)
D. mojavensis
D. arizonae
Hybrid A
Hybrid B
GTWIN (bp)
Copia1_Dmoj (bp)
C)
Hybrid A piRNA overlap
Hybrid B piRNA overlap
Frequency (%)
overlap (nt)
D)
Hybrid A
Hybrid B
GTWIN (bp)
Copia1_Dmoj (bp)

Figure 5


A)
D. mojavensis
D. arizonae
Hybrid
expression
conserved
D. mojavensis dominant
over-dominant
expression
additive
D. arizonae dominant
under-dominant
B)
C)
HybridA
HybridB
Common
Number of loci
7500
5000
2500
0
cons.
add.
moj dom.
ariz dom.
over
under
Number of TE
25
20
15
10
5
0

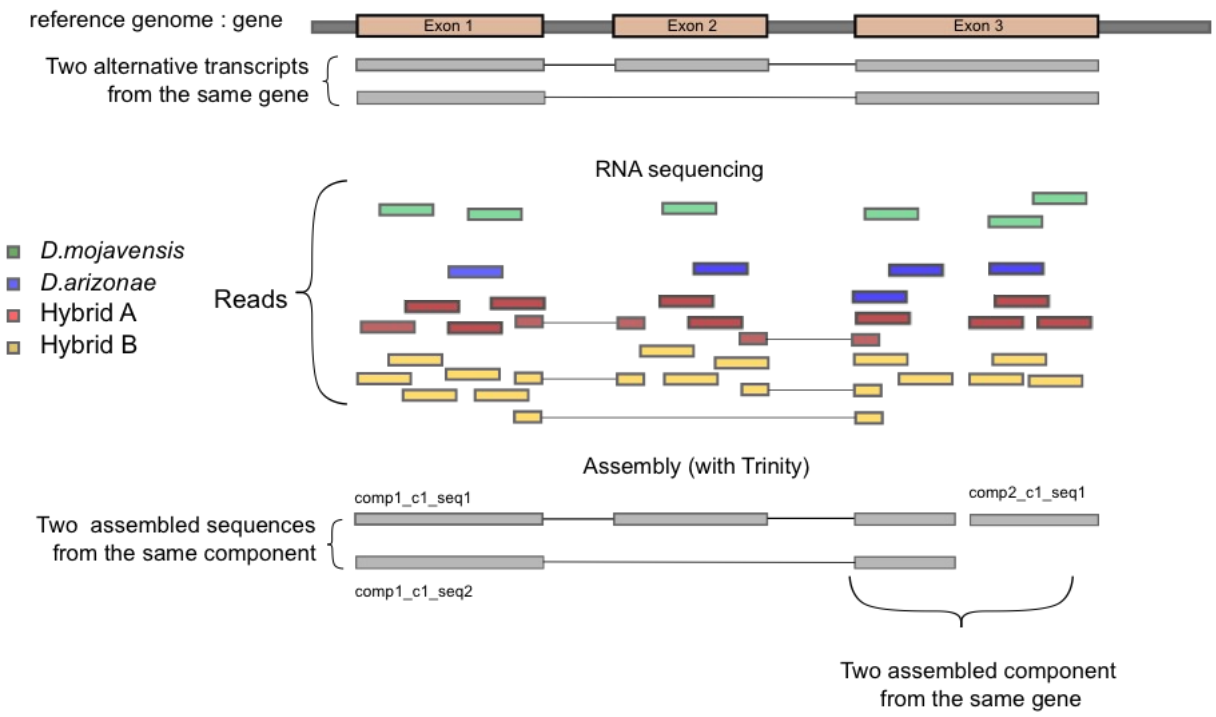
3 Supplementary Information

Title: **Identification of misexpressed genetic elements in hybrids between Drosophila-related species**

Hélène Lopez-Maestre[1,2], Elias A. G. Carnelossi[3], Vincent Lacroix[1,2], Nelly Burlet[1], Bruno Mugat[4], Séverine Chambeyron[4], Claudia M. A. Carareto[3], Cristina Vieira[1]*

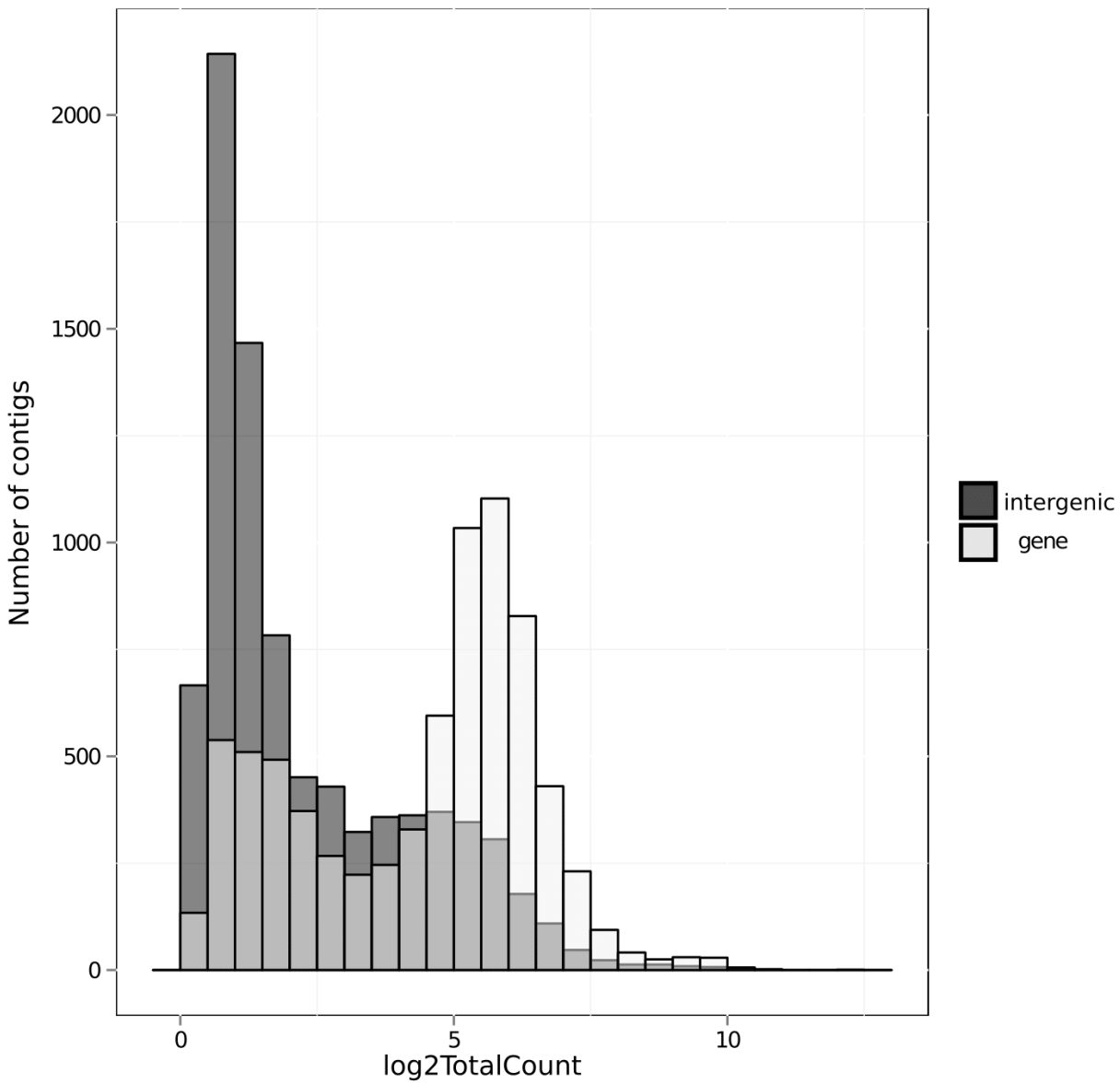
Supplementary Material

Supplementary Figures

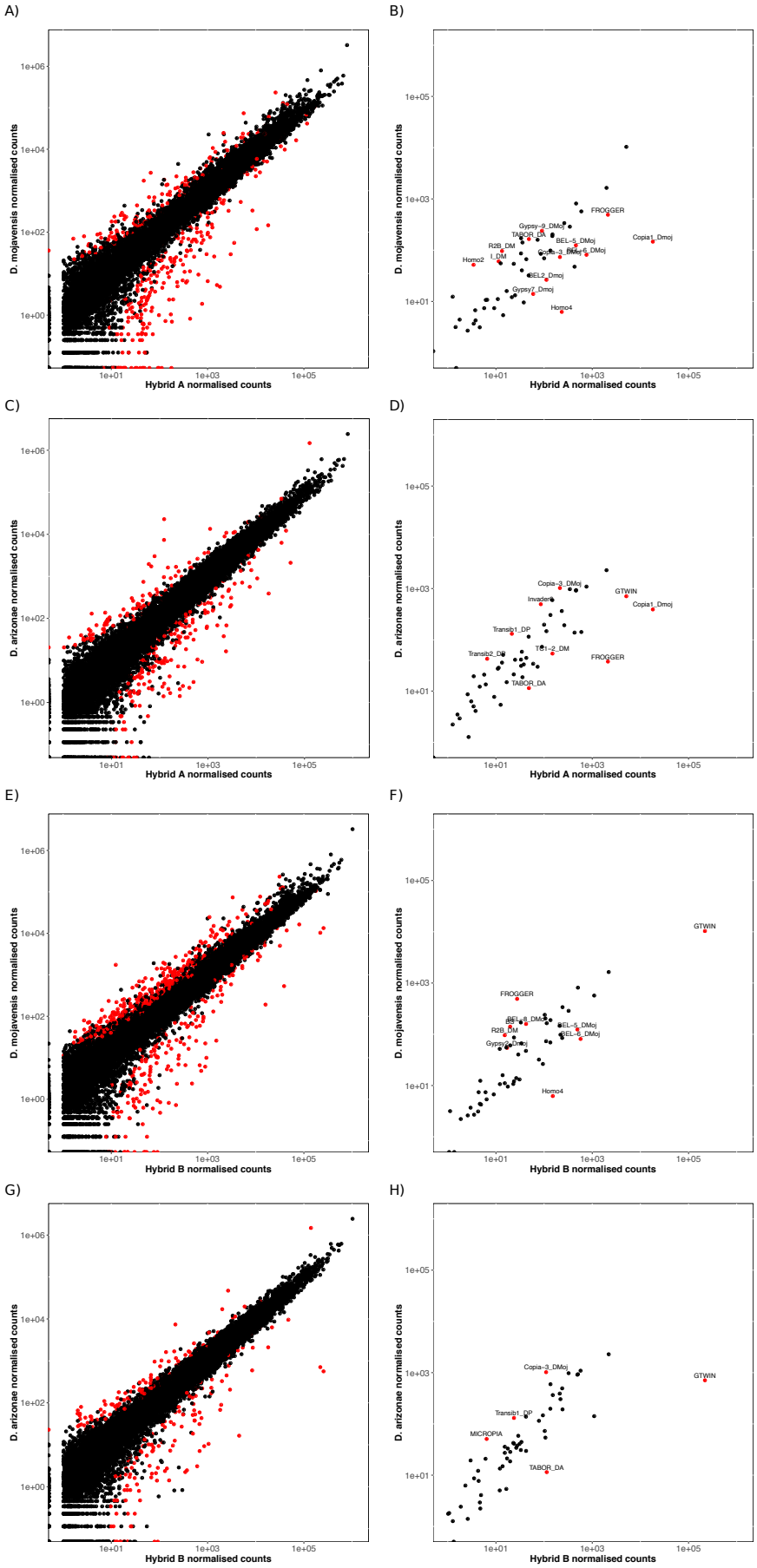


Supplementary Figure 1: Co-assembly result, example on one gene.

In this example, a gene has two alternative transcripts. Due to the coverage heterogeneity in RNAseq data, there is a lack of reads in the third exon of the gene. Thus the assembler fails in the reconstruction of the transcripts and assembled two components for one gene. The first component has two alternative sequences that cover the splicing event present in the real transcripts.

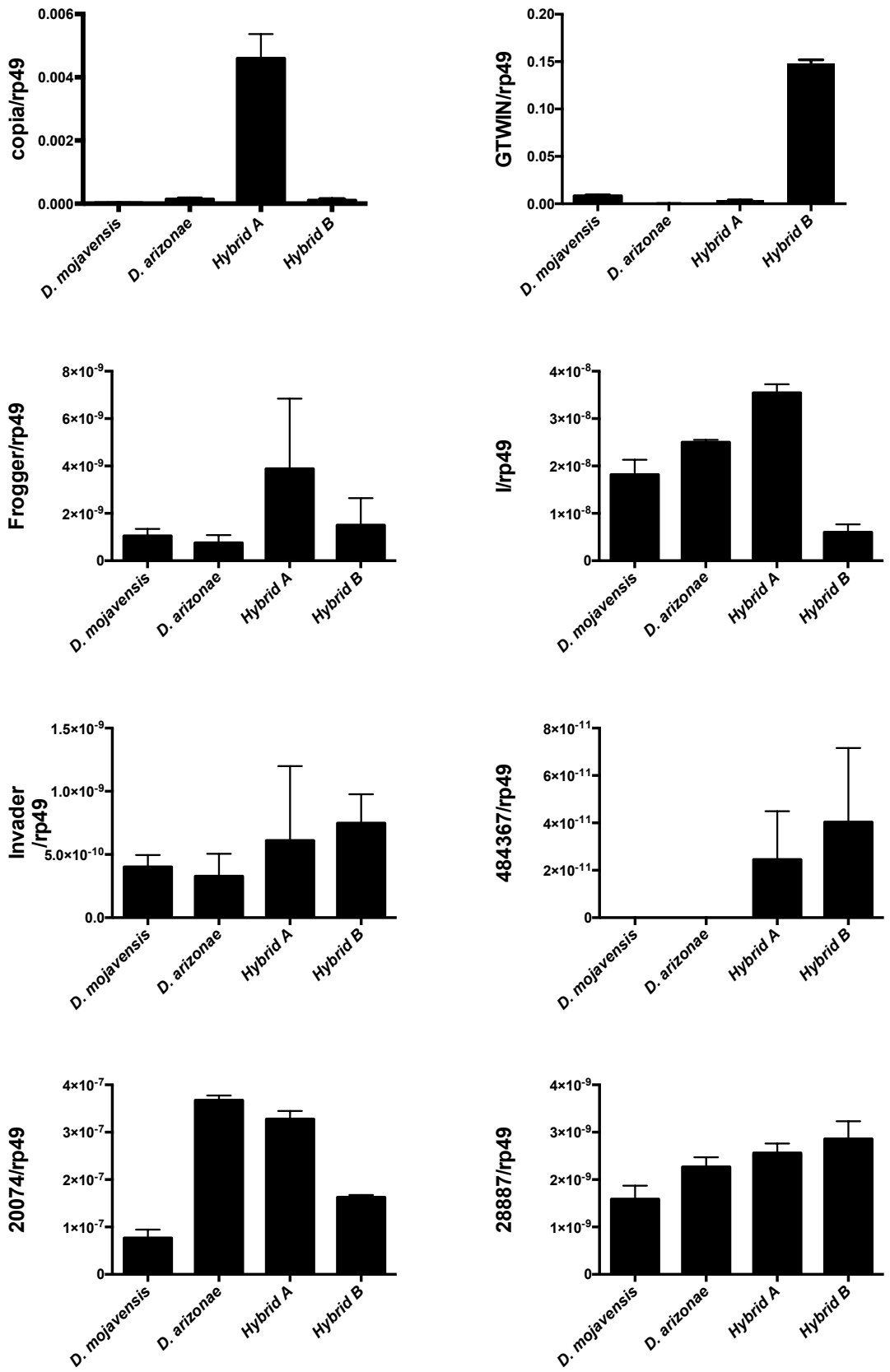


Supplementary Figure 2: Distribution of the total expression from all the samples (in log2FPKM) of the assembled components corresponding to protein-coding genes (white) or components corresponding to potential non coding RNA (darkgrey). There are two modes in this distribution, suggesting that half of the genes are highly expressed, whereas the other half are lowly expressed and could be interpreted as transcription noise, which has been previously reported with transcriptome data.

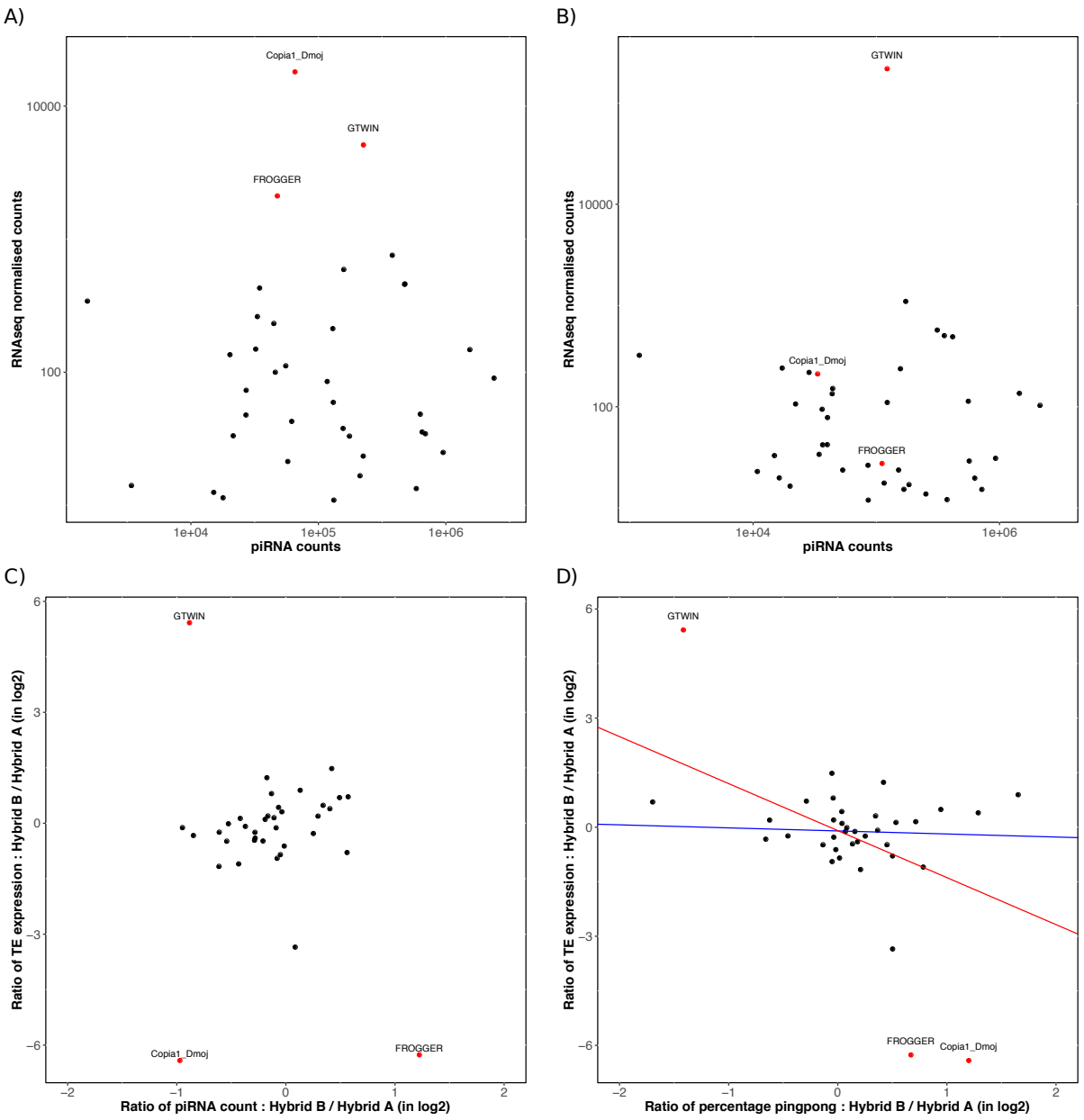

A)
B)
C)
D)
E)
F)
G)
H)
D. mojavensis normalised counts
D. arizonae normalised counts
Hybrid A normalised counts
Hybrid B normalised counts
1e+00
1e+02
1e+04
1e+06
1e+01
1e+03
1e+05
FROGGER
Copia1_Dmoj
Gypsy-9_DMoj
TABOR_DA
BEL-5_DMoj
R2B_DM
I_DM
Homo2
BEL2_Dmoj
Gypsy7_Dmoj
Homo4
GTWIN
Copia-3_DMoj
Invader1
Transib1_DP
Transib2_DP
TE1-2_DM
MICROPIA

1

Supplementary Figure 3: Scatter plot of the mRNA normalized counts of the parental line vs the hybrids, for genes (left A, C, E, G) and transposable elements (rigth B, D, F, H). Each dot represents a gene or a TE. Red dots correspond to differentially expressed genes or TEs.



Supplementary Figure 4: RTq-PCR experiments for Copia, GTWIN, Frogger, I and invader TEs, and for genes #48436, #20074 and #2887 in parental lines and hybrids. The results were in agreement with the differential expression obtained with the RNAseq data.



1

Supplementary Figure 5: piRNA analysis. Scatter plots representing le relation between piRNA amount and mRNA from TE (A- hybrid A, B hybrid B). Ratio between hybrid A/B for total piRNA (C). Ratio between hybrid A/B for secondary piRNA (C). Dots represent TEs. Red dots indicate the ones that are differentially expressed between hybrid A and B. The red regression line panel D suggest a negative correlation between the ratio of mRNA and the ration of secondary RNA (IC (r) =[-0.58 ; -0.16] with 95% confidence, the p-value associated to the t-test is 0.0011), that disappears when data from GTWIN, copia and frogger are

removed (blue line).

Supplementary tables

Supplementary Table 1: Sequences obtained by Trinity with a co-assembling procedure or without co-assembling for the four transcriptomes.

	Co-assembly	Split assembly			
		D. mojavensis	*D. arizonae*	HybridA	HybridB
# of components	21889	15807	15521	15352	14556
# of components not aligned on D. mojavensis reference genome[a]	815 (4%)	728 (5%)	1227 (8%)	908 (6%)	872 (6%)
N50[b]	2695	2562	2664	2630	2636
Coverage of the genome[c]	24.0 Mb	19.6 Mb	19.6 Mb	19.3 Mb	18.6 Mb
Mapping Back Rate[d]	98.5	98.5	98.5	98.1	98.3

a) Number of components (and %) that do not align on the reference genome of D. mojavensis with 80% of identity and 80% of their length (QC). This may correspond to chimeric sequences.
b) N50 of the assembly: The N50 length is the shortest sequence length at 50% of the assembled sequences.
c) To calculate the total length we take into account only the longest sequence per component assembled by trinity.

d) The mapping back rate corresponds to the proportion of reads mapping back to the assembled transcriptome. For the co-assembly we mapped all the reads from all the species and hybrids back to the transcriptome. For the single assemblies we mapped back only the reads from the corresponding species or hybrid.

Supplementary Table 2: Top 30 genes differentially expressed between the parental lines.

Gene or component name	Fold Change	FDR	UP	
comp15874_c0	208,6	2,5E-66	Moj	-
comp56409_c0	127,6	1,8E-52	Moj	FBgn0051075 (pyruvate metabolic process) FBgn0051076
comp21101_c1	121,3	9,9E-41	Arz	-
comp20866_c17	112,9	2,5E-32	Arz	-
comp3770_c0	111,6	1,7E-39	Arz	-
comp23924_c0	102,4	8,9E-33	Arz	-
comp23663_c0	98,9	3,8E-61	Moj	-
comp23953_c0	88,7	9,8E-79	Arz	-
FBgn0134916	80,0	5,5E-33	Moj	-
FBgn0141615	75,7	5,4E-41	Moj	FBgn0259247 (laccase 2 , chitin-based cuticle development)
comp15637_c0	70,3	5,6E-60	Moj	FBgn0041241 (sensory perception of taste)
comp20866_c10	66,5	3,3E-27	Arz	-
comp129_c1	66,0	3,3E-27	Moj	-
comp129_c0	61,0	3,6E-23	Moj	-
comp20327_c1	59,0	1,5E-29	Moj	-
comp15675_c1	57,8	1,7E-30	Moj	-
comp15950_c0	56,2	6,2E-39	Moj	FBgn0019982 (wound healing, metabolic process) FBgn0040705 (mitochondrial electron transport, NADH to ubiquinone)
comp22800_c7	53,8	7,5E-37	Arz	FBgn0264908 (neurogenesis)
FBgn0141106	53,4	4,5E-25	Arz	FBgn0033058 (neuropeptide signaling pathway)
comp17207_c1	51,1	6,3E-24	Arz	-
comp410692_c0	49,8	1,5E-20	Arz	-
comp3221_c0	49,6	6,7E-22	Arz	-

comp19601_c0	47,2	5,4E-24	Arz	-
FBgn0138205	45,9	1,2E-35	Moj	FBgn0265413
comp22637_c8	45,8	2,2E-18	Moj	-
comp20485_c2	43,4	3,9E-25	Arz	-
comp18701_c0	43,2	4,3E-35	Moj	FBgn0032536 (protelysis) FBgn0051716 (regulation of JAK-STAT cascade, centrosome organization)
comp3149_c0	42,7	1,4E-18	Arz	-
comp19494_c0	42,3	5,1E-19	Arz	FBgn0030699 (adult somatic muscle development, regulation of transcription) FBgn0261545 (determination of adult lifespan) FBgn0031897 FBgn0040005 (regulation of GTPase activity)
comp21818_c14	41,7	4,5E-20	Arz	-

Supplementary Table 3: Genes differentially expressed between the parental lines

	Total	Up in *D.mojavensis*	Up in *D.arizonae*
All Genes	1229	684	546
Unique Genes	486	270	166
Multi Genes	138	82	56
intergenic	534	264	270
not in genome	71	18	53

Supplementary Table 4: TEs differentially expressed between *D. mojavensis* and *D.arizonae*

Type of TE	TE Name	Fold Change	FDR	UP
TIR	Homo4	27.7	4.7E-29	Moj
LTR	FROGGER	12.0	1.6E-26	Arz
LTR	Copia-3_DMoj	11.8	1.0E-20	Moj
LTR	GTWIN	11.6	2.3E-19	Arz
LTR	TABOR_DA	10.4	2.5E-15	Arz
LTR	BEL-6_DMoj	9.3	2.6E-10	Moj
TIR	Transib1_DP	8.0	1.2E-07	Moj
TIR	PARIS	6.5	1.0E-05	Arz
LTR	MICROPIA	6.0	8.1E-05	Moj
LTR	BEL-5_DMoj	5.8	2.7E-07	Moj
LINE	BS	5.7	2.1E-05	Arz
LTR	Invader6	5.1	5.5E-06	Moj
TIR	Homo2	4.9	3.9E-03	Arz
LTR	BEL-8_DMo	4.7	1.4E-09	Arz
LTR	BEL2_Dmoj	4.4	1.6E-04	Moj
TIR	Homo1	4.1	3.5E-03	Arz
LTR	TC1-2_DM	3.5	2.5E-09	Arz
LTR	Gypsy-9_DMoj	3.0	2.3E-04	Arz
Helitron	Helitron-1N1_DVir	2.9	5.6E-03	Moj
TIR	TRANSIB1	2.9	5.4E-04	Moj

Supplementary Table 5: Genes differentially expressed between the hybrid lines

	Total	Up in hybrid A	Up in hybrid B
All Genes	89	62	27
Unique Genes	40	32	8
Multi Genes	8	5	3
intergenic	34	19	15
not in genome	7	5	2

Supplementary Table 6: Top 30 genes differentially expressed between the hybrid lines

Gene or component name	Fold Change	FDR	UP	*D.melanogaster* ortholog (function)
comp23953_c0	94.2	5.9E-82	B	-
comp23750_c1	28.2	6.9E-42	B	-
comp16843_c1	15.6	2.6E-08	A	-
comp20770_c4	14.8	1.2E-07	B	-
comp23819_c0	13.6	2.0E-12	A	-
FBgn0136755	13.3	3.7E-08	B	-
FBgn0143900	13.1	3.2E-11	A	FBgn0054038
comp18846_c0	12.2	2.0E-07	A	-
comp20770_c6	12.1	1.6E-06	B	-
comp22800_c7	11.5	4.4E-14	B	FBgn0264908 (neurogenesis)
comp20770_c2	10.8	8.1E-05	B	-
comp16289_c1	9.9	1.2E-05	A	-
comp24148_c0	8.9	5.3E-05	B	-
comp20770_c7	8.7	6.7E-04	B	-
comp21101_c1	8.6	3.6E-08	B	-
comp21244_c9	8.5	9.6E-06	A	-
comp16249_c1	8.1	9.1E-05	A	-
comp24122_c5	8.0	3.3E-04	B	-
FBgn0143905	7.9	5.9E-10	A	FBgn0053680
FBgn0146209	7.7	1.1E-03	A	-
comp15590_c0	7.7	1.4E-03	A	-
comp15705_c0	7.5	2.6E-03	A	-
comp20019_c0	7.4	2.3E-09	B	-
FBgn0134214	7.3	1.6E-14	A	FBgn0265296 (neuron projection morphogenesis)
comp23342_c9	7.1	3.2E-04	A	-
comp16949_c1	6.8	4.6E-03	B	-
comp14298_c0	6.7	3.7E-03	B	-
comp22234_c9	6.6	6.2E-06	A	-
comp18846_c2	6.5	1.3E-03	A	-
FBgn0132849	6.4	2.7E-09	A	FBgn0039201

Supplementary Table 7: TEs differentially expressed between hybrids

Type of TE	TE Name	Fold Change	FDR	normalized counts in Hybrid A	normalized counts in Hybrid B	UP
LTR	FROGGER	62.1	5.5E-76	2110	27	A
LTR	Copia1_Dmoj	55.3	2.7E-38	18053	211	A
LTR	GTWIN	32.7	1.5E-38	5092	218154	B

Supplementary Table 8: Expression data on thirty genes implicated on piRNA biogenesis. None was differentially expressed between hybrids, and only two were differentially expressed between the parental lines. (* p< 0.05).

	Normalised counts in *D.mojavensis*	Normalised counts in *D.arizonae*	Normalised counts in Hybrid A	Normalised counts in Hybrid B	FoldChange between parental lines	FoldChange between hybrids
archipelago	41957	54588	54636	73544	1.3	1.3
Armitage	35341	38519	51264	51201	1.1	1.0
Aubergine	6991	8613	7047	6518	1.2	1.1
Brother_of_Yb	9394	9603	10509	11508	1.0	1.1
cutoff	39168	34880	28427	28756	1.1	1.0
helicase_at_25 E	23051	32452	19748	22519	3.2	1.1
Hen1	19535	24719	37437	36735	1.4	1.1
interruptus_cubitus	777	686	759	446	1.2	1.0
Krimper	3106	2415	4889	4319	1.1	1.6
maelstrom	14300	21927	14430	17035	1.3	1.1
minotaur	363	108	129	181	1.5	1.2
PanoramixA	3362	1759	1852	1987	*3.1	1.4
PanoramixB	3157	4638	4468	4910	1.8	1.1
piwi	39997	62132	48770	50212	1.4	1.1
qin	9917	15763	17512	14859	1.5	1.0
shutdownA	4736	5394	4053	4476	1.6	1.2
shutdownB	2850	3851	2838	2840	1.1	1.1
Sister_of_Yb	3254	1060	2752	1829	1.3	1.0
spindle_E	11133	12011	14216	15057	2.7	1.4
tapas	14454	15546	18392	22548	1.1	1.1
tejas	4499	3708	4100	4505	1.1	1.2
tudor	14	1	1	1	1.2	1.1
vret	14989	36641	21552	22272	*2.3	1.0
Yb	721	590	794	527	1.2	1.5
zucA	6021	4560	2460	3114	1.3	1.2

3

Détection de SNP dans les données RNAseq sans génome de référence

Sommaire

1 Avant-propos 39

2 Article 1 : Identification of misexpressed genetic elements in hybrids between Drosophila-related species 40

3 Supplementary Information 80

1 Avant-propos

KiSplice est une méthode initialement développée pour détecter des variants d'épissage dans des données RNA-seq sans génome de référence (et donc pour des espèces modèles ou non modèles). Bien que principalement développé à Lyon par l'équipe Baobab du *Laboratoire de Biométrie et Biologie Évolutive* (LBBE), KisSplice est issu d'une collaboration de plusieurs équipes de recherches dans le cadre de l'ARN *Colib'read*, qui propose des développements de méthodes basées directement sur les lectures séquencées pour répondre à différents problèmes biologiques (détection de SNP, d'épissage, d'inversions génomiques etc.)

KisSplice commence par construire un graphe de de Bruijn, à partir des lectures séquencées, et recherche des motifs spécifiques, des "bulles", créées par la présence de variants possédant un contexte commun (d'au moins k nucléotides). Selon les caractéristique de cette "bulle", on peut différencier celles correspondant à des variants d'épissages, des indels ou des SNP.

Mon rôle dans ce projet a été de clarifier les points forts et les limites de l'utilisation de KisSplice pour l'identification des SNP sur différents jeux de données réels. J'ai également participé au développement de KisSplice2RefTranscriptome (K2RT), un outil de post-traitement des SNP trouvés par KisSplice, permettant de prédire leur impact sur les séquences protéiques. Cette étude est issue d'une collaboration entre différentes équipes du LBBE : l'équipe Baobab qui a développé et testé KisSplice sur les données humaines, les équipes *Génétique et Évolution des interactions Hôtes-Parasites* et *Éléments transposables, Évolution, Populations* qui ont permis de tester la pipeline sur des données réelles et d'effectuer des validations expérimentales de SNP prédits, ainsi que l'équipe *Statistique en Grande Dimension pour la Génomique* pour la modélisation statistique.

Nous proposons dans l'article qui suit une pipeline utilisant les données RNA-seq permettant d'identifier et quantifier des SNP, de prédire leur impact sur la séquence d'acide aminés, mais aussi d'identifier les SNP spécifiques d'une condition lorsque l'on compare plusieurs conditions biologiques. Nous avons utilisé des données humaines, issues des projets 1000 Genomes et Geuvadis, pour estimer la sensibilité et la précision de KisSplice, et plus généralement de l'ensemble du pipeline.

2 Article 2 : SNP calling from RNA-seq data without a reference genome : identification, quantification, differential analysis and impact on the protein sequence

Nucleic Acids Research Advance Access published July 25, 2016

Nucleic Acids Research, 2016 **1**
doi: 10.1093/nar/gkw655

SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence

Hélène Lopez-Maestre[1,2], Lilia Brinza[3], Camille Marchet[4], Janice Kielbassa[5], Sylvère Bastien[1,2], Mathilde Boutigny[1,2], David Monnin[1], Adil El Filali[1], Claudia Marcia Carareto[6], Cristina Vieira[1,2], Franck Picard[1], Natacha Kremer[1], Fabrice Vavre[1,2], Marie-France Sagot[1,2] and Vincent Lacroix[1,2,*]

[1]Université de Lyon, F-69000, Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622 Villeurbanne, France, [2]EPI ERABLE - Inria Grenoble, Rhône-Alpes, [3]PT Génomique et Transcriptomique, BIOASTER, Lyon, France, [4]Université de Rennes, F-35000 Rennes; équipe GenScale, IRISA, Rennes, [5]Synergie-Lyon-Cancer, Universite Lyon 1, Centre Leon Berard, Lyon, France and [6]Department of Biology, UNESP - São Paulo State University, São José do Rio Preto, São Paulo, Brazil

Received December 22, 2015; Revised July 05, 2016; Accepted July 11, 2016

ABSTRACT

SNPs (Single Nucleotide Polymorphisms) are genetic markers whose precise identification is a prerequisite for association studies. Methods to identify them are currently well developed for model species, but rely on the availability of a (good) reference genome, and therefore cannot be applied to non-model species. They are also mostly tailored for whole genome (re-)sequencing experiments, whereas in many cases, transcriptome sequencing can be used as a cheaper alternative which already enables to identify SNPs located in transcribed regions. In this paper, we propose a method that identifies, quantifies and annotates SNPs without any reference genome, using RNA-seq data only. Individuals can be pooled prior to sequencing, if not enough material is available from one individual. Using pooled human RNA-seq data, we clarify the precision and recall of our method and discuss them with respect to other methods which use a reference genome or an assembled transcriptome. We then validate experimentally the predictions of our method using RNA-seq data from two non-model species. The method can be used for any species to annotate SNPs and predict their impact on the protein sequence. We further enable to test for the association of the identified SNPs with a phenotype of interest.

INTRODUCTION

Understanding the genetic basis of complex phenotypes remains a central question in biology. A classical approach consists in genotyping a large number of individuals in a population based on a pre-specified catalog of variants, and in associating their genotypes to the studied phenotype. This type of approach can be applied to many loci at once, or even genome wide, through what has been called genome wide association studies (GWAS). These methods have been successfully adopted for human and model species. However, the total cost of GWAS remains very high, and the current framework cannot be applied to non-model species for which genomic resources are sparsely or not available. The recent progress in sequencing technologies together with the recent developments in assembly algorithms are largely changing this view. It can now be envisioned to search for variants associated with a phenotype using NGS data only, without relying on pre-existing genomic resources (that have potential limitations). A possible procedure, applicable to model or non-model species, consists in: (i) sequencing the genome; (ii) assembling it; (iii) identifying the SNPs; (iv) genotyping individuals and (v) associating genotypes with phenotypes. However, such a procedure remains costly and still presents the classical problems of sequential pipelines, namely the potential to accumulate experimental and computational errors at each step.

If the purpose of the study is to identify the variants related to a phenotype, the procedure can be simplified in many ways. First, SNPs can be called *de novo* from the reads, without separating the steps of assembly and SNP calling. Second, cost effective methods like exome or transcriptome sequencing may be adopted as the full genome is not al-

*To whom correspondence should be addressed. Email: vincent.lacroix@univ-lyon1.fr

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on September 12, 2016

ways necessary. Third, pooling individuals may be an attractive option if genotyping is not required. These options have been explored individually and give promising results. *De novo* assembly of SNPs is now computationally possible (1–3). The clear advantage is that it can be applied to non-model species, where no reference genome is available. Even in the case where a reference genome is available, these methods still give good results compared to mapping-based approaches, compensating their lower sensitivity by an ability to call more variants in repeated regions. Transcriptome sequencing is already used in several projects, both in the context of model species (4) and non-model species (5–7). In both cases, it was shown that the SNP calling methods could be tailored to have a good precision, meaning that most of the reported SNPs are true SNPs. However, their recall (i.e. capacity to exhaustively report all SNPs) remains to be clearly determined. Clearly, only SNPs from transcribed regions can be targeted, but they arguably correspond to those with a more direct functional impact. Using RNA-seq technology largely reduces the cost of the experiment, and the obtained data concurrently mirror gene expression, the most basic molecular phenotype. RNA-seq experiments may also provide very high depth at specific loci and therefore allow to discover infrequent alleles in highly expressed genes. Finally, pooling samples is already extensively used in DNA-seq (sometimes termed Pool-seq) (8). The main advantage of this method is that it clearly decreases costs, as library preparation for bar-coding is nowadays approximately the same price as sequencing. The drawback is that genotypes cannot be derived anymore. Instead, we have access to the allele frequency in the population, a result known as the allelotype. In this work, we present a method for the de novo identification, differential analysis and annotation of variants from RNAseq data in non-model species. It takes as input RNA-seq reads from at least two conditions (e.g. the modalities of the phenotype) with at least two replicates each, and outputs variants associated with the condition. The method does not require any reference genome, nor a database of SNPs. It can therefore be applied to any species for a very reasonable cost. We first evaluated our method using RNA-seq data from the human Geuvadis project (9). The great advantage of this dataset is that SNPs are well annotated, since the selected individuals were initially included in the 1000 genomes project (10). This enables to clarify what is the precision and recall of our method, and how it compares to methods which require a reference genome or a reference transcriptome.

We then applied our method in the context of non-model species. First we focused on *Asobara tabida*, an hymenoptera that exhibits contrasted phenotypes of dependence to its symbiont. Using RNA-seq data from two extreme modalities of the phenotype, we were able to establish a catalog of SNPs, stratify them by their impact on the protein sequence, and assess which SNPs had a significant change of allele frequency across modalities. We further selected cases for experimental validation, and were able to confirm that the SNPs were indeed condition specific. We then applied our method on two recently diverged *Drosophila* species, *D. arizonae* and *D. mojavensis*. These species can still produce hybrids that are sterile. In this case, our method identifies differences of 1 nt, which are not SNPs but divergences. On this system also, we were able to validate experimentally that the loci we identify were truly divergent.

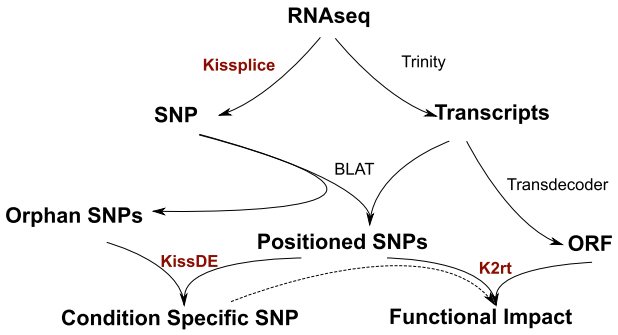


Figure 1. With fasta/fastq input from an RNA-seq experiment, SNPs are found by KISSPLICE without using a reference. As KISSPLICE provides only a local context around the SNPs, a reference can be built with TRINITY, and SNPs can be positioned on whole transcripts. Some SNPs that do not map on the transcripts of TRINITY, called orphan SNPs, are harder to study but can still be of interest. We propose a statistical method, called KISSDE, to find condition-specific SNPs (even if they are not positioned) out of all SNPs found. Finally, we can also predict the amino acid change for the positioned SNPs, and intersect these results with condition-specific SNPs using our package KISSPLICE2REFTRANSCRIPTOME (K2RT).

We outline that, even though the case studies presented in this paper include two replicates, the method can be applied to any number of replicates. Larger cohorts can be helpful to narrow down the list of SNPs likely to be really causal for the phenotype. Our key contribution is that we are able to produce a list of SNPs stratified by their impact on the protein sequence, and ranked by difference of expressed allele frequency across conditions. This list can be further mined for candidates to follow up experimentally.

All the methods presented in this paper are implemented in software that are freely available at http://kissplice.prabi.fr/TWAS. In particular, the statistical procedure that we developed is available through an R package, KISSDE, which is of general interest for researchers who have obtained read counts for pairs of variants in a set of conditions and wish to test if these counts reflect the specificity of the variant in a particular condition.

MATERIALS AND METHODS

Overview

We present here a collection of methods which can be used together to produce, from RNA-seq data alone, a list of condition-specific SNPs, stratified by their predicted impact on the protein. Figure 1 summarises the different steps.

TRINITY, TRANSDECODER and BLAT are third-party software. KISSPLICE was published recently (11), KISSDE and KISSPLICE2REFTRANSCRIPTOME (K2RT) are methods we introduce in this paper.

De novo identification of SNPs

KISSPLICE (11) is a software initially designed to find alternative splicing events (AS) from RNA-seq data, but which

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on September 12, 2016

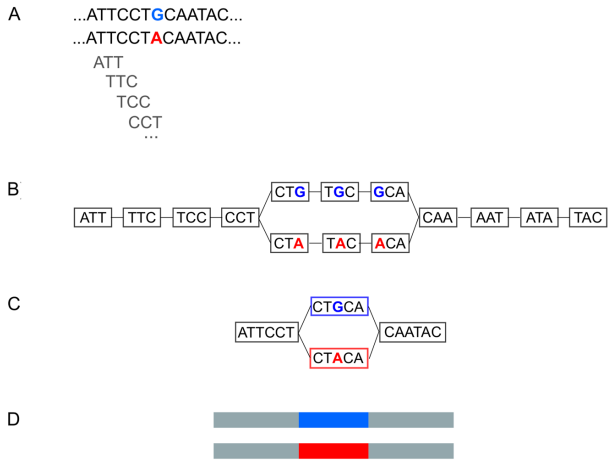


Figure 2. (**A**) A SNP present in two alleles in the data. (**B**) The de Bruijn Graph derived from the data. For the sake of simplicity of exposition, we draw here with $k = 3$. In practice, $k = 41$. (**C**) A compressed de Bruijn graph can be obtained by merging nodes with a single outgoing edge with nodes with a single incoming edge. This compression step is lossless. (**D**) The two paths in the compressed de Bruijn graph correspond to the two alleles of the SNP.

also outputs indels and SNPs. We present here its functionality for SNP detection. The key concept, initially introduced in Peterlongo *et al.* (12) and later used in Iqbal *et al.* and Uricaru *et al.* (1,2) is that a SNP corresponds to a recognisable pattern, called a *bubble*, in a de Bruijn graph (DBG) built from the reads. De Bruijn graphs are widely used data structures in de novo assembly (13–15), as they are well tailored for large amounts of short reads. In our case, DBGs are especially appealing because they model explicitly each nucleotide, a required feature to capture SNPs. The nodes of the graph are words of length k, called k-mers. There is an edge between two nodes if the suffix of length $k - 1$ of the first k-mer is identical to the prefix of length $k - 1$ of the second k-mer. The DBG that is built from two alleles of a locus will therefore correspond to a pair of vertex-disjoint paths in the graph, which form the bubble. Unlike AS events and indels, bubbles generated by SNPs have two paths of equal length (Figure 2B). Linear paths of the DBG can be further compressed in a single node without loss of information (Figure 2C).

In the special case where there are two SNPs located less than k nt apart on the genome, they will be reported in the same bubble (Supplementary Figure S1). In the case where the two SNPs are perfectly linked, a single bubble is reported. If they are partially linked, each haplotype will correspond to a path, and KISSPLICE will report all pairs of paths. In this case, the number of bubbles does not correspond to the number of SNPs, but to the number of pairs of observed haplotypes. Supplementary Figure S2 illustrates the case of two SNPs and four haplotypes.

KISSPLICE consists in essentially three steps: (i) building the DBG from the RNA-seq reads; (ii) enumerating all bubbles in this graph and (iii) mapping the reads to each path of each bubble to quantify the frequency of each variant. Particular attention was paid to both the memory (16,17) and time (18) requirements of the pipeline. KISSPLICE was able to process 200M reads of 2×75 nt in 20 hours, with less than 16GB of RAM.

Filtering out sequencing errors and inexact repeats

SNPs correspond to bubbles in the de Bruijn graph derived from the reads. However, not all bubbles in the DBG correspond to SNPs. Essentially two types of false positives can be found: sequencing errors and inexact repeats. RNA editing sites may also be mistaken for SNPs but in practice, these correspond to a few cases only, that we discuss in the Results section.

Sequencing errors may generate bubbles in the DBG. A distinctive feature that helps to discriminate them from true variants is that one path of the bubble is expected to be poorly covered. In practice, a common way to filter out sequencing errors when dealing with DNA-seq data is to remove all rare k-mers (seen less than a given number of times) prior to the DBG construction. This simple strategy, implemented for instance in DISCOSNP, is however not sufficient when dealing with RNA-seq data. Since the coverage depends on gene expression, it is therefore very unequal across genes, and the cut-off should be adapted to each gene. To account for this constraint, we introduced a relative cut-off, which enables to remove edges in the DBG that are supported by less than a percentage of all counts outgoing from (or incoming to) the same node. This enables to remove sequencing errors even in highly expressed genes (Figure 3). Clearly, the drawback of these cut-off strategies is that rare variants will be filtered out because they will be mistaken for sequencing errors. Our ability to detect rare variants is therefore limited by this critical parameter. We set the cut-off to 5%. This cut-off corresponds to a good trade-off between precision and recall (Supplementary Figure S3).

Inexact genomic repeats may also generate bubbles in the DBG (Figure 4). This is the case for instance for recently diverged paralogs which still share a lot of sequence similarity and hence may differ locally by one nucleotide flanked by k conserved nucleotides. This is also the case for other types of repeats, including inexact tandem repeats or transposable elements which may be present in the UTRs and introns of genes. In principle, introns are not present in RNA-seq data, but in practice, whatever the protocol used to filter out pre-mRNA, a proportion of at least 5% remains (19).

The question of discriminating SNPs from inexact repeats has already been addressed in the literature in the case of unpooled data. Romiguier *et al.* (5) propose to use the idea that loci corresponding to recently diverged paralogs should present an excess of heterozygous sites. This idea cannot be employed in our case since we want our method to be able to deal with pooled data, where we cannot genotype individuals.

Repeats present in a large number of copies (like transposable elements, or large families of paralog genes) generate a large number of bubbles which are false positives. However, these bubbles have a specific feature that we can use to discriminate them from the others: they are branching (Figure 4). The more (inexact) copies in the repeat family, the higher the number of branches in each bubble. In order to filter them out, we introduced a parameter b, which corresponds to the maximum number of branches allowed.

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on September 12, 2016

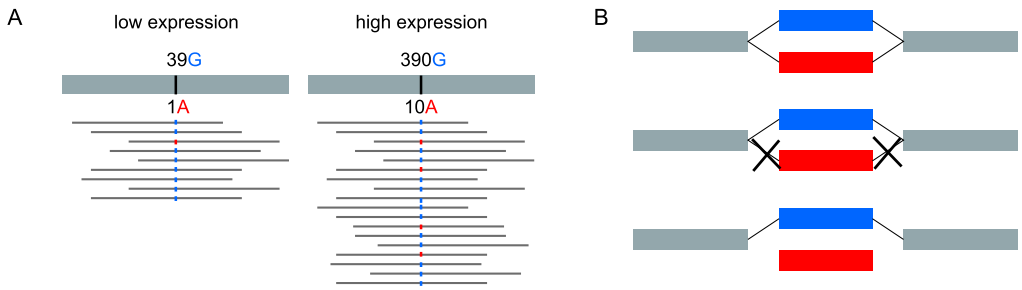


Figure 3. Sequencing errors and rare variants generate bubbles in DBGs with very unbalanced path coverage. (**A**) For ease of exposition of the concept, we represent here the reads mapping to a reference genome. Applying an absolute cutoff would remove the sequencing error for a poorly expressed gene, but not for a highly expressed gene. (**B**) Applying a relative cutoff of 5% in the DBG removes one or two edges from the red path and hence prevents this bubble from being found.

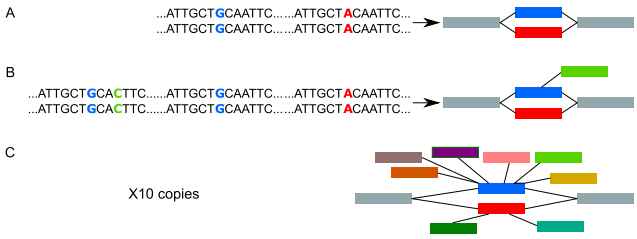


Figure 4. Two inexact repeats give rise to a pattern in the DBG that resembles a SNP (**A**). Very often, repeats are present in more than two copies (**B**) and therefore generate branching bubbles. Bubbles with more than five branches (**C**) are filtered out.

If one path of the bubble has more than b branches, then the bubble is filtered out. In practice, we set this parameter to 5, which appeared to be a good trade-off between recall and precision as shown in Supplementary Figure S3.

Repeats present in a small number of copies are not filtered out by this criterion. Some can be filtered by focusing on bubbles whose path length is strictly $2k + 1$, not larger. We found that this simple strategy was efficient and we used it in this work. It can however be modified in KISSPLICEwith the s parameter, which we recommend if the purpose is to find multiple SNPs. In any case, most inexact repeats are actually filtered out at the next step of the pipeline, when we test for the enrichment of one variant in one condition (as described in the Statistical analysis section). Indeed, most repeats do not have expression levels that are condition-specific. The ones that are *not* filtered out at this step correspond to paralogous genes, where one copy is more expressed in the first condition and the second copy is more expressed in the other condition. Although these are not SNPs, we can argue that they are still relevant candidates for an association study aiming at proposing causes for the difference of phenotype.

Predicting the impact of SNPs on the protein sequence

KISSPLICE predicts SNPs, but outputs only a very local context around the SNP. In order to predict the amino acid change it causes, if any, we need to place the SNP in a larger genomic context. For this, we relied on a widely used global transcriptome assembler: TRINITY (15), which takes as input RNA-seq reads and outputs contigs that correspond to either full-length transcripts (if the expression level of the transcript is sufficient) or to fragments of transcripts. The results of KISSPLICE were aligned onto the transcripts predicted by TRINITY using BLAT (20). Concurrently, we Fdsearched for coding potential in the transcripts using TRANSDECODER. Once we had the location of the SNP within the transcript and the location of the open reading frame (ORF), we could assess if the SNP was located within the CDS or not, and if so, if it was a synonymous or non synonymous SNP. In the case where no ORF was predicted for the transcript, we concluded that the SNP was within a non coding region. In practice, this can correspond to a non coding RNA, a UTR or an intron. Prediction of the amino acid change of a SNP was included in a Python package, called KISSPLICE2REFTRANSCRIPTOME (K2RT), which takes as input a set of predicted ORFs (bed format), the output of KISSPLICE (fasta format), and a mapping of the results of KISSPLICE to the transcripts (psl format). Importantly, TRINITY, TRANSDECODER and BLAT are third party software which can be replaced by others, provided the exchange formats are respected (bed and psl).

In the case where a SNP mapped to several TRINITY transcripts, we reported the amino acid change of the SNP in each transcript. This happened in particular when a SNP was located in a constitutive exon of a gene that gave rise to multiple alternative transcripts through alternative splicing. We further show in the Results Section that our ability to call SNPs both in constitutive exons and alternative exons is a strong advantage of our method against others that first map the reads to the assembled transcriptome and then call SNPs using a genotyper.

In the case where a SNP mapped to no transcript, then it could not be treated by K2RT and it was filtered out. Those SNPs were called orphan SNPs. They were mostly located in poorly expressed genes and/or highly repeated regions. Indeed, repeated regions are notoriously difficult to assemble. When repeated regions are located within genes, they may either generate chimeric transcripts in the assembly if the assembler is too permissive, or a series of truncated short contigs if the assembler is too conservative. By default, TRINITY does not output contigs shorter than 200 nucleotides. Because these contigs are highly enriched in repeats and poorly expressed genes, it explains the origin of the majority of our orphan SNPs.

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on September 12, 2016

As mentioned in the model section, the number of bubbles does not always correspond to the number of SNPs. In the case of SNPs located less than k nucleotides apart, the number of bubbles corresponds to the number of pairs of haplotypes out of the total number of haplotypes. The same SNP may therefore be present in multiple bubbles. When mapping the bubbles to a reference transcriptome, it is possible to remove this redundancy and count the true number of SNPs. Indeed, if two bubbles map to the same transcript at the same location, then it means that they refer to the same SNP, and we count it only once.

The software versions that we used were: TRINITY r20140717, TRANSDECODER v2.0.1, BLATSUITE36, KISSPLICE v2.4, KISSPLICE2REFTRANSCRIPTOME v1.0.

All were used with default parameters. We set the minimum query coverage to 90% in K2RT. Changing this from 70% to 90% only marginally affected our results.

A critical parameter in de novo assembly is the k-mer size. In TRINITY, this value is set to 25 and cannot be modified. In KISSPLICE the default value is 41 as we found it is a good compromise between recall and precision. We also tested 25 and this resulted in an increase of 10% in recall but a decrease of 10% in precision (Supplementary Figure S3). For advanced users interested in obtaining a more exhaustive list of candidates (hence optimising recall), we recommend to decrease the value of k in KISSPLICE.

Statistical analysis

Testing the association between a variant and a condition. Given the number of SNPs (n) and the number of replicates (m), our data set is a count matrix of size $2n \times m$, with two lines corresponding to one SNP (upper and lower path representing the two different alleles with one nucleotide differing between both paths). For each individual, we aimed to compare read counts per allele and per condition. As we worked with biological replicates, several sources of variance were added and the variance parameter of the Poisson distribution was in general not flexible enough to describe the data (21,22). Hence, our statistical analysis adopted the framework of count regression with Negative Binomial distribution.

We considered a two-way design with interaction, with *alleles* and *experimental conditions* as main effects. Following the Generalized Linear Model framework, the expected intensity of the signal was denoted by λ_{ijk} and was decomposed as:

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

where μ is the local mean expression of the transcript that contains the SNP, α_i the effect of allele i on the expression, β_j the contribution of condition j to the total expression, and $(\alpha\beta)_{ij}$ the interaction term. In order to properly model the variability of the data that are characterised by overdispersion (as in any RNASeq data (21,22)), we considered the Negative Binomial distribution. In this setting, Y_{ijk} denotes the counts of a sample k with allele i in condition j. We assume that:

$$Y_{ijk} \sim NB(\lambda_{ijk}, v_{ijk}),$$

with λ_{ijk} defined as above. With this model, the variance of the observations becomes:

$$v_{ijk} = \lambda_{ijk} + \phi \times \lambda_{ijk}^2,$$

with ϕ the over-dispersion, which is the excess of variance seen in the data in comparison to a Poisson distribution.

Due to numerical instabilities associated with the estimation of Negative Binomial parameters, we adopted a model selection approach to determine which model was best suited to handle the over-dispersion parameter ϕ. Our strategy was first to estimate a model without over-dispersion using the GLMNET package (model $\mathcal{M}(\phi = 0)$). We then considered two different estimation methods for the parameter ϕ, namely a global estimation approach using the package AOD (model $\mathcal{M}(\phi = \phi_{\text{global}})$), and a SNP-specific parameter using the DSS package (model $\mathcal{M}(\phi = \phi^i_{\text{DSS}})$. We used a BIC to choose the best model out of the three. Before comparing the allele read counts from different libraries, the count data were normalised by library sizes as proposed in the DESEQ package (23). This software has been shown to be the most efficient according to a recent normalisation comparison study (24). Pseudo-counts (*i.e.*, systematic random allocation of ones) were considered for SNPs showing many zeros to avoid singular hessian matrices while fitting the generalised linear model. Some events were then filtered out based on their counts: if global counts (for all replicates and all conditions) for both variants were too low (less than 10 counts), we considered that we did not have enough power to conclude on this event and we did not test it.

We then performed the core test on the association between variant and condition. The target hypothesis was H_0: $\{(\alpha\beta)_{ij} = 0\}$, i.e. no interaction between the allele and the condition. If this interaction term is not null, a differential usage of an allele across conditions occurred. The test was performed using a Likelihood Ratio Test with one degree of freedom, which corresponds to the supplementary interaction parameter that is included in the second model and not in the first (25). To account for multiple testing, p-values were adjusted with a 5% false discovery rate (FDR) following a Benjamini–Hochberg procedure (26).

Quantifying the magnitude of the effect. When a variant is found to be differentially represented in two populations, one remaining difficulty is to quantify the magnitude of this effect. Indeed, significant ($P < 0.05$) but weak effects are often detected, especially in RNA-seq data in which some genes are very highly expressed (and hence have very high read counts).

A natural measure for quantifying the magnitude of the effect would be the difference of allele frequencies between the two conditions. In practice, the true difference of allele frequencies is not known, and we estimated it using the RNA-seq counts. The precision of this estimation is discussed in the Results Section.

We denote by f_e the estimation of the allele frequency based on RNA-seq counts:

$$f_e = \frac{\#counts_variant_1}{\#counts_variant_1 + \#counts_variant_2}.$$

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on September 12, 2016

The value of f_e was computed for each replicate of each condition. We then took the mean of these values for all replicates within each condition. Finally, we calculated the difference across conditions and obtained the magnitude of the effect: $Df_e = f_{e_{cond1}} - f_{e_{cond2}}$. In the special case where the two variants had low counts (less than 10) within one replicate, then f_e was not calculated. Finally, if at least half of the replicates of one condition had low counts, Df_e was not computed either. Overall, this prevented from over-interpreting large magnitudes obtained from low counts.

Our method is embedded and distributed in an R package, called KISSDE, which can take as input either the output file of KISSPLICE or any count matrix with two lines representing an event.

Methodology for testing and validating our approach

We first evaluated our method in human, because it is a species for which a reference genome is available and SNPs are well annotated. We then used our method on a non-model species: *Asobara tabida*, an hymenoptera that exhibits contrasted phenotypes and for which no reference genome is available. Finally, we applied our method on a different evolutionary timescale, working on two recently diverged *Drosophila* species, *D. mojavensis* and *D. arizonae*, for which a draft reference genome is available only for *D. mojavensis*.

The Geuvadis dataset. Our method enables to find SNPs from RNA-seq data. In order to assess if the SNPs we find are correct, and if the list we output is exhaustive, we chose to test our method on RNA-seq data from the Geuvadis project. Indeed, the individuals whose transcriptome was sequenced in this project were already included in the 1000 genome project. Hence, their SNPs have already been well annotated. We downloaded fastq files from SRA (see Data access) and selected 10 Toscans and 10 Central Europeans. We sampled 10M reads for each individual and concatenated the fastq files in pools of five individuals.

Definition of the set of true SNPs and their genotypes. We downloaded the vcf file from the 1000G webpage. For each SNP called in the 1000 Genomes project, we had at our disposal the genotype of each individual. We focused on the genotypes of the 20 individuals selected for our analysis. Whenever only one allele was represented in the 20 individuals, we filtered out this SNP, as it simply cannot be discovered based on these 20 individuals only.

Whenever one SNP was covered by less than 5 reads out of the total number of reads in the 20 individuals, we considered that the SNP was located in a too poorly expressed region and could not be discovered by RNA-seq. Other levels of poorly/medium/highly expressed regions are discussed in the Results section. The read coverage was computed using SAMTOOLS depth, on the .sam file obtained after mapping the reads with STAR (v2.3.0) (27).

Calling SNPs from reads mapped to a reference genome: GATK-GENOME. In order to clarify if the performances of our method were on par with other methods, we chose to benchmark against GATK, which is the most widely used method for variant calling in eukaryote samples when a reference genome is available.

We employed the GATK Best Practices workflow for SNP and indel calling on RNA-seq data (https://www.broadinstitute.orggatkguidearticle?id=3891 posted on 6 March 2014, last updated on 31 October 2014) which considers the following steps: (i) mapping to the reference genome with the STAR aligner, 2-pass method (28) with the suggested parameters allowing to obtain the best sensitivity for the variant call task, where during the second pass of STAR a new reference index is created from the splice junction information determined during the first step alignment and a new alignment step is done with the new index reference; (ii) adding read group information, sorting, marking duplicates and indexing, using Picard's tools; (iii) splitting reads into exon segments (removing Ns but maintaining grouping information) and hardclipping sequences overhanging into the intronic regions, using the SplitNCigarReads GATK tool; (iv) realigning indels and recalibrating Base quality; (v) calling variant with HAPLOTYPECALLER, and finally filterimg the variants with VARIANTFILTRATION.

Calling SNPs from reads mapped to a reference transcriptome MPILEUP-TRANSCRIPTOME. The reference transcriptome was assembled using TRINITY (as described previously) and reads were mapped to this reference using BOWTIE2 (29). We then used MPILEUP and BCFTOOLS (30) to call SNPs from the mapped reads. TRINITY, BOWTIE2 and MPILEUP were used with default parameters. BCFTOOLS was used with the options –multiallelic-caller and –variants-only.

As outlined in the Results section, this pipeline performs poorly in the context of alternative splicing, as it misses most of the SNPs located in exons shared by several transcripts.

A way to deal with this issue is to filter the redundancy caused by alternative splicing. The first approach we considered was described in Pante *et al.* (31) and consists in applying CD-HIT (7), a widely used greedy clustering method, to the transcriptome assembled by TRINITY. The second approach we considered was described in Van Belleghem *et al.*, 2012 (6) and consists in keeping only the longest isoform for each gene assembled by TRINITY.

In both cases, we obtained a filtered transcriptome, with reduced redundancy, and we then used BOWTIE2, MPILEUP and BCFTOOLS to call SNPs.

Comparison of genome-based and transcriptome-based approaches. In order to compare the SNPs predicted by KISSPLICE with our set of true SNPs, we needed to obtain a genomic position for each of our predictions. To this purpose, we aligned each variant of each bubble to the reference genome using STAR (v2.3.0). In the case where a variant mapped to several locations, we used the default behaviour of STAR, which is to assign the variant to the location with the fewer number of mismatches. In case of ties, we kept all equally good locations, and if at least one of the possible locations corresponded to an annotated SNP, we considered that the prediction of KISSPLICE was correct.

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on September 12, 2016

For MPILEUP, we aligned the transcripts assembled by TRINITY on the reference genome with BLAT.

Asobara tabida *lines, RNA sequencing and SNP verification.* *Asobara tabida* (Hymenoptera: Braconidae) is a parasitoid species which develops on *Drosophila* hosts. *A. tabida* is naturally infected by three strains of *Wolbachia*, among which one (*w*Atab3) is necessary for oogenesis completion (32,33). However, when *Wolbachia* are removed by antibiotic treatment, the degree of oogenetic defect exhibits genetic variation within populations (34). We thus founded two lineages of *A. tabida* from a natural population (Sainte Foy-les-Lyon, France) based on their extreme phenotype after elimination of *Wolbachia*: the SFR2 lineage whose females do not produce any eggs and the SFR3 lineage whose females produce half the normal content of eggs. In both cases, dependence is complete as the eggs produced are sterile. These two lineages were founded by three females and were kept for 15 generations (three founders at each generation) before RNA extraction.

The experimental design for RNA-seq sequencing aimed at describing the transcriptomic changes associated with the presence / absence of *Wolbachia*, and the variations observed in the two *A. tabida* lineages exhibiting an extreme phenotype. To this purpose, cDNA libraries were constructed from infected and non-infected ovaries in these two lineages. Because these RNA-seq data were issued from two distinct lineages from a non-model species, we exploited this dataset to validate the method developed here and to discover biologically relevant SNPs, using libraries obtained from infected ovaries. The samples used for RNA extraction were young female (0–1 day old) ovaries dissected in a drop of A-buffer (two replicates of 30 ovaries per lineage). RNA was extracted as described in Kremer *et al.* (35). These RNA extracts were used to generate corresponding cDNA libraries, following the recommendations given by the manufacturer of the SMARTer PCR cDNA synthesis and BD Advantage two PCR kits (Clontech). These cDNA libraries were then purified with the Qiaquick kit (Qiagen) and their quality checked. Sequencing of cDNA was performed by the Genoscope (Evry), on an Illumina GA-IIx instrument, to obtain 1x75bp reads. These data were trimmed using the ShortRead package with default parameters and then used as input of the pipeline defined in Figure 1.

Based on these results, 34 SNPS were chosen for verification. For each SNP, primers were designed on the corresponding transcript to amplify the surrounding genomic region. PCRs were performed from an aliquot of the purified cDNA libraries. The reaction was performed in a total volume of 25 µl, and the mixture consisted in 2.5 µl of 5× green DreamTaq mastermix, 200 nM of dNTP, forward and reverse primers (see Supplementary Table S1 for primer sequences), and 5U of DreamTaq DNA polymerase (ThermoFisher). PCR amplification was performed on a Tetrad thermocycler (Biorad) as follows: 2 min at 94°C, 35 times (30 s at 94°C, 30 s at 58°C, 30s at 72°C), and 10 min at 72°C. The PCR products were sequenced using the Sanger method from forward and reverse primers by the Biofidal company. The sequences were aligned and their respective chromatograms analysed by the CLC Main workbench.

Drosophila *strains, RNA sequencing and SNP verification.* *D. mojavensis* and *D. arizonae* are two Drosophila species that are endemic of the arid southwestern United States and Mexico. These species diverged recently (less than 1 MYA) (36,37). In the laboratory, hybridisation of these two species is possible while in nature it does not occur (or is very rare). The ovarian transcriptome of these two species (and their reciprocal crosses) was sequenced to investigate the first step of hybrid incompatibility and look for deregulated genes in hybrids. In this paper, we did not study the transcriptomes of the hybrids, we only used the transcriptomes of the parents to test for the validity of our pipeline at a different evolutionary scale. The sequenced strains were *Drosophila mojavensis* from the Anza Borrego Desert, CA (stock number: 15081–1352.01) and *Drosophila arizonae*, from Metztitlan – Hidalgo, Mexico (stock number: 15081–1271.17), both obtained in the US San Diego Drosophila Stock Center. Virgin female flies were collected after hatching and isolated until they reached ten days. The RNA was extracted from a pool of 30 ovaries of 10-days-old flies for each line. The extractions were performed using the RNeasy kit (Qiagen) and samples were then treated with DNase (DNA-free Kit, Ambion) and stored at −80°C. The samples were quantified by fluorescence in the Bioanalyser 2100 (Agilent), according to pre-established criteria by the sequencing platform. For each line, the extracted RNA was divided into two parts in order to generate two cDNA libraries (two replicates per condition). RNA was sequenced by Illumina Technology, in the IlluminaHiseq 2000. We sequenced 2 × 51 bp paired-end reads and the medium size of the inserts was 300 bp. We used URQT (38) with the default parameters to remove the low quality bases and the polyA tail from the dataset before running the pipeline described in Figure 1. The protocol for SNP verification is identical to the one used for *Asobara tabida* (see Supplementary Table S2 for primer sequences).

Data access

The human data used in this study can be found through the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress/) under the accession number E-GEOD-29342 and we used the individuals named NA20808, NA20809, NA20810, NA20811, NA20812, NA20813, NA20814, NA20815, NA20819, NA20826, NA06984, NA11840, NA06986, NA06989, NA06994, NA07346, NA07357, NA10851, NA11829 and NA11832.

The RNAseq libraries from *D. mojavensis* and *D. arizonae* are available through the NCBI Sequence Read Archive (SRA : http://www.ncbi.nlm.nih.gov/sra) under the accession no. SRX1272419 and SRX1277353.

The *A. tabida* dataset is available through the NCBI Sequence Read Archive (SRA : http://www.ncbi.nlm.nih.gov/sra) under the accession no. SRX1701817, SRX1701824, SRX1701826 and SRX1701855.

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on September 12, 2016

RESULTS

Validation of the SNP calling method using available data from a model species

Identification of variants. In order to evaluate the performance of our method, we needed to test it in the case where we knew which SNPs should be found. We thus focused on a dataset from human in which SNPs were already annotated. We selected two populations (Toscans and Central Europeans) from the Geuvadis project (39), and downloaded the RNA-seq data of 10 individuals in each population. We sampled 10M reads from each individual and pooled individuals 5 × 5, to obtain two replicates of five pooled individuals per population. We ran KISSPLICE and TRINITY on these four read sets and we aligned the variants of KISSPLICE to the TRINITY transcripts using BLAT (with at least 90% query coverage and 90% identity). Out of the 64824 bubbles initially found by KISSPLICE, 53494 (82%) mapped to TRINITY-assembled transcripts, 8024 partially aligned, and 3306 did not align. As explained in the Methods Section, SNPs located near other SNPs may be enumerated more than once, but with different contexts (see Supplementary Figure S2). After removing this redundancy, we ended up with 51,235 bubbles.

To assess whether these bubbles were true SNPs, we first aligned the sequences of the variants (i.e. each path of the bubble) to the human reference genome and compared their genomic positions to a set of SNPs downloaded from the 1000 genome project webpage. We also benchmarked our method against two software: GATK, a widely used method to call SNPs in the presence of a reference genome and MPILEUP, part of the SAMTOOLS/BCFTOOLS, used here to call SNPs on the transcriptome assembled by TRINITY using the same RNAseq data.

GATK was run with parameters recommended from the GATK web page for RNA-seq data. MPILEUP was run on top of BOWTIE2, both on the transcriptome assembled by TRINITY (MP-TRANSCRIPTOME), and on the reduced transcriptome. In the latter case, we either kept the longest isoform for each gene (MP-LONG-TRANS) as described in Van Belleghem *et al.* (6), or we applied CD-HIT to cluster similar isoforms (MP-CD-HIT) as described in Pante *et al.* (31).

For each method, we calculated the Precision, i.e. the number of true SNPs out of the total number of predicted SNPs, and the Recall, i.e. the number of predicted SNPs out of the total number of true SNPs.

As outlined in Figure 5, the recall of all methods is extremely low if no filter is applied to the set of true SNPs (True SNPs minimum coverage set to 0). This is an expected result, because true SNPs were identified using DNA-seq data and recovering them using RNA-seq data requires that they are located in sufficiently expressed regions. The higher the expression, the higher the recall of all methods. For SNPs located in regions covered by at least 100 reads, the best recall is reached for GATK-GENOME (42%), which is better than KISSPLICE (35%) and MP-TRANSCRIPTOME (28%). The low recall of MP-TRANSCRIPTOME is essentially due to its poor ability to find SNPs in constitutive exons, a limitation which can be adressed using MP-LONG-TRANS (but not MP-CD-HIT). The recall of KISSPLICE can also be improved by modifying its relative threshold parameter from 5% to 2%. Interestingly, it even slightly outperforms GATK-GENOME. The reason is that KISSPLICE finds more SNPs located in repeated regions of the genome, while GATK filters them out based on their low mapping quality. Finally, we show that a large number of SNPs are still not found by any method. The majority of those are rare alleles (Supplementary Figure S4) and the remaining are SNPs located in repeated regions or very polymorphic genes, like immune genes.

As outlined in Figure 5, with the exception of KISSPLICE, the precision of all methods was very poor if no filter was applied on the number of reads supporting each prediction. This is an interesting advantage of KISSPLICE. Its predictions can be taken as is, and the precision will already be 80%. If we now focus on predicted SNPs supported by at least 100 reads, then GATK-GENOME was the best and reaches a precision of almost 90%, while MP-TRANSCRIPTOME was the worst with a precision of 70%.

The false positives of all methods can essentially be divided into two categories: sequencing errors, and inexact repeats. The impact of RNA editing was minor (less than 5% of cases were annotated in RADAR v2 (40)).

Filtering out SNPs supported by few reads effectively deals with the issue of sequencing errors, but this consequently affects the ability to find true SNPs in poorly expressed regions.

The issue of inexact repeats affects mostly transcriptome-based methods, not genome-based methods. While KISSPLICE partially deals with this issue with the branching parameter and the filtering of long bubbles, MP-TRANSCRIPTOME does not address this problem.

Overall, we conclude that, although we do not use a reference genome, the recall and precision of our method are comparable to those which use one, such as GATK. Furthermore, we show that our method has a better ability to call SNPs in the context of alternative splicing and a more efficient way to filter out inexact repeats than methods which call SNPs after mapping reads to an assembled transcriptome.

Quantification of variants and statistical differential analysis. The quantification we obtain for variants called from pooled RNA-seq data reflects both the allele frequency of the variant in the pool and the expression level of the gene. An 'expressed' allele frequency can be derived from these counts, by simply taking the ratio, but the obtained frequency is expected to be distorted compared to the allele frequency estimated from DNA-seq data. Several causes may be listed. First, within a heterozygous individual, one allele may be more expressed than the other, a process known as Allele Specific Expression (ASE). Second, RNA expression from different individuals (hence possibly different genotypes) can be variable within a pool, thus distorting the allelotype. In order to evaluate the magnitude of this distortion, we computed within each pool the correlations between the true allelic frequencies, and the estimated allele frequencies. To obtain the true allelic frequency within a pool, we took advantage of the availability of the genotypes of each individual from the Geuvadis dataset, and we simply summed up the number of alternative alleles over the

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on September 12, 2016

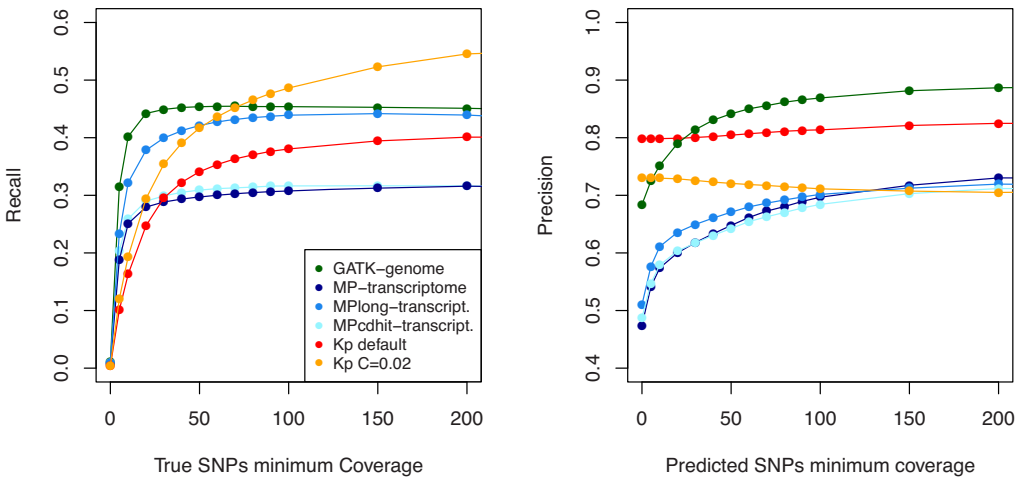


Figure 5. Precision and recall of KISSPLICE, GATK-GENOME, MP-transcriptome and MP-LONG-TRANS as a function of the expression level of the locus. For the recall, all predictions are taken into account, but the set of true SNPs is restricted to those covered by at least a given number of reads. For the precision, only SNPs supported by at least a given number of reads are taken into account.

total number of alleles within the pool. The expressed allele frequencies were obtained from KISSPLICE calls, summing the alternative allele counts of each individual over all allele counts of the pool.

We found that the distortion highly depends on the expression levels (Supplementary Figure S5). While the correlation was weak (0.65) for poorly expressed loci (less than 3 reads), it increased steadily with the expression level up to a plateau of 0.98. When we restricted to loci with at least 10 reads, the correlation reached 0.95.

We therefore conclude that, whenever a locus was sufficiently expressed (at least 10 reads), the expressed allele frequency was a good predictor of the true allele frequency.

If we now compute the difference of allele frequencies across conditions (denoted by *df*), and compare it to the difference of expressed allele frequencies across conditions (denoted by *dfe*), the correlations remain high, but are weaker, reaching a plateau of 80% for highly expressed loci. The reason is that most SNPs do not have a significant difference of allele frequencies across our two populations, hence these correlations are contaminated by SNPs with (almost) equal allele frequencies. In this case, the difference of allele frequencies is just a random fluctuation. When considering all SNPs, the correlation between *df* and *dfe* is significant but weak (Figure 6-A)

If we restrict to SNPs that are found as condition specific by KISSDE, then the correlation is much stronger (Figure 6B). Finally, if we restrict to SNPs covered by a total of at least 100 reads (an average of 25 reads per sample), then the correlation is again higher (Figure 6C). The more a gene is expressed, the higher the fit between *df* and *dfe*. A few SNPs ($n = 22$), however, exhibited a large difference between *df* and *dfe* (>0.3). A detailed analysis of these cases reveals that they are located in immune genes ($n = 5$), in genes showing a very variable expression across individuals ($n = 9$), or in genes exhibiting an allele specific expression ($n = 8$).

Overall, we conclude that, provided we restrict to condition specific SNPs, the metric we output with KISSDE for the difference of expressed allele frequencies, that is *dfe*, can largely be interpreted as a measure of the true difference of allele frequencies.

Prediction of the amino acid change. When no reference genome is available, it is not possible to obtain a genomic location for each SNP and therefore to apply SNPEFF (41), or POLYPHEN (42), which are widely used software for assessing the impact of a SNP on the protein sequence. In the absence of any reference genome, a reference transcriptome can nevertheless be obtained, using a full-length transcriptome assembler like TRINITY (15). Based on this transcriptome, it is possible to assess the coding potential of each transcript using TRANSDECODER, to position the predicted SNPs onto the assembled transcripts using BLAT (20), and finally to assess the impact of each SNP on its transcript(s). In the end, each positioned SNP is classified as coding or non coding. In the case where the SNP is located in the coding region, it is then classified as synonymous or non-synonymous (See Methods).

Out of 47,243 positioned SNPs (those which aligned to TRINITY transcripts), 14,804 cases (31%) fell in CDSs and the other 32,439% fell in non-coding regions (including UTRs). Among the ones falling in CDSs, we found that 53% (7788) were synonymous, while the other half (7016) were non synonymous.

To validate our predictions, we then intersected the genomic positions of our predicted SNPs with the genomic positions of SNPs in dbSNP, for which the impact on the protein sequence is known. Out of the 47,243 SNPs we predict, 39313 could be assigned a genomic position which matched a SNP annotated in dbSNP. Out of those 39313 cases, 2725 have no functional annotation in dbSNP, 35,141 had a correct prediction and 1447 cases wrongly predicted. A thorough examination of the 1447 cases wrongly predicted reveals that in most cases, the transcript predicted by TRINITY was very partial and was overlapping an intron (this happens when pre-mRNA is sampled together with mRNA at the RNA extraction step, despite selection

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on September 12, 2016

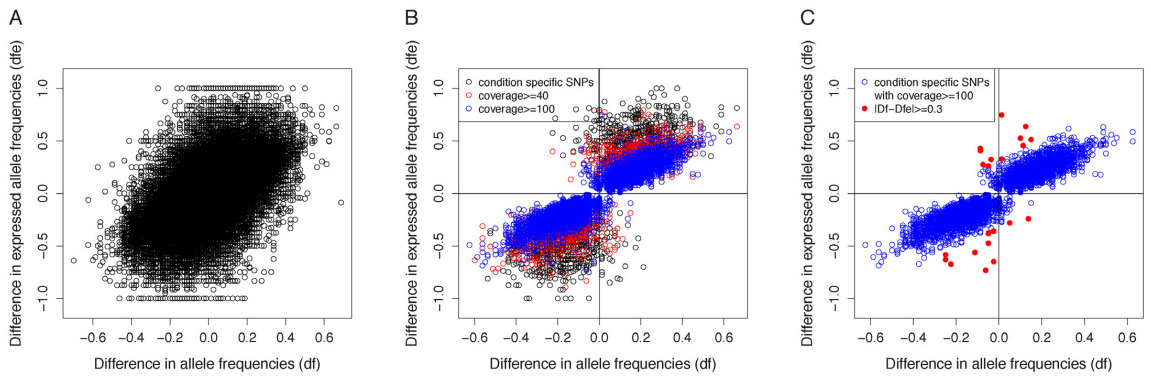


Figure 6. Difference of allele frequencies (df) Vs Difference of expressed allele frequencies (dfe). (**A**) All SNPs. (**B**) Condition-specific SNPs. (**C**) Conditions-specific SNPs covered by at least 100 reads.

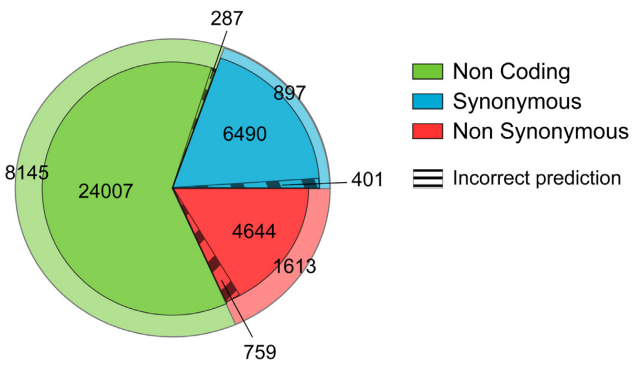


Figure 7. Results of KISSPLICE2REFTRANSCRIPTOME The green, red and blue areas correspond respectively to non-coding, synonymous and non-synonymous SNPs. The dashed area corresponds to errors of our predictions of the impact on the protein sequence. The outer area corresponds to SNPs that are not in dbSNP or for which the prediction cannot be evaluated due to a lack of annotation in dbSNP.

of polyA+RNAs). In this case, the ORF predictor can over-predict coding regions, and our pipeline therefore tends to over-predict non synonymous cases. Figure 7 summarises our results for the prediction of the impact on the protein sequence. Overall, when SNPs can be evaluated, the precision of K2RT is 96% (35,141 out of 36,588).

Performance of the full pipeline. In the previous section, we evaluated our capacity to predict the impact on the protein independently of the remaining of our pipeline. We now turn to its evaluation within the full pipeline. Two situations can be discussed here. First, if only one experimental condition is considered, then no differential analysis is carried out. SNPs are identified and their impact on the protein is predicted. In this case, the prediction inherits from the errors made at the identification step. Out of 47,243 predicted SNPs, 39313 were in dbSNP and 35,141 had a correctly predicted impact. In the worst-case scenario, if we consider that the 7930 SNPs for which there was no dbSNP entry and the 2715 SNPs for which the dbSNP entry is incomplete were false positives, the precision of the pipeline was 74%. In practice, dbSNP is not exhaustive, and the true precision is between 74% and 96%. Second, if two conditions were considered (which is the original purpose of this study), then many of the false positives of the identification step were filtered out. Out of the 47,243 predicted SNPs, 5518 were condition-specific, and 5309 had a correct prediction of the impact on the protein sequence. Hence the precision, in the worst-case scenario, for condition-specific SNPs was 96% (5309 out of 5518).

Application of the method using biological data from species without any reference genome

From our study on the human dataset, we conclude that our method has a precision and recall similar to methods which require a reference genome. We now turn to the application of our method to non-model species.

Application to intraspecific polymorphism: the case of Asobara tabida. We first applied our method to *Asobara tabida*, for which RNA-seq data from two lineages (SFR2 and SFR3) were available. These lineages come from the same population, but they differ by their phenotype of dependence to their symbiont *Wolbachia*. In the absence of *Wolbachia*, SFR2 individuals produce no eggs, while SFR3 produce some. Consequently, we suspect a low but significant genetic differentiation between lineages that could be associated with the phenotypes, or to genetic drift associated with maintenance in the laboratory. While the experimental design, with a single lineage for each phenotype, does not enable us to separate between these two effects, we think that this dataset is still well tailored for a validation of our method because: (a) no reference genome is available for this species; (b) individuals were pooled for RNA extraction and (c) replicates are available for each lineage.

The transcriptomes of two replicates of pools of 30 individuals were sequenced through RNA-seq for each lineage, leading to 15M reads for each replicate. We ran our pipeline and found a total of 18609 positioned SNPs out of which 17,031 are condition-specific. The large proportion of condition-specific SNPs is largely due to the fact that most of them are fixed in at least one lineage. Indeed, 21% of them are fixed in both lineages, 63% are fixed in one lineage and polymorphic in the other, and 7% are polymorphic in both lineages (Supplementary Figure S6B).

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on September 12, 2016

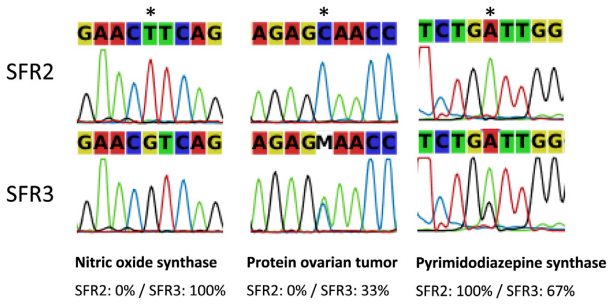


Figure 8. Three examples of SNPs validated by Sanger sequencing. The first is fixed in both the SFR2 and SFR3 lineages. The second and third are polymorphic in SFR3 but fixed in SFR2. In the third case, the base caller does not reflect the polymorphism but it can be seen from the chromatogram

Out of the 17,031 condition-specific variants, we found that 5608 (32%) were non coding, 6137 (36%) were synonymous and 3876 (22%) were non-synonymous.

Based on these results, we selected 27 cases for experimental validation: 10 were cases where the two lineages were fixed for a different nucleotide, 15 were cases where one lineage was fixed and the other polymorphic, 2 were cases where the two lineages were polymorphic. For all the 10 first cases, we were able to validate that the SNP was real and that the two lineages were indeed fixed for a different nucleotide (Supplementary Table S1, Figure 8). Out of the 17 remaining cases, we were able to validate that the SNP was real in all cases, but only in 9 cases were we able to validate that the site was polymorphic in one lineage (Supplementary Table S1, Figure 8). The rate of validation of the polymorphic status of the site within a lineage largely depended on the frequency of the minor allele (Supplementary Figure S5). Rare variants were harder to validate in terms of polymorphism detection. These rare variants could be false positives of our method, but they may also very well be true variants, not detectable experimentally using a direct sequencing technique without cloning. Importantly, although we could not always validate the fact that one site is polymorphic within a lineage, we systematically confirmed that the SNP was real, and that each lineage had a specific major allele. Therefore, we validated the condition-specificity of all SNPs.

As discussed earlier, our method outputs SNPs that are found by no other method. In order to test if these SNPs were true, we further tested specifically 7 such cases, and were able to validate all seven SNPs (Supplementary Table S1).

Because our RNA-seq data were initially obtained to compare the transcriptome of these two lineages, the design was not optimized for QTL analysis. In particular, each phenotype is represented by a single inbred genotype, making it difficult to separate the SNPs linked to the phenotype from those linked to drift. Despite this issue, we further characterised the impact on the protein sequence of the condition-specific SNPs. Among all these genes, some called our attention regarding their possible implication in the dependence phenotype. For instance, some genes, such as *Dorsal* and *Hypoxia up-regulated protein 1*, presented SNPs in their UTRs and were differentially expressed between lineages. These genes are involved in immunity and oxidative stress homeostasis, two functions that have been shown as particularly important in this biological system. Another example concerns genes involved in oogenesis, like *OTU-domain containing protein* or *Female sterile*, that exhibit non-synonymous SNPs in their CDS regions. These few examples show how the suite we propose in this paper rapidly allows to link the SNPs detected to their impact on the protein sequence, thus permitting to pinpoint candidate genes involved in phenotypic variation. Validation of these genes could involve either genetic studies (*e.g.*, knockdown experiments) and/or other linkage analyses targeted to these candidates.

Application to Interspecific Divergence: the case of Drosophila mojavensis *and* Drosophila arizonae. Similarly to the *Asobara* dataset, the drosophila dataset corresponds to non-model species, where individuals had to be pooled prior to RNA sequencing. In this case however, the two modalities of the phenotypes are not two populations of the same species, but two recently diverged species. This therefore enabled us to assess if our method also applies to a very different evolutionary scale, where differences of one nucleotide are no longer SNPs, but divergences. Additionally, the availability of the reference genome for *D. mojavensis* (and not *D. arizonae*) enabled us to study in depth the case of condition-specific inexact repeats.

D. mojavensis and *D. arizonae* are two closely related species that diverged 1MYA. We sequenced through RNA-seq the ovarian transcriptomes of two replicates of pools of 30 individuals for each species. We obtained 55M paired-end reads per replicate. We ran our pipeline on the data and obtained 51,730 positioned SNPs, and most of them (51,135) were condition-specific.

The condition-specific SNPs were mostly in coding regions (60%, i.e. 40,674 SNPs). We could classify 34,382 of them as synonymous, and the other 6292 SNPs as non-synonymous.

We selected 11 cases for experimental validation, six of which were divergent sites, and five were cases where the site was polymorphic in one species and fixed in the other. We were able to validate that the variation was condition-specific for all the divergent sites, and for four cases out of five for the polymorphic cases. Additionally, for two cases out of these four, we were able to amplify the two alleles in the species where the site was predicted to be polymorphic (Supplementary Table S2).

In most cases, an observed variation in the transcriptome is caused by the presence of two alleles at one locus. However, it is also possible that two mono-allelic loci, if they exhibit the same sequence except for one nt, generate a variation that resembles a SNP. In order to quantify this phenomenon, we explicitely selected in the results of KISSPLICE, the variations for which one path was mapping to one locus and the other path was mapping to another locus. This was only possible because we had at our disposal a draft genome of *D. mojavensis*. We selected explicitely cases where we knew that the variation we detected was potentially caused by two loci. There were only 224 cases like this,

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on September 12, 2016

which is very few compared to the total number of variations detected. We however tested three of them experimentally, and we were able to validate all of them. These cases are not true SNPs, but they correspond to recent paralog genes where one copy is more expressed in *D. arizonae*, and the other copy is more expressed in *D. mojavensis*.

CONCLUSION AND PERSPECTIVES

We present a method that can discover condition-specific SNPs from raw RNA-seq data. The individuals may be pooled, which decreases the costs of library preparation, while still enabling to allelotype and to find variants specific to one condition. As no reference genome is required, the range of applications of the method is very large. We first evaluated our method in human, where a reference genome is available and SNPs are extensively annotated. We show that our method has similar performances in terms of precision and recall, compared to GATK, a widely used mapping-based approach. We then evaluated our method on two non-model species.

In both cases, we were able to call variants, to classify them, and to discuss their impact. We selected a fraction of them for experimental validation through RT-PCR + Sanger sequencing. In all cases, we were able to validate that the variant was condition-specific. However, when the locus was predicted to be polymorphic in one condition, we were able to validate the presence of the two alleles only in cases where the minor allele frequency was at least 15%.

This work is a first approach toward transcriptome-wide association studies in non-model species. The method can readily be applied to RNA-seq data from any species, whenever two phenotypes are clearly identified and the goal is to find candidates for their genetic bases. In the case of continuous phenotypes, like height, the statistical framework can be generalised to quantitative trait loci (QTL).

This work focuses on SNP identification and analysis and does not address the question of the experimental design of a transcriptome-wide association study. A systematic evaluation of the optimal design is beyond the scope of this paper, but we would like to provide here briefly some basic advice.

First, in all the case studies presented here, we considered only two replicates, which is the minimum required by our method. We clearly advise that for a pre-determined cost, it is wiser to have a low coverage for each replicate, but to increase the number of replicates. Second, the type of replicates to choose is probably a more central issue. In the case of *Asobara*, we sequenced two biological replicates, but both replicates were derived from the same lineage. Having replicates when extracting RNA is useful, but not as useful as replicates at the line-establishment step. Only this type of replicate can allow to discriminate between SNPs in the original population and genetic drift in the lab. Finally, if pooling is envisioned, the number of individuals per pool should be as large as possible, especially for very polymorphic species. The larger the pool, the more representative of the population it is.

From the point of view of our method itself, there is of course also room for improvement. In particular, we found that, while easy SNPs are identified by all methods, a large amount of difficult SNPs are currently being overseen. This is the case of SNPs located in repeated regions of the genome, and that are notoriously difficult to annotate. SNPs located very close to each other are also challenging to annotate. Without a reference genome, we found that they are particularly difficult to tell apart from inexact repeats. Finally, SNPs located within very polymorphic regions of the genome, like immune genes, are also very challenging, even for mapping-based approaches. The use of a single reference genome is clearly limiting. De novo assembly methods are a promising direction for these, but still need to be optimised.

For future work, we see two lines of research, which could ultimately be combined. First, we could take advantage of the availability of long reads coming from third generation sequencing platforms (Pacbio, Minion). In principle, long reads have the potential to solve most of the issues we mentioned, but currently, the error rates are too high (10–15%) and the sequencing depth is not sufficient to apply to RNA-seq. In the meantime, it seems still relevant to keep on working in the context of short reads, but we think that the best resolution we can achieve for the prediction of difficult SNPs is not well captured by sequences. Graphs could instead well represent close SNPs and a partial quantification of their phasing.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was performed using the computing facilities of the CC LBBE/PRABI.
The authors would like to thank Marie-Christine Carpentier and Delphine Charif for help on the analysis of the *Asobara* dataset; Gustavo Sacomoto for help on the analysis of the Geuvadis dataset; Sebastien Deraison for help on experimental validation of the candidates for the Drosophila dataset, and Vincent Miele and Alice Julien-Laferrière for help on developing KISSPLICE and KISSDE.

FUNDING

Agence Nationale de la Recherche [ANR-12-BS02-0008, ANR-11-BINF-0001-06, ANR-2010-BLAN-170101]; São Paulo Research Foundation – FAPESP/Brazil [2010/10731-4 to C.M..C.]; European Research Council under the European Community's Seventh Framework Programme [FP7 /2007–2013)/ERC Grant Agreement No. [247073]10]. Funding for open access charge: INRIA.
Conflict of interest statement. None declared.

REFERENCES

1. Iqbal,Z., Caccamo,M., Turner,I., Flicek,P. and McVean,G. (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.*, **44**, 226–232.
2. Uricaru,R., Rizk,G., Lacroix,V., Quillery,E., Plantard,O., Chikhi,R., Lemaitre,C. and Peterlongo,P. (2015) Reference-free detection of isolated SNPs. *Nucleic Acids Res.*, **43**, e11.
3. Leggett,R.M., Ramirez-Gonzalez,R.H., Verweij,W., Kawashima,C.G., Iqbal,Z., Jones,J.D.G., Caccamo,M. and Maclean,D. (2013) Identifying and classifying trait linked

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on September 12, 2016

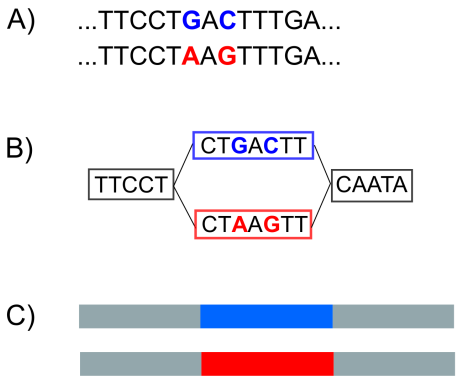
polymorphisms in non-reference species by walking coloured de bruijn graphs. *PLoS One*, **8**, e60058.
4. Piskol,R., Ramaswami,G. and Li,J.B. (2013) Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.*, **93**, 641–651.
5. Romiguier,J., Gayral,P., Ballenghien,M., Bernard,A., Cahais,V., Chenuil,A., Chiari,Y., Dernat,R., Duret,L., Faivre,N. *et al.* (2014) Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, **515**, 261–263.
6. Van Belleghem,S.M., Roelofs,D., Van Houdt,J. and Hendrickx,F. (2012) De novo transcriptome assembly and SNP discovery in the wing polymorphic salt marsh beetle Pogonus chalceus (Coleoptera, Carabidae). *PLoS One*, **7**, e42605.
7. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
8. Schlötterer,C., Tobler,R., Kofler,R. and Nolte,V. (2014) Sequencing pools of individuals – mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.*, **15**, 749–763.
9. Lappalainen,T., Sammeth,M., Friedländer,M.R., 't Hoen,P.A.C., Monlong,J., Rivas,M.A., Gonzàlez-Porta,M., Kurbatova,N., Griebel,T., Ferreira,P.G. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.
10. Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T. and McVean,G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
11. 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.
12. Peterlongo,P., Schnel,N., Pisanti,N., Sagot,M.F. and Lacroix,V. (2010) Identifying SNPs without a reference genome by comparing raw reads. https://hal.inria.fr/inria-00514887/document.
13. 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.
14. Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.
15. 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.
16. Chikhi,R. and Rizk,G. (2013) Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorith. Mol. Biol: AMB*, **8**, 22.
17. Salikhov,K., Sacomoto,G. and Kucherov,G. (2014) Using cascading Bloom filters to improve the memory usage for de Brujin graphs. *Algorit. Mol. Biol.*, **9**, 2.
18. Sacomoto,G., Sinaimeri,B., Marchet,C., Miele,V., Sagot,M.-F. and Lacroix,V. (2014) Navigating in a Sea of Repeats in RNA-seq without Drowning. *Lect. Notes Bioinformatics*, **8701**, 82–96.
19. Tilgner,H., Knowles,D.G., Johnson,R., Davis,C.A., Chakrabortty,S., Djebali,S., Curado,J., Snyder,M., Gingeras,T.R. and Guigó,R. (2012) Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.*, **22**, 1616–1625.
20. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
21. Lu,J., Tomfohr,J.K. and Kepler,T.B. (2005) Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics*, **6**, 1.
22. Robinson,M.D. and Smyth,G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.
23. Anders,S. (2010) Analysing RNA-Seq data with the DESeq package. *Mol. Biol.*, **43**, 1–17.
24. Dillies,M.-A., Rau,A., Aubert,J., Hennequet-Antier,C., Jeanmougin,M., Servant,N., Keime,C., Marot,G., Castel,D., Estelle,J. *et al.* (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinformatics*, **14**, 671–683.
25. Bullard,J.H., Purdom,E., Hansen,K.D. and Dudoit,S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.
26. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)*, 289–300.
27. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
28. Engström,P.G., Steijger,T., Sipos,B., Grant,G.R., Kahles,A., Alioto,T., Behr,J., Bertone,P., Bohnert,R., Campagna,D. *et al.* (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods*, **10**, 1185–1191.
29. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
30. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
31. Pante,E., Rohfritsch,A., Becquet,V., Belkhir,K., Bierne,N. and Garcia,P. (2012) SNP detection from de novo transcriptome sequencing in the bivalve Macoma balthica: marker development for evolutionary studies. *PLoS One*, **7**, e52302.
32. Dedeine,F., Vavre,F., Fleury,F., Loppin,B., Hochberg,M.E. and Bouletreau,M. (2001) Removing symbiotic Wolbachia bacteria specifically inhibits oogenesis in a parasitic wasp. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 6247–6252.
33. Dedeine,F., Vavre,F., Shoemaker,D.D. and Boulétreau,M. (2004) Intra-individual coexistence of a Wolbachia strain required for host oogenesis with two strains inducing cytoplasmic incompatibility in the wasp Asobara tabida. *Evolution*, **58**, 2167–2174.
34. Kremer,N., Dedeine,F., Charif,D., Finet,C., Allemand,R. and Vavre,F. (2010) Do variable compensatory mechanisms explain the polymorphism of the dependence phenotype in the Asobara tabida-wolbachia association? *Evolution*, **64**, 2969–2979.
35. Kremer,N., Voronin,D., Charif,D., Mavingui,P., Mollereau,B. and Vavre,F. (2009) Wolbachia interferes with ferritin expression and iron metabolism in insects. *PLoS Pathog.*, **5**, e1000630.
36. Matzkin,L.M. (2004) Population genetics and geographic variation of alcohol dehydrogenase (Adh) paralogs and glucose-6-phosphate dehydrogenase (G6pd) in Drosophila mojavensis. *Mol. Biol. Evol.*, **21**, 276–285.
37. Reed,L., Nyboer,M. and Markow,T. (2007) Evolutionary relationships of Drosophila mojavensis geographic host races and their sister species Drosophila arizonae. *Mol. Ecol.*, **16**, 1007–1022.
38. Modolo,L. and Lerat,E. (2015) UrQt: an efficient software for the Unsupervised Quality trimming of NGS data. *BMC Bioinformatics*, **16**, 137.
39. Lappalainen,T., Sammeth,M., Friedländer,M.R., 't Hoen,P.A.C., Monlong,J., Rivas,M.A., Gonzàlez-Porta,M., Kurbatova,N., Griebel,T., Ferreira,P.G. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.
40. Ramaswami,G. and Li,J.B. (2014) RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res.*, **42**, D109–D113.
41. Cingolani,P., Platts,A., Wang,L.L., Coon,M., Nguyen,T., Wang,L., Land,S.J., Lu,X. and Ruden,D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, **6**, 80–92.
42. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on September 12, 2016

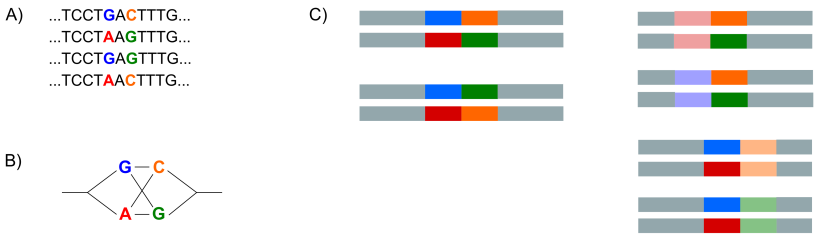
3 Supplementary Information

Supplementary Figures File

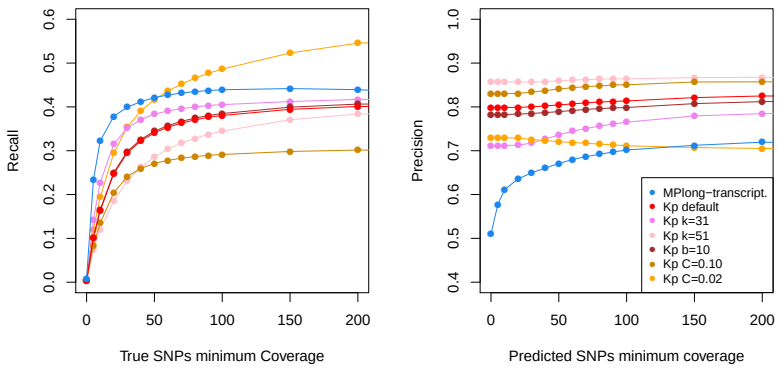
1



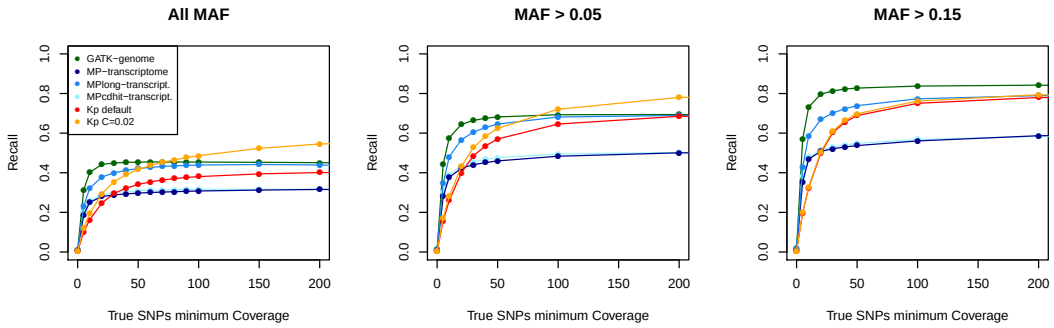
Supplementary Figure 1: Two SNPs separated by less than k nucleotides will be reported in the same bubble. If the SNPs are linked, only one bubble is reported.



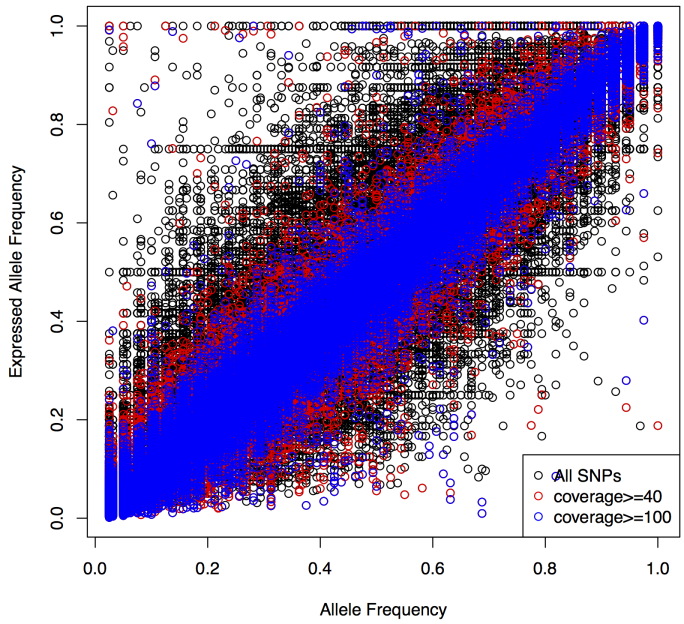
Supplementary Figure 2: Two SNPs separated by less than k nucleotides, but with no linkage, can correspond to 4 haplotypes. They will generate 6 bubbles.



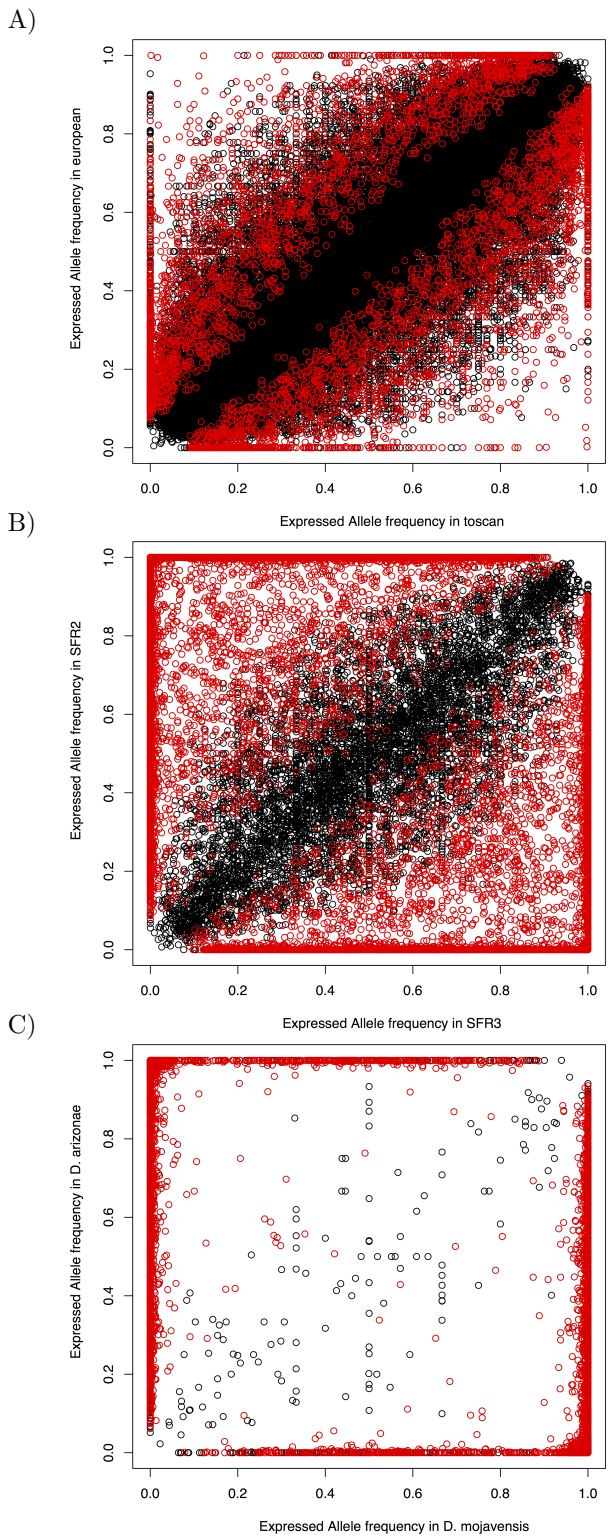
Supplementary Figure 3: Influence of the parameters k, C, and b on the recall and precision of KisSplice. k is the kmer size. C is the relative coverage cutoff. b is the maximum number of branches allowed in a bubble. The default values are k=41, C=0.05 and b=5. Increasing k, increasing C or decreasing b results in a better precision but a worse recall. We also indicate the recall and precision of mp-long. The best recall is reached for C=0.02. The best precision is reached for k=51.



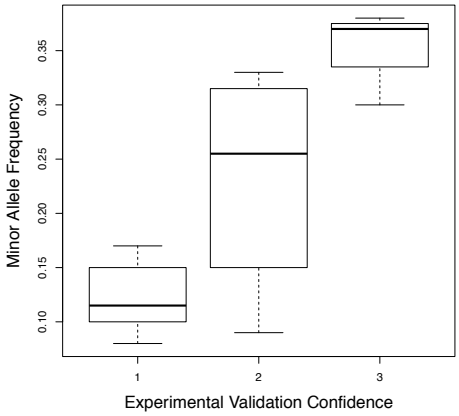
Supplementary Figure 4: Impact of minor allele frequency cut-off on the recall of all methods. The larger the MAF, the easier it is to detect the SNP.



Supplementary Figure 5: Allele frequency estimated using RNA-seq data Vs the true allele frequency. The higher the expression, the higher the correlation.



Supplementary Figure 6: Expressed allele frequencies of one lineage/population Vs expressed allele frequency of the other. Red dots are conditions specific SNPs. Black dots are SNPs whose allele frequency is not different across populations. A) Human TSI Vs CEU B) *Asobara tabida* SFR2 Vs SFR3 C) *Drosophila mojavensis* Vs *Drosophila arizonae*



Supplementary Figure 7: Confidence in the experimental validation depends on the minor allele frequency. A scale ranging from 1 to 3 indicates the confidence degree of the experimental validation process; a number of 3 corresponding to the highest confidence.

Supplementary Table 1: List of SNPs predicted by KisSplice in *Asobara tabida* SFR2 and SFR3 lines. The 27 first cases were chosen for experimental validation because they covered a wide range of MAF and they fell in genes whose function was related to the constrasted phenotypes. The last 7 cases were chosen because they were found by KisSplice only.

Supplementary Table 2: List of divergent sites, SNPs and inexact repeats predicted by KisSplice in *Drosophila mojavensis* and *Drosophila arizonae*. They were chosen for experimental validation because they covered a wide range of MAF and were located in sufficiently expressed loci (at least 100 reads).

Supplementary Table 1

Locus_ID	SNP_ID	Position	CDS/UTR	Codon	S/NS	Allele frequency SFR2
c12624_g2_i1	bcc_11996|Cycle_0	1487	UTR	NA	NA	0,55
c14371_g1_i2	bcc_8887|Cycle_0	3187	CDS	TCG/TTG	NS (S/L)	1,00
c14371_g1_i2	bcc_8886|Cycle_0	3245	CDS	CCC/GCC	NS (P/A)	0,00
c14371_g1_i2	bcc_8885|Cycle_0	3299	CDS	AGT/GGT	NS (S/G)	1,00
c13827_g1_i1	bcc_6853|Cycle_1	645	CDS	TTC/TTG	NS (F/L)	0,08
c13827_g1_i1	bcc_6861|Cycle_0	2853	CDS	GGG/GGT	S	1,00
c13827_g1_i1	bcc_6859|Cycle_0	2538	CDS	AGA/AGG	S	0,99
c13827_g1_i1	bcc_6858|Cycle_0	2236	CDS	AAA/GAA	NS (K/E)	1,00
c13827_g1_i1	bcc_6851|Cycle_0	251	UTR	NA	NA	0,11
c6670_g1_i1	bcc_27099|Cycle_0	878	UTR	GTC/TTC	NA	0,00
c13376_g1_i2	bcc_8960|Cycle_0	2706	CDS	TTG/TCG	NS (L/S)	1,00
c13376_g1_i2	bcc_8956|Cycle_2	651	CDS	GAA/GCA	NS (E/A)	0,00
c11237_g1_i1	bcc_15963|Cycle_0	1937	CDS	GAA/GCA	NS (E/A)	0,00
c11237_g1_i1	bcc_15962|Cycle_0	2004	CDS	TTG/CTG	S	1,00
c12429_g1_i2	bcc_626|Cycle_0	1295	CDS	CCC/CCT	S	0,00
c12429_g1_i2	bcc_623|Cycle_0	762	CDS	GTT/CTT	NS (V/L)	0,00
c12429_g1_i2	bcc_622|Cycle_0	488	UTR	NA	NA	0,00
c12429_g1_i2	bcc_624|Cycle_0	959	CDS	GTA/GTG	S	1,00
c13389_g1_i1	bcc_18926|Cycle_0	2851	UTR	NA	NA	0,09
c13389_g1_i1	bcc_18930|Cycle_0	2084	CDS	CCC/CCA	S	0,17
c13389_g1_i1	bcc_18929|Cycle_0	2267	CDS	CAG/CAA	S	0,21
c13389_g1_i1	bcc_18932|Cycle_0	1385	CDS	GGC/GGT	S	1,00
c13389_g1_i1	bcc_18931|Cycle_0	1591	CDS	CTG/TTG	S	1,00

c13389_g1_i1	bcc_18926|Cycle_6	2890	UTR	NA	NA	0,70
c6099_g1_i1	bcc_612|Cycle_0	421	CDS	GAG/GAC	NS (E/D)	0,37
c6099_g1_i1	bcc_613|Cycle_0	197	CDS	ATT/GTT	NS(I/V)	1,00
c6730_g1_i2	bcc_15558|Cycle_0	841	CDS	ACG/AAG	NS(T/K)	0,19
c14181_g1_i1	bcc_11633|Cycle_0	1921	CDS	GGA/GGC	S	0,00
c14181_g1_i1	bcc_11634|Cycle_0	2065	CDS	CAG/CAA	S	0,00
c14181_g1_i1	bcc_11627|Cylcle_0	3133	CDS	GAA/GAG	S	0,00
c13139_g1_i1	bcc_9194|Cycle_0	406	Non-coding	NA	NA	0,00
c10697_g1_i1	bcc_19641|Cycle_0	1078	CDS	CCA/CCG	S	0,00
c14411_g1_i5	bcc_28533|Cycle_0	2102	CDS	GTT/GTC	S	1,00
c14411_g1_i4	bcc_6950|Cycle_0	1374	CDS	CCC/CAC	NS(P/H)	0,00

Supplementary Table 1

Allele frequency SFR3	Detection of inter-population polymorphism	Detection of intra-population polymorphism	Confidence after sequencing	Primer F (5'-3')	Primer R (5'-3')
0,00	yes	yes	4	CTATACGTCACTAATCTCCCG	TTTATCGCCTCTTGTGCCT
0,00	yes	NA	4	AGAGAAGACAGAGGGCCA	ACCAGGTCCATTCCTCCA
1,00	yes	NA	4	AGAGAAGACAGAGGGCCA	ACCAGGTCCATTCCTCCA
0,00	yes	NA	4	AGAGAAGACAGAGGGCCA	ACCAGGTCCATTCCTCCA
1,00	yes	no	1	ACAAATCGAGCCAAACACA	CAACTCCTCCAATTTTTCCC
0,12	yes	no	1	CAGAAAAGGGCAATGAGAC	CTTGGGTTTTGGGGATTT
0,10	yes	no	1	CAGAAAAGGGCAATGAGAC	CTTGGGTTTTGGGGATTT
0,15	yes	no	1	CAGAAAAGGGCAATGAGAC	CTTGGGTTTTGGGGATTT
0,92	yes	no	1	ACAAATCGAGCCAAACACA	CAACTCCTCCAATTTTTCCC
1,00	yes	NA	4	CCTCCTTGTCCGTCATTT	CATCTCCTCATCTCCACT
0,38	yes	yes	3	GAAAGAAAGAGAACATCAGGG	CACGGATGGAGCAAACAA
0,33	yes	yes	3	TTCGTGATGTTCGATGCTT	GGAGGGAGATCTTTGAGTTG
1,00	yes	NA	4	CTCATTCCTCTCCCTCTC	CAAGCTCACATCCAAATCC
0,00	yes	NA	4	CTCATTCCTCTCCCTCTC	CAAGCTCACATCCAAATCC
1,00	yes	NA	4	AAAACCGAAAGCCTAGCA	CATCTCCACCCACAAGAAAA
1,00	yes	NA	4	AAAACCGAAAGCCTAGCA	CATCTCCACCCACAAGAAAA
1,00	yes	NA	4	AAAACCGAAAGCCTAGCA	CATCTCCACCCACAAGAAAA
0,00	yes	NA	4	AAAACCGAAAGCCTAGCA	CATCTCCACCCACAAGAAAA
1,00	yes	yes	2	ACCACAACTCTCCAGAAA	CGAAAAACCCCGCAAATAA
1,00	yes	no	1	GAGGTTATGGGGATGTGG	GGAGGGCGGATAAATTGG
0,99	yes	yes	2	GAGGTTATGGGGATGTGG	GGAGGGCGGATAAATTGG
0,30	yes	yes	3	AATCCATCATACCGTCCA	CCACCACTATCGATCTCAA
0,37	yes	yes	3	AATCCATCATACCGTCCA	CCACCACTATCGATCTCAA

1,00	yes	yes	2	ACCACAACTCTCCAGAAA	CGAAAAACCCCGCAAATAA
1,00	yes	yes	3	TCAAGCTCCACCTCCTCT	CACCACGGCCAAATCATCA
0,67	yes	yes	2	TCAAGCTCCACCTCCTCT	CACCACGGCCAAATCATCA
0,00	yes	yes	2	AACATGAAGATGCAGAGG	GGAGACGGATAATGAAGAA
1,00	yes	NA	4	TCAAGCTTCCGAAATAATCACA	CCAAAGAACACCCTTCCAGT
1,00	yes	NA	4	TCAAGCTTCCGAAATAATCACA	CCAAAGAACACCCTTCCAGT
1,00	yes	NA	4	TTGATCTGTTGTCGGTTCCA	TTGAGTGACCCATTTCGATG
1,00	yes	NA	4	GGGAGGCGTGATTACAAGAA	GCTTTGCGGGTACGATTTT
1,00	yes	NA	4	CCCTGAGTCTCGGTTACTCG	ATTGCCGAAGTTGTATGGGA
0,00	yes	NA	4	AGCATGGAATACTGGGAGCA	AGTGGAGAGAGGCGAATGG
1,00	yes	NA	4	ACCGGAAGTGGATGTAGACG	CAGAATCGGCCAATAGCAA

Supplementary Table 1

Type	Annotation (Best hit on CDS)
Polymorphic in one lineage	sex-lethal
Fixed in both lineages	piwi-like protein 1 (Argonaute)
Fixed in both lineages	piwi-like protein 1 (Argonaute)
Fixed in both lineages	piwi-like protein 1 (Argonaute)
Polymorphic in one lineage	hypoxia up-regulated protein 1 isoform X1
Polymorphic in one lineage	hypoxia up-regulated protein 1 isoform X1
Polymorphic in one lineage	hypoxia up-regulated protein 1 isoform X1
Polymorphic in one lineage	hypoxia up-regulated protein 1 isoform X1
Polymorphic in one lineage	hypoxia up-regulated protein 1 isoform X1
Fixed in both lineages	nitric oxide synthase
Polymorphic in one lineage	protein ovarian tumor (OTU)-like
Polymorphic in one lineage	protein ovarian tumor (OTU)-like
Fixed in both lineages	OTU domain-containing protein 6B
Fixed in both lineages	OTU domain-containing protein 6B
Fixed in both lineages	peptidoglycan-recognition protein LE
Fixed in both lineages	peptidoglycan-recognition protein LE
Fixed in both lineages	peptidoglycan-recognition protein LE
Fixed in both lineages	peptidoglycan-recognition protein LE
Polymorphic in one lineage	Transcription factor p65/Dorsal
Polymorphic in one lineage	Transcription factor p65/Dorsal
Polymorphic in one lineage	Transcription factor p65/Dorsal
Polymorphic in one lineage	Transcription factor p65/Dorsal
Polymorphic in one lineage	Transcription factor p65/Dorsal

Polymorphic in one lineage	Transcription factor p65/Dorsal
Polymorphic in one lineage	pyrimidodiazepine synthase-like
Polymorphic in one lineage	pyrimidodiazepine synthase-like
Polymorphic in one lineage	caspase 1
Fixed in both lineages	tRNA (adenine(58)-N(1))-methyltransferase non-catalytic subunit TRM6
Fixed in both lineages	tRNA (adenine(58)-N(1))-methyltransferase non-catalytic subunit TRM6
Fixed in both lineages	tRNA (adenine(58)-N(1))-methyltransferase non-catalytic subunit TRM6
Fixed in both lineages	NA
Fixed in both lineages	uncharacterized protein
Fixed in both lineages	uncharacterized protein
Fixed in both lineages	uncharacterized protein

Supplementary Table 1

hit species	e-value	found by Kissplice	found by MP-transcriptome	found by Mplong-Transcriptome
Fopius arisanus	1E-143	yes	yes	yes
Fopius arisanus	0.0	yes	no	yes
Fopius arisanus	0.0	yes	no	yes
Fopius arisanus	0.0	yes	no	yes
Fopius arisanus	0.0	yes	no	yes
Fopius arisanus	0.0	yes	yes	yes
Fopius arisanus	0.0	yes	yes	yes
Fopius arisanus	0.0	yes	yes	yes
Fopius arisanus	0.0	yes	yes	yes
Fopius arisanus	0.0	yes	yes	yes
Athalia rosae	3E-115	yes	no	yes
Athalia rosae	3E-115	yes	no	yes
Fopius arisanus	3E-151	yes	yes	yes
Fopius arisanus	3E-151	yes	yes	yes
Microplitis demolitor	3E-097	yes	yes	yes
Microplitis demolitor	3E-097	yes	yes	yes
Microplitis demolitor	3E-097	yes	yes	yes
Microplitis demolitor	3E-097	yes	yes	yes
Fopius arisanus	0.0	yes	yes	yes
Fopius arisanus	0.0	yes	yes	yes
Fopius arisanus	0.0	yes	yes	yes
Fopius arisanus	0.0	yes	yes	yes
Fopius arisanus	0.0	yes	yes	yes

Fopius arisanus	0.0	yes	yes	yes
Fopius arisanus	1E-154	yes	yes	yes
Fopius arisanus	1E-154	yes	yes	yes
Fopius arisanus	0.0	yes	no	yes
Diachasma alloeum	0.0	yes	no	no
Diachasma alloeum	0.0	yes	no	no
Diachasma alloeum	0.0	yes	no	no
NA	NA	yes	no	no
Diachasma alloeum	1,00E-92	yes	no	no
Diachasma alloeum	0.0	yes	no	no
Diachasma alloeum	0.0	yes	no	no

Supplementary Table 2

Locus_ID	SNP_ID	Position	Codon
c8406_g1_i1	bcc_76854|Cycle_19	830	NA
c4329_g1_i1	bcc_61683|Cycle_0	3762	NA
c8352_g1_i1	bcc_85264|Cycle_0	2210	CTG/CTA
c8033_g1_i1	bcc_80693|Cycle_6	1911	GAT/GAC
c8254_g20_i1	bcc_55710|Cycle_4	1783	TAT/TAC
c2924_g1_i1	bcc_76573|Cycle_0	3345	ACC/CCC
c5390_g1_i1	bcc_33707|Cycle_0	227	GCA/GCG
c8206_g2_i1	bcc_77662|Cycle_0	550	NA
c8218_g35_i1	bcc_67156|Cycle_0	1545	GAC/GAT
c8386_g3_i1	bcc_3630|Cycle_0	4692	NA
c10563_g1_i1	bcc_23843|Cycle_0	536	CCT/CCC
c8308_g30_i1	bcc_3040|Cycle_0	1202	TCG/TCT
c8368_g2_i1	bcc_23212|Cycle_7	2604	GTT/GTC
c8221_g24_i2	bcc_57994|Cycle_51	421	CAA/CAG

Supplemenary Table 2

CDS/UTR	S/NS	Allele frequency Moj	Allele frequency Arz	Detection of divergence
UTR	NA	0	0,76	yes
UTR	NA	1	0,21	yes
CDS	S	0,44	1	no
CDS	S	0,65	0	yes
CDS	S	0,63	0	yes
CDS	NS (T/P)	1	0	yes
CDS	S	1	0	yes
UTR	NA	1	0	yes
CDS	S	1	0	yes
CDS	NA	0	1	yes
CDS	S	0	1	yes
CDS	S	0,35	1	yes
CDS	S	0,5	1	yes
CDS	S	0,5	1	yes

Supplemenary Table 2

Detection of intra-species polymorphism	**Primer F (5'-3')**
yes	TGTTTTGAGCAGAGAGTATGTCG
no	TGAAGACCACTGCGTACTCG
no	GGATGTGGACGAGAAGGAAA
yes	CGCGATAAATTCCAAGAGGA
no	TTGCACCATTGTTGAGTTTCTT
NA	GGAGGTGCCCGTCGAG
NA	GAAACCAAAAGCCACTGAGG
NA	TAGGTGATTGTTGCCTGTGC
NA	ATGCTGATGTGGGCTATGAA
NA	GACAATGGTGCGTTATCTCG
NA	AGCAGCATGACCTTCAAAAA
yes	ATAAAAAGCCCCAACGGACT
yes	CGATCGTCTTGTCACCTTGA
yes	AGTTCGGACGCGTCTACTTG

Supplemenary Table 2

Primer R (5'-3')	**Type**
CTCTCCGGTATGGATGTGGT	Polymorphic in one species
GCTCGATTGTTTGTAATTCTGC	Polymorphic in one species
TAAAGTTAATGCCCGCCTCA	Polymorphic in one species
GAGGCTAGTAAGCGCCTTGA	Polymorphic in one species
AGCAGGAGCAACAGGATCTC	Polymorphic in one species
TCAGCATCCTCAACGTCAT	Divergent
GGCGCCTTCTTTACGTTCTT	Divergent
CTCAGCCCCAGGGTTAGTTC	Divergent
TTATCCCGATTCCACTCCAG	Divergent
TGGTCAGTCCCAGTTCCTTT	Divergent
AGCCGAATCACTTGCTTGTT	Divergent
ACGAGATCATGGTGCCTTTC	Inexact Repeat
GCAGTTATAGGACCCGTTGG	Inexact Repeat
ATGAGCAGACCAGCCAAAGT	Inexact Repeat

4

Conclusion et Perspectives

Sommaire

1 Avant-propos 97

2 Article 2 : SNP calling from RNA-seq data without a reference genome : identification, quantification, differential analysis and impact on the protein sequence 98

3 Supplementary Information 112

Durant ma thèse, je me suis intéressée à l'analyse de données RNA-seq chez les espèces non-modèles, en me confrontant d'une part à de l'analyse d'expression des éléments transposables et des gènes, et d'autre part à l'aspect méthodologique de la détection de variants nucléotidiques à partir des graphes de de Bruijn.

1 Hybrides

Les données RNA-seq provenant *D. mojavensis* et *D. arizonae* et des hybrides réciproques issus de leur croisement ont donné accès à l'expression des gènes et des éléments transposables dans ces quatre lignées.

Dansa cette étude, nous avons choisi d'utiliser les lectures issues du séquençage de lignées parentales et hybrides pour assembler un transcriptome utilisé comme référence (co-assemblage). Nous avons montré l'apport de cette stratégie pour notre étude puisqu'elle elle nous permet d'augmenter artificiellement la profondeur de séquençage et d'assembler plus de gènes et d'ET. Ceci est d'autant plus important que les niveaux d'expression des ET sont en générale faible. Cette approche est rendue possible par le faible taux de divergence entre les génomes les lignées parentales. Néanmoins, une question reste ouverte concernant la généralisation du co-assemblage à d'autres espèces, en particulier : jusqu'à quel taux de divergence entre les espèces séquencées le co-assemblage est-il encore possible et intéressant ? Nous avons pu tester l'apport de cette stratégie dans notre cas, mais je n'ai pas pu poursuivre cette étude par la réalisation de simulations.

L'analyse de ces niveaux d'expression nous a permis d'identifier les gènes et éléments dérégulés chez les hybrides. Nous avons ainsi vu que la majorité des ET sont régulés chez les hybrides, dans les deux sens de croisement, et qu'il n'y a donc pas de dérégulation globale. Seuls quelques rares ET présentent des niveaux d'expression particulièrement importants chez les hybrides. L'élément Copia1 est largement sur-exprimé chez les hybrides issus d'une mère *D. mojavensis*. Un élément de la famille des gypsys est lui très fortement exprimé chez les hybrides issus d'une mère *D. arizonae*. L'analyse du séquençage des piRNA chez les lignées hybrides semblent montrer que la sur-expression de ces deux éléments est associée à une diminution des piRNA secondaires. Le séquençage des piRNA issus des lignées parentales est nécessaire pour une analyse plus poussée. De même, il faudrait des réplicats biologiques pour le séquençage des piRNA. Cela nous permettrait de comparer l'abondance des piRNA entre les lignées hybrides et parentales.

Le pipeline développé ici pourra être réutilisé pour l'analyse des transcriptomes chez les mâles de *D. mojavensis*, *D. arizonae* et leurs hybrides. Contrairement aux femelles, on observe un stérilité des mâles issus du croisement d'une femelle *D. arizonae* avec un mâle *D. mojavensis*. On s'attend donc à des différences plus importantes entre les hybrides liées à la différence de phénotype. L'analyse de données transcriptomiques issues des individus mâles pourraient permettre d'identifier quelles sont différences entre les hybrides à l'origine de la variabilité observée.

De plus, selon les lignées parentales choisies, on observe différentes intensités de stérilité des hybrides. Dans le cadre de l'ARN Exhyb, il est prévu de croiser la lignée de *D. arizonae* utilisée dans ce travail avec trois autres lignées de *D. mojavensis* qui conduisent à des niveaux de stérilité variables chez les hybrides.

Par ailleurs, le séquençage du génome des lignées parentales (en cours) devrait nous permettre d'identifier plus précisément les divergences de séquences entre *D. mojavensis* et *D. arizonae*, mais aussi le nombre de copies d'ET présentes au sein de chaque espèce, en particulier pour GTWIN et Copia1.

J'ai également eu l'opportunité de collaborer avec Valèria Romero Soriano en travaillant sur un autre modèle biologique permettant d'étudier l'impact de l'hybridation inter-spécifique sur la stabilité des génomes. *Drosophila buzzatii* et *Drosophila koepferae* sont deux espèces proches, ayant divergé il y 4 à 5 millions d'années (Gómez and Hasson [2003]). Des hybrides issus de femelles *D. koepferae* ont pu être observés dans la nature (Franco et al. [2010]). Une mobilisation des ET a déjà été détectée chez ces hybrides (Labrador et al. [1999]; Vela et al. [2014]). Dans l'étude présentée en annexe (cf. Annexes, section 1), nous avons séquencé les transcriptomes (ARNm et piRNA) extrait des ovaires des lignées parentales ainsi que de la lignée hybride (F1) et des individus issus du rétro-croisement de ces hybrides avec des mâles *D. buzzatii* (BC1). Les transcriptomes (ARNm et piRNA) extrait des testicules d'individus *D. buzzatii* et F1 ont également séquencés. L'analyse de l'expression gènes et des éléments transposables, ainsi que l'abondance des piRNA, montrent que la divergence entre les piRNA des lignées parentales associée à une divergence (nucléotidique et d'expression) des gènes de la voie des piRNA pourraient être à l'origine des dérégulations d'ET et des instabilités génomiques observées chez les hybrides.

2 Détection des variants nucléotidiques

Dans un second temps, j'ai travaillé sur la détection de SNP à partir de données RNA-seq sans génome de référence. J'ai pour cela utilisé le logiciel KisSplice, qui permet de trouver différents types de variants (épissages, SNP, indels) directement dans le graphe de de Bruijn construit à partir des lectures séquencées. J'ai clarifié les points forts et les limites de cette approche sur des données réelles, en la comparant à des méthodes basées sur l'alignement des lectures sur un génome de référence ou sur un transcriptome assemblé. J'ai également participé au développement de KisSplice2RefTranscriptomequi permet de prédire l'impact des SNP sur les séquences des protéines.

Nous avons montré, sur des données RNA-seq humaines, que les performances de KisSplice, en terme de sensibilité et précision, sont comparables à celles obtenues par des méthodes d'alignement sur génome de référence (comme GATK). La sensibilité et la précision du pipeline sont également meilleures que celles obtenues par alignement des lectures sur transcriptome assemblé. Le pipeline que nous proposons a donc de meilleures performances que les méthodes sans génome de référence, qui (comme nous) utilisent uniquement les données RNA-seq pour l'identification des SNP.

Nous avons appliqué l'ensemble du pipeline sur deux autres jeux de données réels pour lesquelles nous n'avons pas de génome de référence : chez la drosophile ainsi que sur *Asobara tabida*.Nous avons sélectionné plusieurs cas de SNP, qui ont ensuite été validés par rt-PCR et séquençage.

Un des enjeux majeurs est de différencier un vrai SNP présent dans les données de deux types d'"erreurs" : les erreurs de séquençages et les répétition inexactes.

Dans le cas des erreurs de séquençages, on a choisi dans KisSplice de les filtrer à l'aide de deux paramètres, en fonction de leur abondance (cf Chapitre 3, Figure 3). Un premier filtre, généralement utilisé par de nombreuses approches (assemblage ou alignement, en génomique ou transcriptomique) consiste en l'élimination des chemins couverts par trop peu de lectures (par défaut 2). Le second filtre supprime quant à lui les bulles pour lesquelles la quantification relative d'un des chemins est trop faible (par défaut 5%). Ces filtres éliminent néanmoins de vrais SNP, les SNP rares et/ou peu couverts. La valeur par défaut choisie est un compromis entre la nécessité d'obtenir le plus de vrais SNP possibles (une bonne sensibilité) et celle d'avoir le moins de faux positifs (un bonne précision) en sortie de KisSplice.

Les variants liés à des répétitions inexactes créent eux aussi des bulles semblables aux SNP dans le graphe de de Bruijn. La stratégie mise en place dans KisSplice pour les filtrer est basée sur le nombre de branches dans la bulle (cf Chapitre 3, Figure 4). Si un des chemins est branchant, de plus de b branches (par défaut $b = 5$) alors la bulle n'est pas sortie par KisSplice. En faisant cela, on suppose en réalité que les répétitions inexactes sont présentes en un nombre suffisant de copies assez divergentes entre elles pour créer des régions trop branchantes pour être sorties par KisSplice. Ce filtre supprime également des vrais SNP. En effet les SNP présents dans des régions fortement polymorphes sont à l'origine de bulles ayant les même caractéristiques et les répétitions inexactes filtrées. Chez l'Homme, c'est par exemple le cas pour certains gènes de l'immunité (HLA, AbParts). De plus, dans cette étude, nous nous sommes intéressés uniquement aux SNP isolés trouvés par KisSplice, suffisamment distants d'autres variants (distance minimale d'au moins k nucléotides). Certains SNP proches, distants de moins de k nucléotides et dont les bulles sont suffisamment peu branchantes, peuvent également être trouvés par KisSplice, dans un fichier à part (non inclus dans l'étude présentée dans le Chapitre 3). Néanmoins, chez l'Homme la précision de KisSplice sur cette sortie est assez faible, car elle contient des répétitions inexactes elles aussi suffisamment peu branchantes pour être énumérées. De manière générale, KisSplice mais également pour les autres méthodes de détections de SNP, basées sur l'alignement des lectures contre un génome ou un transcriptome de référence, ont des difficultés à détecter les SNP proches et les SNP dans des régions fortement polymorphes. Identifier de tels variants reste donc un problème méthodologique ouvert.

Durant les derniers mois de ma thèse j'ai également commencé à comparer KisSplice avec DiscoSNP (Uricaru et al. [2015]). DiscoSNP et KisSplice ont été développés conjointement dans le cadre d'une collaboration entre plusieurs équipes (Colib'read), ils sont basés sur le même modèle (détection d'une bulle dans un graphe de de Bruijn). DiscoSNP a néanmoins été pensé pour travailler sur des données génomiques (DNA-seq) tandis que KisSplice a été développé pour des données transcriptomiques (RNA-seq) et identifie également les épissages alternatifs. Concernant la détection des SNP, ils diffèrent essentiellement sur leur politique de branchement : DiscoSNP n'autorise que des branchement symétriques. L'hypothèse sous-jacente est qu'un branchement symétrique est indicateur d'une région fortement polymorphe, alors qu'un branchement asymétrique est indicateur d'une erreur de séquençage ou d'une répétition inexacte. Ce présupposé n'a pas été

testé explicitement et mériterait de l'être. La poursuite d'une comparaison des performances de DiscoSNP et KisSplice, à la fois en terme de sensibilité et précision, mais également en temps de calcul et utilisation mémoire, sur des données DNA-seq et RNA-seq, pourrait mettre en évidence les avantages et limites de chaque méthode et de leur politique de branchement.

Un développement possible autour de KisSplice pourrait également permettre d'étudier conjointement les SNP et les épissages. Les SNP proches de variants d'épissage sont théoriquement sortis par KisSplice, mais difficiles à identifier dans la sortie actuelle. Des développements méthodologiques supplémentaires sont nécessaires pour permettre la détection de ce type de variants dans la sortie de KisSplice, ou bien directement dans le graphe de de Bruijn. Dans certains cas, notamment, lorsque les SNP sont distants de plus de k nucléotides du site d'épissage, KisSplice produit une bulle correspondant au SNP et une autre bulle correspondant aux deux variants d'épissage. Il serait possible de faire le lien entre ces deux types de variants, par exemple en les alignant sur une référence (génome ou transcriptome assemblé). Cette possibilité n'est pour le moment pas implémentée dans K2RT mais semble réalisable. Des développements méthodologiques seraient également nécessaires pour tester un éventuel lien entre la présence d'un variant nucléotidique et un variant d'épissage.

Enfin, si en RNA-seq les répétitions sont problématiques pour l'identification de variants dans les graphes de de Bruijn, les identifier pourrait permettre non seulement d'aider à résoudre les problèmes liées à ces zones du graphe, mais aussi à analyser et quantifier les éléments transposables directement dans ce graphe. Une perspective à long terme serait de chercher à identifier les sous-graphes correspondant à des répétitions (familles d'ET mais aussi familles de gènes) pour les quantifier collectivement dans le graphe et analyser la diversité des copies exprimées.

Bibliographie

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*, 11(10) :R106. 30

Anders, S., Pyl, P. T., and Huber, W. (2014). Htseq—a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2) :166. 29

Au, K. F., Jiang, H., Lin, L., Xing, Y., and Wong, W. H. (2010). Detection of splice junctions from paired-end rna-seq data by splicemap. *Nucleic acids research*, 38(14) :4570–4578. 23

Baack, E. J., Whitney, K. D., and Rieseberg, L. H. (2005). Hybridization and genome size evolution : timing and magnitude of nuclear dna content increases in helianthus homoploid hybrid species. *New Phytologist*, 167(2) :623–630. 39

Bao, Z. and Eddy, S. R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. *Genome research*, 12(8) :1269–1276. 34

Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B. J., and Frey, B. J. (2010). Deciphering the splicing code. *Nature*, 465(7294) :53–59. 13

Bastide, M. and McCombie, W. R. (2007). Assembling genomic dna sequences with phrap. *Current Protocols in Bioinformatics*, pages 11–4. 25

Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J. P., and Lander, E. S. (2002). Arachne : a whole-genome shotgun assembler. *Genome research*, 12(1) :177–189. 25

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(7218) :53–59. 16

Bernard, E., Jacob, L., Mairal, J., and Vert, J.-P. (2014). Efficient rna isoform identification and quantification from rna-seq data with network flows. *Bioinformatics*, page btu317. 27

Biémont, C. and Vieira, C. (2003). [the influence of transposable elements on genome size]. *Journal de la Societe de biologie*, 198(4) :413–417. 31

Biemont, C. and Vieira, C. (2006). Genetics : Junk DNA as an evolutionary force. *Nature*, 443(7111) :521–524. 31, 32

Bourque, G. (2009). Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Current Opinion in Genetics & Development*, 19(6) :607 – 612. 32

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5) :525–527. 30

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(6) :1089–1103. 35

Cáceres, M., Ranz, J. M., Barbadilla, A., Long, M., and Ruiz, A. (1999). Generation of a widespread drosophila inversion by a transposable element. *Science*, 285(5426) :415–418. 32

Callinan, P. and Batzer, M. (2006). Retrotransposable elements and human disease. In *Genome and disease*, volume 1, pages 104–115. Karger Publishers. 32

Casacuberta, E. and Gonzalez, J. (2013). The impact of transposable elements in environmental adaptation. *Molecular Ecology*. 31, 32, 33

Chaisson, M. J., Brinza, D., and Pevzner, P. A. (2009). De novo fragment assembly with short mate-paired reads : Does the read length matter ? *Genome research*, 19(2) :336–346. 26

Chen, J.-M., Stenson, P. D., Cooper, D. N., and Férec, C. (2005). A systematic analysis of line-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Human genetics*, 117(5) :411–427. 32

Chénais, B. (2015). Transposable elements in cancer and other human diseases. *Current cancer drug targets*, 15(3) :227–242. 32

Chénais, B., Caruso, A., Hiard, S., and Casse, N. (2012). The impact of transposable elements on eukaryotic genomes : from genome size increase to genetic adaptation to stressful environments. *Gene*, 509(1) :7–15. 31

Collins, F. S., Brooks, L. D., and Chakravarti, A. (1998). A dna polymorphism discovery resource for research on human genetic variation. *Genome research*, 8(12) :1229–1231. 28

DeBarry, J., Liu, R., and Bennetzen, J. (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(1) :235. 34

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star : ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1) :15–21. 23, 29

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput dna sequencing. *Nucleic acids research*, 36(16) :e105–e105. 21

Dupressoir, A., Vernochet, C., Bawa, O., Harper, F., Pierron, G., Opolon, P., and Heidmann, T. (2009). Syncytin-a knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. *Proceedings of the National Academy of Sciences*, 106(29) :12127–12132. 32

Edgar, R. C. and Myers, E. W. (2005). Piler : identification and classification of genomic repeats. *Bioinformatics*, 21(suppl 1) :i152–i158. 34

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). Ltrharvest, an efficient and flexible software for de novo detection of ltr retrotransposons. *Bmc Bioinformatics*, 9(1) :18. 34

Feschotte, C. and Pritham, E. J. (2007). DNA Transposons and the Evolution of Eukaryotic Genomes. *Annual Review of Genetics*, 41(1) :331–68. 31

Franco, F. F., Silva-Bernardi, E. C. C., Sene, F. M., Hasson, E. R., and Manfrin, M. H. (2010). Intra-and interspecific divergence in the nuclear sequences of the clock gene period in species of the drosophila buzzatii cluster. *Journal of Zoological Systematics and Evolutionary Research*, 48(4) :322–331. 135

Gómez, G. A. and Hasson, E. (2003). Transpecific polymorphisms in an inversion linked esterase locus in drosophila buzzatii. *Molecular biology and evolution*, 20(3) :410–423. 135

González, J., Lenkov, K., Lipatov, M., Macpherson, J. M., and Petrov, D. A. (2008). High rate of recent transposable element–induced adaptation in drosophila melanogaster. *PLoS Biol*, 6(10) :e251. 32

Goubert, C., Modolo, L., Vieira, C., Moro, C. V., Mavingui, P., and Boulesteix, M. (2015). De novo assembly and annotation of the asian tiger mosquito (aedes albopictus) repeatome with dnapipete from raw genomic reads and comparative analysis with the yellow fever mosquito (aedes aegypti). *Genome biology and evolution*, 7(4) :1192–1205. 34

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. *Nature biotechnology*, 29(7) :644–652. 26

Grentzinger, T. and Chambeyron, S. (2014). Fast and accurate method to purify small non-coding rnas from drosophila ovaries. *PIWI-Interacting RNAs : Methods and Protocols*, pages 171–182. 21

Guio, L., Barrón, M. G., and González, J. (2014). The transposable element bari-jheh mediates oxidative stress response in drosophila. *Molecular ecology*, 23(8) :2020–2030. 32

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nature biotechnology*, 28(5) :503–510. 27

Han, Y. and Wessler, S. R. (2010). Mite-hunter : a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic acids research*, 38(22) :e199–e199. 34

Hancks, D. C. and Kazazian, H. H. (2016). Roles for retrotransposon insertions in human disease. *Mobile DNA*, 7(1) :1. 32

Hansen, K. D., Brenner, S. E., and Dudoit, S. (2010). Biases in illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*, 38(12) :e131–e131. 21

Huang, S., Zhang, J., Li, R., Zhang, W., He, Z., Lam, T.-W., Peng, Z., and Yiu, S.-M. (2011). Soapsplice : genome-wide ab initio detection of splice junctions from rna-seq data. *Frontiers in genetics*, 2 :46. 23

Huang, X. and Madan, A. (1999). Cap3 : A dna sequence assembly program. *Genome research*, 9(9) :868–877. 25

Huang, X. and Yang, S.-P. (2005). Generating a genome assembly with pcap. *Current Protocols in Bioinformatics*, pages 11–3. 25

Hughes, J. F. and Coffin, J. M. (2005). Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome. *Genetics*, 171(3) :1183–1194. 32

Illumina (2016). An introduction to next-generation sequencing technology. http ://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf. 18

Jin, Y., Tam, O. H., Paniagua, E., and Hammell, M. (2015). Tetranscripts : A package for including transposable elements in differential expression analysis of rna-seq datasets. *Bioinformatics*, page btv422. 36

Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nature methods*, 7(12) :1009–1015. 27

Kelleher, E. S., Edelman, N. B., and Barbash, D. A. (2012). Drosophila interspecific hybrids phenocopy pirna-pathway mutants. *PLoS Biol*, 10(11) :e1001428. 39

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, 115(1) :49–63. 31

Kidwell, M. G. and Lisch, D. R. (2000). Transposable elements and host genome evolution. *Trends in ecology & evolution*, 15(3) :95–99. 31

Kim, D., Langmead, B., and Salzberg, S. L. (2015). Hisat : a fast spliced aligner with low memory requirements. *Nature methods*, 12(4) :357–360. 23

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 research*, 8(5) :464–478. 31

Koch, P., Platzer, M., and Downie, B. R. (2014). Repark—de novo creation of repeat libraries from whole-genome ngs reads. *Nucleic acids research*, page gku210. 34

Labrador, M., Farré, M., Utzet, F., and Fontdevila, A. (1999). Interspecific hybridization increases transposition rates of osvaldo. *Molecular Biology and Evolution*, 16(7) :931–937. 39, 135

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4) :357–359. 29

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 biology*, 10(3) :1. 29

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. 34

Li, B. and Dewey, C. N. (2011). Rsem : accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1) :1. 29

Li, H. (2011). A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21) :2987–2993. 28

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11) :1851–1858. 28

Li, J. J., Jiang, C.-R., Brown, J. B., Huang, H., and Bickel, P. J. (2011). Sparse linear modeling of next-generation mrna sequencing (rna-seq) data for isoform discovery and abundance estimation. *Proceedings of the National Academy of Sciences*, 108(50) :19867–19872. 27

Li, R., Ye, J., Li, S., Wang, J., Han, Y., Ye, C., Wang, J., Yang, H., Yu, J., Wong, G. K.-S., and Wang, J. (2005). Reas : Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Comput Biol*, 1(4) :e43. 34

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome research*, 20(2) :265–272. 26

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., et al. (2012). Comparison of the two major classes of assembly algorithms : overlap–layout–consensus and de-bruijn-graph. *Briefings in functional genomics*, 11(1) :25–37. 24, 25

Lisch, D. (2013). How important are transposons for plant evolution ? *Nature Reviews Genetics*, 14(1) :49–61. 32

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell*, 133(3) :523–536. 16

Liu, Y., Zhou, J., and White, K. P. (2014). Rna-seq differential expression studies : more sequence or more replication ? *Bioinformatics*, 30(3) :301–304. 20

Lynch, M. and Conery, J. S. (2003). The origins of genome complexity. *science*, 302(5649) :1401–1404. 31

Makalowski (2001). The human genome structure and organization. *Acta Biochim Pol*, 48 :587–598. 31

Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. (2012). The gem mapper : fast, accurate and versatile alignment by filtration. *Nature methods*, 9(12) :1185–1188. 23

Mateo, L., Ullastres, A., and González, J. (2014). A transposable element insertion confers xenobiotic resistance in drosophila. *PLoS Genet*, 10(8) :e1004560. 32

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, 36(6) :344–355. 31, 32

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit : a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9) :1297–1303. 28

Metcalfe, C. J., Bulazel, K. V., Ferreri, G. C., Schroeder-Reiter, E., Wanner, G., Rens, W., Obergfell, C., Eldridge, M. D., and O'Neill, R. J. (2007). Genomic instability within centromeres of interspecific marsupial hybrids. *Genetics*, 177(4) :2507–2517. 39

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics*, 11(1) :31–46. 19

Mi, S., Lee, X., Li, X.-P., Veldman, G. M., Finnerty, H., Racie, L., Lavallie, E., Tang, X.-Y., Edouard, P., Howes, S., et al. (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, 403(6771) :785–789. 32

Mishra, B. (2008). Transposable element-driven duplications during hominoid genome evolution. *eLS*. 32

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7) :621–628. 16

Mullikin, J. C. and Ning, Z. (2003). The phusion assembler. *Genome research*, 13(1) :81–90. 25

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., et al. (2000). A whole-genome assembly of drosophila. *Science*, 287(5461) :2196–2204. 25

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, 320(5881) :1344–1349. 16

NHGRI (2014). Dna, alternative splicing. http ://www.genome.gov/Images/EdKit/bio2j_-large.gif. National Human Genome Research Institute. 14

Novak, P., Neumann, P., and Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, 11(1) :378. 34

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12) :1413–1415. 13

Patro, R., Mount, S. M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from rna-seq reads using lightweight algorithms. *Nature biotechnology*, 32(5) :462–464. 30

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nature biotechnology*, 33(3) :290–295. 27

Philippe, N., Salson, M., Commes, T., and Rivals, E. (2013). Crac : an integrated approach to the analysis of rna-seq reads. *Genome biology*, 14(3) :1. 23, 27, 28

Roberts, A. and Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature methods*, 10(1) :71–73. 29

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. *Nature methods*, 7(11) :909–912. 26

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger : a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1) :139–140. 30

Sacomoto, G. A., 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(6) :1. 28

Saito, K. and Siomi, M. C. (2010). Small rna-mediated quiescence of transposable elements in animals. *Developmental cell*, 19(5) :687–697. 33

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of molecular biology*, 94(3) :441–448. 13

SanMiguel, P., Tikhonov, A., Jin, Y. K., Motchoulskaia, N. Zakharov, D., Melake-Berhan, A., Springer, P. S., Edwards, K. J., Lee, M., Avramova, Z., , and Bennetzen, J. L. (1996). Nested retrotransposons in the intergenic regions of the maize genome. *Science*, 274 :765–768. 31

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(5956) :1112–1115. 31

Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. (2012). Oases : robust de novo rna-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8) :1086–1092. 26

Sela, N., Mersch, B., Gal-Mark, N., Lev-Maor, G., Hotz-Wagenblatt, A., and Ast, G. (2007). Comparative analysis of transposed element insertion within human and mouse genomes reveals alu's unique role in shaping the human transcriptome. *Genome biology*, 8(6) :1. 37

Sendler, E., Johnson, G. D., and Krawetz, S. A. (2011). Local and global factors affecting rna sequencing analysis. *Analytical biochemistry*, 419(2) :317–322. 22

Shen, S., Park, J. W., Huang, J., Dittmar, K. A., Lu, Z.-x., Zhou, Q., Carstens, R. P., and Xing, Y. (2012). Mats : a bayesian framework for flexible detection of differential alternative splicing from rna-seq data. *Nucleic acids research*, page gkr1291. 27

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., and Birol, I. (2009). Abyss : a parallel assembler for short read sequence data. *Genome research*, 19(6) :1117–1123. 26

Siomi, M. C., Sato, K., Pezic, D., and Aravin, A. A. (2011). Piwi-interacting small rnas : the vanguard of genome defence. *Nature reviews Molecular cell biology*, 12(4) :246–258. 33

Sniegowski, P. D. and Charlesworth, B. (1994). Transposable element numbers in cosmopolitan inversions from a natural population of drosophila melanogaster. *Genetics*, 137(3) :815–827. 32

Tilgner, H., Knowles, D. G., Johnson, R., Davis, C. A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T. R., and Guigó, R. (2012). Deep sequencing of subcellular rna

fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncrnas. *Genome research*, 22(9) :1616–1625. 37

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). Tophat : discovering splice junctions with rna-seq. *Bioinformatics*, 25(9) :1105–1111. 23

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. *Nature biotechnology*, 28(5) :511–515. 27

Uricaru, R., Rizk, G., Lacroix, V., Quillery, E., Plantard, O., Chikhi, R., Lemaitre, C., and Peterlongo, P. (2015). Reference-free detection of isolated snps. *Nucleic acids research*, 43(2) :e11–e11. 137

van de Lagemaat, L. N., Gagnier, L., Medstrand, P., and Mager, D. L. (2005). Genomic deletions and precise removal of transposable elements mediated by short identical dna segments in primates. *Genome research*, 15(9) :1243–1249. 32

Vela, D., Fontdevila, A., Vieira, C., and Guerreiro, M. P. G. (2014). A genome-wide survey of genetic instability by transposition in drosophila hybrids. *PloS one*, 9(2) :e88992. 39, 135

Villesen, P., Aagaard, L., Wiuf, C., and Pedersen, F. S. (2004). Identification of endogenous retroviral reading frames in the human genome. *Retrovirology*, 1(1) :1. 32

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., et al. (2010). Mapsplice : accurate mapping of rna-seq reads for splice junction discovery. *Nucleic acids research*, 38(18) :e178–e178. 23

Wanunu, M. (2012). Nanopores : A journey towards dna sequencing. *Physics of life reviews*, 9(2) :125–158. 19

Wei, Z., Wang, W., Hu, P., Lyon, G. J., and Hakonarson, H. (2011). Snver : a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic acids research*, 39(19) :e132–e132. 28

Wu, T. D. and Nacu, S. (2010). Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7) :873–881. 23

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., et al. (2014). Soapdenovo-trans : de novo transcriptome assembly with short rna-seq reads. *Bioinformatics*, 30(12) :1660–1666. 26

Zerbino, D. R. and Birney, E. (2008). Velvet : algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5) :821–829. 26

Zhang, Y., Lameijer, E.-W., AC't Hoen, P., Ning, Z., Slagboom, P. E., and Ye, K. (2012). Passion : a pattern growth algorithm-based pipeline for splice junction detection in paired-end rna-seq data. *Bioinformatics*, 28(4) :479–486. 23

Zhang, Z. and Wang, W. (2014). Rna-skim : a rapid method for rna-seq quantification at transcript level. *Bioinformatics*, 30(12) :i283–i292. 30

Zytnicki, M., Akhunov, E., and Quesneville, H. (2014). Tedna : a transposable element de novo assembler. *Bioinformatics*, page btu365. 34

A

Articles annexes

Sommaire

1 **Hybrides** 134

2 **Détection des variants nucléotidiques** 136

1 Hybrides et éléments transposables

Manuscripts submitted to Genome Biology and Evolution

OXFORD UNIVERSITY PRESS | Genome Biology and Evolution

Transposable element misregulation is linked to the divergence between parental piRNA pathways in Drosophila hybrids

Journal:	*Genome Biology and Evolution*
Manuscript ID	GBE-161024
Manuscript Type:	Research Article
Date Submitted by the Author:	21-Oct-2016
Complete List of Authors:	Romero-Soriano, Valèria; Universitat Autònoma de Barcelona, Departament de Genètica i Microbiologia Modolo, Laurent; Université Lyon 1 - CNRS, lab. Biométrie et biologie Evolutive, UMR CNRS 5558 López-Maestre, Hélène; Université Lyon 1 - CNRS, lab. Biométrie et biologie Evolutive, UMR CNRS 5558 Mugat, Bruno; Institut de Genetique Humaine Pessia, Eugénie; Université Lyon 1 - CNRS, lab. Biométrie et biologie Evolutive, UMR CNRS 5558 Chambeyron, Séverine; Institut de Genetique Humaine Vieira, Cristina; University Lyon 1, UMR CNRS 5558 Garcia Guerreiro, Maria Pilar; Universitat Autònoma de Barcelona, Departament de Genètica i Microbiologia
Keywords:	transposable elements, piRNAs, interspecific hybridization, RNA-seq, Drosophila buzzatii, Drosophila koepferae

SCHOLARONE™ Manuscripts

Transposable element misregulation is linked to the divergence between parental piRNA pathways in *Drosophila* hybrids

Authors:

Valèria Romero-Soriano[1], Laurent Modolo[2], Hélène Lopez-Maestre[2], Bruno Mugat[3], Eugénie Pessia[2], Séverine Chambeyron[3], Cristina Vieira[2] and Maria Pilar Garcia Guerreiro[1*] .

[1] Grup de Genòmica, Bioinformàtica i Biologia Evolutiva, Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona. 08193 Bellaterra, Barcelona, Spain.

[2] Laboratoire de Biométrie et Biologie Evolutive, UMR5558, Université Claude Bernard Lyon 1, Villeurbanne, France.

[3] Institut de Génétique Humaine, CNRS, UPR1142, 34396 Montpellier Cedex 5, France.

* Corresponding author: mariapilar.garcia.guerreiro@uab.cat

Data deposition: Sequence data from this article is currently being submitted to Sequence Read Archive (SRA); the process will be finished before acceptation.

Keywords: Transposable elements, piRNAs, interspecific hybridization, RNA-seq, *Drosophila buzzatii*, *Drosophila koepferae*.

Abstract

Interspecific hybridization is a genomic stress condition that leads to the activation of transposable elements (TEs) in both animals and plants. In hybrids between *Drosophila buzzatii* and *Drosophila koepferae*, mobilization of at least 28 TEs has been described. However, the molecular mechanisms underlying this TE release remain poorly understood. To give insight on the causes of this TE activation, we performed a TE transcriptomic analysis in ovaries (notorious for playing a major role in TE silencing) of parental species and their F1 and backcrossed (BC) hybrids. We find that 15.2% and 10.6% of the expressed TEs are deregulated in F1 and BC1 ovaries respectively, with a bias towards overexpression in both cases. While differences between parental piRNA populations explain only partially these results, we demonstrate that piRNA pathway proteins have divergent sequences and are differentially expressed between parental species. Thus, a functional divergence of the piRNA pathway between parental species, together with some differences between their piRNA pools, might be at the origin of hybrid instabilities and ultimately cause TE misregulation in ovaries. These analyses were complemented with the study of F1 testes, where TEs tend to be less expressed than in *D. buzzatii*. This can be explained by an increase in piRNA production, which probably acts as defence mechanism against TE instability in the male germline. Hence, we describe a differential impact of interspecific hybridization in testes and ovaries, which reveals that TE expression and regulation are sex-biased.

Introduction

Transposable elements (TEs) are mobile DNA fragments that are dispersed throughout the genome of the vast majority of both prokaryotic and eukaryotic organisms. Their capacity to mobilize, together with their repetitive nature, confers them a high mutagenic potential. TE insertions can be responsible for the disruption of genes or regulatory sequences, and can also cause chromosomal rearrangements, representing a threat to their host genome integrity (Hedges & Deininger 2007). To mitigate these deleterious effects, mechanisms of TE control are especially important in the germline, where novel insertions (as well as other mutations) can be transmitted to the progeny (Iwasaki et al. 2015; Czech & Hannon 2016).

Animal genomes have developed a TE silencing system, the piRNA (Piwi-interacting RNA) pathway (Klattenhoff & Theurkauf 2008; Brennecke & Senti 2010), that acts in the germline at both post-transcriptional and transcriptional levels (Rozhkov et al. 2013). piRNA templates form specific genomic clusters, whose transcription produces long piRNA precursors that are cleaved to produce primary piRNAs (Brennecke et al. 2007). The resulting piRNAs can initiate an amplification loop called the ping-pong cycle, giving rise to secondary piRNAs (Brennecke et al. 2007; Gunawardane et al. 2007). A third kind of piRNAs are produced by phased cleavage of piRNA cluster transcript remnants that have first been processed during secondary piRNA biogenesis (Han et al. 2015; Mohn et al. 2015). In the soma, another small-RNA mediated silencing system, the endo-siRNA (endogenous small interference RNA) pathway, has been shown to be involved in post-transcriptional silencing of TEs (Ghildiyal et al. 2008).

These strong mechanisms of TE regulation can be relaxed under different stress conditions, leading to unexpected TE mobilization events (García Guerreiro 2012). Hybridization between species causes a genomic stress that can lead to several genome reorganizations that seem to be driven by TEs (Fontdevila 2005; Michalak 2009; García Guerreiro 2014; Romero-

Soriano et al. 2016). In the literature, several cases of TE proliferation in interspecific hybrids have been reported for a wide range of species, including plants (Liu & Wendel 2000; Ungerer et al. 2006; Wang et al. 2010) as well as animals (Evgen'ev et al. 1982; O'Neill et al. 1998; Metcalfe et al. 2007). Studies describing an enhanced TE expression in hybrids suggest that this may be caused by a TE silencing breakdown (Kelleher et al. 2012; Carnelossi et al. 2014; Dion-Côté et al. 2014; Renaut et al. 2014; García Guerreiro 2015). In this work, we propose two possible explanatory hypotheses –not mutually exclusive– to understand this breakdown, since the molecular mechanisms allowing TE release in hybrids remain unknown.

The first hypothesis, that we call the maternal cytotype failure, recalls the hybrid dysgenesis phenomenon (Picard 1976; Kidwell et al. 1977), where an increase of TE activity is observed. This occurs when *Drosophila* females whose genome is devoid of a particular TE are mated with males containing it, and is associated with the absence of specific piRNAs in the maternal cytoplasm (Brennecke et al. 2008), which are crucial to initiate an efficient TE silencing response in the progeny (Grentzinger et al. 2012). In the same logic, differences between parental species piRNA pools could lead to a transcriptional activation of some paternally-inherited TEs in interspecific hybrids. Under this hypothesis, only a subset of TE families, specific to the male species, would be deregulated after hybridization.

The second hypothesis claims that a global failure of the piRNA pathway is responsible for the observed TE activation in hybrids. It has been shown that piRNA pathway effector proteins show adaptive evolution marks (Obbard et al. 2009; Simkin et al. 2013) and their expression levels can significantly differ between different populations of the same *Drosophila* species (Fablet et al. 2014). Thus, genetic incompatibilities involving this pathway could arise even between closely related species. The accumulated functional divergence of these proteins would cause a widespread transcriptional TE derepression, as suggested in *D. melanogaster-D. simulans* artificial (*Hmr*-rescued) hybrids (Kelleher et al. 2012).

In order to test these hypotheses and provide new insight into the mechanisms underlying TE activation in hybrids, we have performed a whole-genome study of TE expression and regulation using the species *D. buzzatii* and *D. koepferae* (*buzzatii* complex, *repleta* group). We chose this species pair as a model because hybridization between them can occur in nature (Gomez & Hasson 2003; Piccinali et al. 2004; Franco et al. 2010), providing a source of genetic variability that makes them particularly interesting for natural hybridization and speciation studies. Contrarily to *D. melanogaster* and *D. simulans*, our species allow backcrosses to be performed (Marín & Fontdevila 1998; Barbash 2010), even if their divergence time appears to be higher: 4.0-5.0 Mya for *D. buzzatii-D. koepferae* (Gomez & Hasson 2003; Laayouni et al. 2003; Oliveira et al. 2012) compared to 1.0-3.0 for *D. melanogaster-D. simulans* (Cutter 2008; Russo et al. 1995; Lachaise & Silvain 2004). Furthermore, several TE mobilization events have previously been detected in our hybrids by *in situ* hybridization (Labrador et al. 1999), amplified fragment length polymorphism (AFLP) markers (Vela et al. 2011) and/or transposon display (Vela et al. 2014). Finally, at least two of the mobilized elements, the retrotransposons *Osvaldo* and *Helena*, present abnormal patterns of expression in hybrids (García Guerreiro 2015; Romero-Soriano & García Guerreiro 2016), pointing to a failure of TE silencing.

We demonstrate that 15.2% of the expressed TE families are deregulated in F1 hybrid ovaries, in most cases overexpressed. This proportion decreases to 10.6% after a generation of backcrossing. However, even if differences between parental piRNA pools can be linked to the misexpression of some TE families, they do not explain the whole pattern of deregulation. Accordingly, our analyses of genomic TE content show that parental TE landscapes are very similar, and hence big differences in their piRNA populations are not expected. On the other hand, we demonstrate that the piRNA pathway proteins are particularly divergent between *D. buzzatii* and *D. koepferae* translated transcriptomes, which seems to lead to dissimilarities in

their piRNA production strategies. Interestingly, a high proportion of the overexpressed TEs do not have associated piRNA populations in parents (nor in hybrids), pointing out a complex TE deregulation network where a failure of the piRNA pathway together with other TE silencing mechanisms would take place. Finally, we show that the effects of hybridization are sex-biased, since in testes (contrarily to ovaries) TE deregulation is globally biased towards underexpression, which can be explained by a higher production of piRNAs in hybrid males.

Material and Methods

Drosophila stocks and crosses

Interspecific crosses were performed between males of *D. buzzatii* Bu28 strain, an inbred line originated by the union of different populations (LN13, 19, 31 and 33) collected in 1982 in Los Negros (Bolivia); and females of *D. koepferae* Ko2 strain, an inbred line originated from a population collected in 1979 in San Luis (Argentina). Both lines were maintained by brother-sister mating for more than a decade and are now kept by mass culturing.

We performed 45 different interspecific crosses of 10 *D. buzzatii* males with 10 *D. koepferae* virgin females (in order to obtain F1 individuals), then 30 backcrosses of 10 *D. buzzatii* males with 10 hybrid F1 females (which gave rise to BC1 females). All stocks and crosses were reared at 25ºC in a standard *Drosophila* medium supplemented with yeast.

RNA extraction, library preparation and sequencing

Flies were dissected in PBT (1× phosphate-buffered saline [PBS], 0.2% Tween 20), 5-6 days after their birth. Total RNA was purified from testes (n=30 pairs per sample for *D. buzzatii* and n=45 pairs per sample for F1 hybrids) or ovaries (n=20 pairs per sample) with the Nucleospin RNA purification kit (Macherey-Nagel). RNA quality and concentration was evaluated using Experion Automated Electrophoresis System (Bio-rad), in order to keep only high quality samples. Two Illumina libraries of 250-300bp fragments were prepared for each kind of sample (*D. buzzatii*, *D. koepferae*, F1 and BC1 ovaries; and *D. buzzatii* and F1 testes), using 2µg of purified RNA. Duplicate libraries correspond to biological replicates (ovaries from different crosses and separate RNA extractions). Sequencing was performed using the Illumina mRNA-seq paired-end protocol on a HiSeq2000 platform, at the INRA–UMR AGAP (Montpellier, France). We obtained 53.5 to 59.1 million paired-end reads for each sample (divided in two replicates) resulting in a total of 332.7 million paired-end reads.

Assembly and annotation

A *de novo* reference transcriptome was constructed for each of our target species using Trinity r2013_08_14 (Grabherr et al. 2011) with options *–group_pairs_distance* 500 and *–min_kmer_cov* 2. All contigs were aligned to *D. buzzatii* genome (Guillén et al. 2015) using BLAT v.35x1 (Kent 2002), with parameters *–minIdentity=80* and *–maxIntron=75000*, in order to identify chimers. Contigs that aligned partially (≤60%) on up to 3 genomic locations with a total alignment coverage of ≥80% were considered chimeric and split consequently.

Finally, to annotate protein-coding genes, all contigs of both transcriptomes were aligned against the *D. buzzatii* predicted gene models and the *D. buzzatii* genome (Guillén et al. 2015) using BLAT v.35x1 (same parameters as before). This approach allows us to identify untranslated regions and double-check the genomic position associated to a contig. Only contigs with alignment coverages ≥70% and whose best hit genomic coordinates overlapped in both alignments were annotated. The same approach was applied to the remaining non annotated contigs with *D. mojavensis*' gene models. The rest of the contigs were clustered using CD-HIT v4.5.4 (Fu et al. 2012) with options *-c 0.8*, -T 0, *-aS 0.8*, *-A 80*, *-p 1*, *-g 1*, *-d 50*; and annotated with the name of the longest sequence of each cluster. Supplementary table S1 depicts a summary of annotation statistics.

TE library construction

Our library is mainly constituted by the list of all TE copies masked in the *D. buzzatii* genome (because *D. koepferae* has not until now been sequenced). In order to have a better representation of *D. koepferae* TE landscape and increase specificity in further analyses, we annotated TE transcripts from our *de novo* assemblies by aligning them to a consensus TE library (the same used to mask *D. buzzatii* genome) using BLAT v.35x1. Contigs whose alignments covered ≥80% of their sequences with a minimum 80% identity and ≥80 bp long

(*three 80 criteria*) were kept as TE transcripts and included in our TE library. To improve our coverage and sensitivity to detect poorly expressed TEs, a third *de novo* assembly, using all the reads from all sequenced samples (from both parents and hybrids) was performed and annotated as described above.

This resulted in 65,772 final TE copies belonging to 699 TE families, which were assigned to only 658 families after two steps of clustering. Clustering was performed using the *three 80 criteria*; manually through BLAT alignments, and automatically using CD-HIT v4.5.4 (same parameters as in gene annotation). These 658 families were divided in 5 categories, following Repbase classification (Jurka et al. 2005): LTR and LINE (class I), DNA and RC (class II) and Unknown (unclassified).

Small RNA extraction, library preparation and sequencing

Small RNA was purified from ovaries (n=70 pairs for all samples) and testes (n=96 pairs for *D. buzzatii* and n=333 pairs for F1 sterile males), following the manual small RNA purifying protocol described by Grentzinger et al. (2013), which significantly reduces endogenous contamination and degradation products abundance. After small RNA isolation, samples were gel-purified and precipitated. A single Illumina library was prepared for each sample and sequencing was performed on an Illumina Hiseq 2500 platform by FASTERIS SA (Switzerland). We obtained a total of 401.1 million reads (21.4 to 58.7 million reads per sample). Reads of 23-30 nucleotides were kept as piRNAs.

TE analyses: read mapping and differential expression

All our sequencing data was trimmed using UrQt (Modolo & Lerat 2015), in order to remove polyA tails (for RNA-seq) and low-quality nucleotides (for both RNA-seq and piRNA-seq). The resulting trimmed reads were aligned to our TE library using Bowtie2 v2.2.4 for RNA-seq (Langmead & Salzberg 2012) and Bowtie1 v1.1.1 for piRNAs (Langmead et al. 2009),

with the default options implemented in TEtools pipeline (the most sensitive option and keeping a single alignment for reads mapping to multiple positions, *--very-sensitive* for Bowtie2 and *-S* for Bowtie). The read count step (built in TE tools: https://github.com/l-modolo/TEtools) was computed per TE family (adding all reads mapped on copies of the same family). Finally, we performed the differential expression analyses between TE families using the R Bioconductor package DESeq2 (Love et al. 2014) on the raw read counts, using the Benjamini-Hochberg multiple test correction (FDR level of 0.1). Statistical summaries of these analyses are available in Supplementary files S1 and S5, including both raw and normalized read count tables. TE families with ≤10 aligned reads per sample are considered to be unexpressed in the text. For piRNA analyses, no significant differences could be detected at the TE family level due to the lack of replicates, leading us to perform the analyses using FC values.

Gene analyses: read mapping, differential expression and GO enrichment

Gene expression analyses were performed following the same approach used for TEs. RNA-seq reads were aligned against the addition of *D. buzzatii* and *D. koepferae* transcriptomes, and read count was computed per annotated gene (by adding all reads mapped on contigs with the same annotation).

Trinity's tool TransDecoder (Haas et al. 2013) was employed to predict ORFs within *D. buzzatii* and *D. koepferae* transcriptomes, using Pfam-A database v.29 (Punta et al. 2012). Then, we performed a functional annotation of the resulting proteomes using GO terms (The Gene Ontology Consortium 2000). For that, we used eggnog-mapper tool (https://github.com/jhcepas/eggnog-mapper): we first mapped our sequences to eggNOG orthologous groups from eukaryotic, bacterial and archaeal databases (Huerta-Cepas et al. 2016) using an e-value of 0.001. Then, we transferred the GO terms of the best orthologous group hit for each gene. GO enrichments for deregulated genes in hybrids were analysed

using the Topology-Weighted method built in Ontologizer (Bauer et al. 2008), with a p-value threshold of 0.01.

Divergence time and TE landscapes of parental species

In order to identify contig pairs between *D. buzzatii* and *D. koepferae*, all sequences ≥2000 bp of the *D. buzzatii de novo* transcriptome were aligned against *D. koepferae*'s using BLAST (McGinnis & Madden 2004). We kept only the best hit for each query and subject, resulting in a total of 2,656 contig pairs, which were translated using EMBOSS getorf (Rice et al. 2000). We used the most likely protein sequences of each contig pair (*i.e.* the longest) to perform codon alignments with MUSCLE (Edgar 2004). Finally, the *dS* rate of each pair was calculated using the codeml program in PAML version 4 (Yang 2007). Divergence time was estimated as in Keightley et al. (2014) using the obtained *dS* mode.

We examined the repeatomes of *D. buzzatii* and *D. koepferae* using dnaPipeTE pipeline (Goubert et al. 2015), which assembles repeats from low coverage genomic NGS data and annotates them with RepeatMasker Open-4.0 (Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. 1996–2010, http://www.repeat- masker.org, last accessed February 24, 2016) and Tandem repeats finder (Benson 1999). We employed Repbase library version 2014-01-31 (Jurka et al. 2005). For both species, two iterations were performed using a read sample size corresponding to a genome coverage of 0.25X (Guillén et al. 2015), according to genome size estimates in Romero-Soriano et al. (2016). Because mitochondrial DNA is usually assembled, all dnaPipeTE contigs were aligned to BLAST nucleotide collection (McGinnis & Madden 2004) to distinguish nuclear from mitochondrial sequences. Reads mapping to mitochondrial contigs were identified using Bowtie2 with default parameters (Langmead & Salzberg 2012) and filtered out. DnaPipeTE was then run without mithocondrial reads (same parameters).

Ping-pong signature identification

The ping-pong cycle is mediated by Aubergine and Ago3 proteins, which cleave the piRNA precursor (or TE transcript) preferentially 10 bp after its 5' end. Thus, sense and antisense reads overlapped by 10 nucleotides are produced during secondary piRNA biogenesis (Klattenhoff & Theurkauf 2008). We aligned our piRNA raw reads (23-30nt, without any trimming step in order to maintain their real size) against the whole TE library using Bowtie1 (*-S* option) and checked for the presence of 10nt-overlapping sense-antisense read pairs using the *signature.py* pipeline (Antoniewski 2014). The same analysis was carried out separately for each of the TE families of the library.

piRNA pathway proteins ortholog search

Proteomes of *D. buzzatii* and *D. koepferae* (see *Gene analyses* section) were aligned against each other using BLAST. Identity percentages of each protein best hit were kept and used to calculate the median identity percentage between *D. buzzatii* and *D. koepferae*.

We identified the orthologs of 30 proteins involved in piRNA biogenesis (Yang & Pillai 2014) in *D. buzzatii* and *D. koepferae* proteomes by reciprocal best blast hit analysis, using their *D. melanogaster* counterparts as seeds (EnsemblMetazoa 27 release, Cunningham et al. 2015), with and e-value cutoff of 1e-05. *D. buzzatii* proteins were aligned against their *D. koepferae* ortholog using BLAST, in order to evaluate their identity percentage.

Results

Qualitative changes in TE expression after interspecific hybridization

We sequenced the ovarian transcriptomes of both parental species and two hybrid generations, the F1 and a first backcross BC1 (Figure 1), and examined their TE expression. We also sequenced and analysed the testicular transcriptomes of *D. buzzatii* (male parental species) and F1 hybrids. Globally, we detected expression of 415 out of 658 candidate TE families (Supplementary file S1). We show that ovaries present significantly higher TE global alignment rate than testes (Figure 2A; Student's t=4.09, p=0.0035) whereas the global TE alignment rate between hybrids and parental species is not significantly different (Student's t=-1.10, p=0.30). At a qualitative level, we observe notable differences between parents and hybrids: LTR proportion is increased in both hybrid testes (from 14.2 to 31.4%) and ovaries (from 7.7-8.3 to 14.4-13.8%), as well as are RC elements (*Helitron*) in F1 testes (from 4.3 to 8.1%, Figure 2B). TE expression profiles are very similar between ovaries of *D. buzzatii* and *D. koepferae*, but parental testes (*D. buzzatii*) present a considerably lower LINE proportion (Figure 2B). In all cases, TE expression is mainly represented by retrotransposons (LINEs are the most expressed category followed by LTRs). Therefore, even if the global amounts of TE expression remain unchanged after interspecific hybridization, we observe differences at the TE family expression level.

TE deregulation in hybrid ovaries is biased towards overexpression

Compared to *D. buzzatii* and *D. koepferae* separately, F1 ovaries present a similar number of differentially expressed TE families (221 and 234, respectively), while in BC1 expression is closer to *D. buzzatii* (149 and 254, Figure 3A). In both cases, hybrid ovaries present a bias towards TE overexpression compared to parental species (Figure 3A), with 55% of the deregulated families (on average) more expressed in hybrids (Supplementary table S2).

When compared to both parental species, 37 TE families are significantly overexpressed in F1 and only 27 in BC1 (most of them are shared between generations, Table 1). Among them, 77% are retrotransposons, and *Gypsy* elements exhibit the highest fold change (FC) values. Surprisingly, we also observe 26 underexpressed families in F1 and 17 in BC1 (Table 2). Underexpressed TE families are also mainly retrotransposons (71%) and their FC values tend to be lower than those of overexpressed families (Tables 1 and 2).

Therefore, after a generation of backcrossing, the global amount of TE deregulation decreases from 15.2 to 10.6% of the 415 expressed families. In the same way, we observe that FC values are often lower in BC1 than in F1 (Tables 1 and 2). All the deregulated TE families are transcriptionally active in both parental species (Figure 3B), but only 21% of them exhibit differences of expression higher than 2-fold between parental species (a total of 16 families; 14 overexpressed and 2 underexpressed, Figure 3B).

Divergence time between parental species and TE landscapes influence deregulation

In a previous study, *D. simulans*-*D. melanogaster* artificial hybrid (*Hmr*-rescued) ovaries displayed a proportion of deregulated TE families of 12.1% (similar to *D. buzzatii*-*D. koepferae* 15.2% in F1), which was considered to be widespread compared to the 0.7% found for protein-coding genes (Kelleher et al. 2012). To evaluate the extent of gene deregulation in our hybrids, we produced a *de novo* transcriptome assembly for each parental species and annotated them using BLAT alignments against gene models of *D. buzzatii* (Guillén et al. 2015) and *D. mojavensis* (*Drosophila* 12 Genomes Consortium 2007) genomes (see Methods).

We annotated 70.9% of the final transcriptome contigs (Supplementary table S1) as 11,190 different protein-coding genes. Among these, 657 are overexpressed and 821 underexpressed in F1 ovaries (Supplementary file S2), reaching a proportion of deregulation of 13.2%. In

BC1, it decreases to 12.3%, with 711 overexpressed and 662 underexpressed genes (Supplementary file S2). Thus, both TE and gene expression are affected at similar levels (~10-15%) in ovaries of *D. buzzatii-D. koepferae* hybrids, but they follow distinct patterns (only TEs are biased towards overexpression). It is noteworthy that F1 and BC1-overexpressed genes have in common three enriched Gene Ontology (GO) terms: response to methotrexate, GABA receptor activity and cation-aminoacid symporter activity (Supplementary table S3). More interestingly, in the case of underexpressed genes, several enriched GO terms related to aminoacid metabolism, ion transport and oogenesis are shared between F1 and BC1 (Supplementary table S3), which may be related to the hybrid loss of fertility.

Alteration of gene expression is remarkably higher in our hybrids than in *D. simulans-D. melanogaster* ones, which might be due to differences in divergence times between these species pairs. We have calculated the most common rate of substitution per synonymous site between our parental species (*dS*=0.139; Supplementary file S3) and estimated their divergence time at 4.96 Mya using Keightley's mutation rate estimate (2014). This result concurs with the few available estimations of divergence between this species pair, that range between 4.02-4.63 Mya (Laayouni et al. 2003; Gomez & Hasson 2003; Oliveira et al. 2012). Using the same formula, *D. melanogaster* and *D. simulans* (with *dS*=0.068, Cutter 2008) would have diverged 2.43 Mya, which is in concordance with the most commonly used estimation (2-3 Mya, Lachaise & Silvain 2004) and confirms that the latter species pair are more closely related.

In spite of being closely related, *D. melanogaster* and *D. simulans* have radically different TE contents: while mostly recent and active TE copies that account for 15% of the genome are found in *D. melanogaster*; *D. simulans* carries mainly old and deteriorated copies, representing 6.9% of the genome (Modolo et al. 2014). We have examined the repeatomes of

our parental species using dnaPipeTE (Goubert et al. 2015), which revealed that both their TE landscapes and abundance are very similar (Supplementary figure S1 and file S4). Both species seem to share similar kinds and proportions of recent and active TEs, suggesting that species divergence (rather than differences in TE content) would cause TE deregulation in our hybrids, which recalls the piRNA pathway failure hypothesis.

Differences in parental piRNA pools cannot fully explain hybrid TE expression

Differences in piRNA pools between parental species ovaries can be at the origin of TE silencing impairment (Brennecke et al. 2008), especially when piRNA levels of a particular TE are lower in the maternal species, *D. koepferae*. To test the maternal cytotype failure hypothesis, we sequenced and analysed the piRNA populations of the samples presented in Figure 1. Globally, antisense regulatory piRNA populations (23-30nt) were detected for 392 out of 658 candidate TE families (Supplementary file S5), mostly retrotransposons. In this case, we performed the differential expression analyses using FC values (see Methods).

A total of 196 TE families present differences higher than 2-fold between *D. buzzatii* and *D. koepferae* ovarian antisense piRNA populations (Figure 4A). Families having lower levels of piRNAs in the maternal species are not always overexpressed: among the 98 TE families that exhibit reduced abundance of piRNAs in *D. koepferae*, only 8 are overexpressed in hybrids (either in F1 or BC1, Figure 4B-i). Reciprocally, families having higher levels of piRNAs in the maternal species are not more commonly underexpressed: only 12 out of 98 families with higher piRNA abundance in *D. koepferae* are classified as underexpressed (Figure 4B-iii). Actually, some deregulated TE families even present the opposite pattern (e.g. *Gypsy6-I* or *Howili1*, Figure 4A). However, this does not mean that differences between piRNA pools cannot account for some specific cases of TE deregulation (e.g. *TART_B1* or *MINOS*, Figure 4A).

Interestingly, 12 of the overexpressed families are among those without associated piRNA populations (Figure 4B-iv), indicating that other TE regulation mechanisms (if any) could be responsible for their regulation in the ovaries.

piRNA production strategies differ between parental species

Artificial hybrids between *D. simulans* and *D. melanogaster* present deficient piRNA production, which displaces the size distribution of ovarian piRNAs (23-30nt) towards miRNAs and siRNAs (18-22 nt) (Kelleher et al. 2012). However, our hybrids present an overall size distribution pattern similar to *D. koepferae* (Figure 5A) and similar (to higher) levels of piRNAs than parental species (Supplementary file S5). Thus, our results show that piRNAs are produced in *D. buzzatii*-*D. koepferae* hybrids.

Interestingly, we note that size distribution of small RNA populations differs between our parental species (Figure 5A): *D. koepferae* exhibits abundant piRNAs and lower levels of miRNAs and siRNAs, whereas the opposite is observed in *D. buzzatii*. These differential amounts of piRNAs between our parental species might be due to a functional divergence in their piRNA biogenesis pathways. To get greater insight into piRNA production strategies, we have assessed the functionality of the secondary biogenesis pathway in our samples. In the germline, mature piRNAs (either maternal or primary) can initiate an amplification loop called the ping-pong cycle, yielding sense and antisense secondary piRNAs (Brennecke et al. 2007; Gunawardane et al. 2007). In this loop, piRNAs are cleaved 10 bp after the 5' end of their template, a feature that is specific to this pathway and can be used to recognize secondary piRNAs. We have determined the ping-pong signature in our sequenced piRNA populations (Antoniewski 2014) and revealed that *D. buzzatii*'s ping-pong fraction is higher than *D. koepferae*'s (Figure 5B), which is in agreement with the idea of divergence in piRNA biogenesis between them.

In hybrids, ping-pong signature levels in F1 and BC1 ovaries are intermediate between parental species (F1 is more similar to D. *koepferae* and BC1 to *D. buzzatii*, Figure 5B), whereas in *D. simulans-D. melanogaster* artificial hybrids, a reduced ping-pong fraction was observed (Kelleher et al. 2012). Therefore, our hybrids differ from *D. melanogaster-D. simulans* model in that they are not characterized by a widespread decrease of piRNA production: although a few TE families present lower levels of piRNAs than both parental species (Supplementary file S6), they do not always coincide with the upregulated ones.

Interestingly, half of the overexpressed TE families (a total of 20, including the 12 without associated piRNA populations described in Figure 4B-iv) do not present traces of ping-pong amplification (Supplementary figure S2). Eleven of them are LINE retrotransposons, of which five belong to the R1 clade, whose members have a high target-specificity for 28S rRNA genes in arthropods (Eickbush et al. 1997; Kojima & Fujiwara 2003). The eight families with associated piRNA populations but without ping-pong signal could possibly be somatic elements, expressed in follicle cells of the ovaries, where secondary piRNA biogenesis does not take place.

piRNA pathway proteins have rapidly evolved

Although the piRNA pathway is highly conserved across the metazoan lineage, some of its effector proteins are encoded by genes bearing marks of positive selection (Simkin et al. 2013). The accumulated divergence between these proteins has been proposed to account for the TE silencing failure in *Hmr*-rescued interspecific hybrids (Kelleher et al. 2012). To elucidate the global failure hypothesis, we have aligned *D. buzzatii* and *D. koepferae* translated transcriptomes (see Methods) against each other and assessed their identity percentage distribution, with a resulting median identity of 97.2% (Figure 6).

We have then identified in *D. buzzatii* and *D. koepferae* translated transcriptomes a total of 30 protein-coding genes known to be involved in TE regulation (Yang & Pillai 2014) as reciprocal best BLAST hits of their *D. melanogaster* putative orthologs (their names and symbols are listed in Table 3). Alignments of all these genes between our parental species exhibit identity percentages lower than the median –their own median equals 92.5%– with the exception of the helicase Hel25E, whose sequence is identical in *D. buzzatii* and *D. koepferae* (Figure 6). Among the 10 most divergent proteins (identity ≤90%), we find factors involved in both piRNA biogenesis (*e.g.* zucchini, tejas) and TE silencing (*e.g.* Panoramix, maelstrom, Hen1 and qin). Thus, protein divergence between our studied species could cause hybrid incompatibilities in both biogenesis and function of piRNAs.

We have also examined the expression of these 30 protein-coding genes and revealed significant differences between our parental species for all of them, with the exception of *Hen1*, *Panoramix* (*Panx*) and *tejas* (*tej*, Table 3). The highest FC (log2FC=5.0) is attributed to *krimper* (*krimp*, more expressed in *D. buzzatii*), known to participate in the ping-pong amplification process (Sato et al. 2015; Webster et al. 2015). Moreover, the two main genes involved in secondary piRNA biogenesis, *Aubergine* (*Aub*) and *Argonaute3* (*Ago3*), are also more expressed in *D. buzzatii* (Table 3). Altogether, these results are consistent with the higher ping-pong fraction reported in this species (Figure 5B). Therefore, divergence in piRNA production between our parental species can be explained by the accumulated divergence in their piRNA pathway effector proteins as well as by the important differences in their expression levels.

When comparing hybrids to both parental species (Table 3), we observe significant underexpression of *Hen1* (involved in primary and secondary piRNA biogenesis) and *Sister of Yb* (*SoYb*, involved in primary piRNA biogenesis) in both F1 and BC1. On the other hand, significant overexpression of *Panx* (involved in transcriptional silencing) also occurs in both

hybrid generations. Those three genes are among the most divergent between parental species (identity≤90%, Figure 6) and their altered expression could also partially account for TE deregulation.

Interspecific hybridization has sex-biased effects on TE deregulation

An enhanced piRNA production may cause TE underexpression in hybrid testes

F1 testes present 256 differentially expressed TE families compared to *D. buzzatii* (more than any hybrid-parent comparison in ovaries, Figure 7A), and, as in ovaries, most of them are retrotransposons (Supplementary file S7). Although we cannot compare hybrids to both parental species, we observe that TE underexpression in hybrid testes prevails over their overexpression (Supplementary table S2), showing that TE deregulation exhibits sex-biased patterns.

Regarding piRNA populations, the global piRNA production seems to be enhanced in F1 hybrids compared to *D. buzzatii* (Figure 7B), and the ping-pong fraction is also increased (Figure 7C). Besides, there is a bias towards piRNA overexpression of TE families in hybrids: 130 TE families exhibit more piRNAs in hybrids than in *D. buzzatii*, whereas 87 families have lower piRNA levels in hybrids (considering ≥2-fold differences, Supplementary file S7). Therefore, in the case of males, the bias towards TE underexpression seems to be explained by a higher production of piRNAs.

TE expression and piRNA production are sex-biased

The described sex-biased TE deregulation patterns are consistent with the remarkable differences in TE expression observed between testes and ovaries. Our results show that opposite sex samples always present more differences than samples of the same sex (Supplementary table S2). In particular, testes tend to present higher TE expression than ovaries (Supplementary table S2): for instance, 303 TE families present differential

expression between ovaries and testes of *D. buzzatii*, of which 164 are more expressed in males than in females (Figure 7A). piRNA production also differs between sexes in *D. buzzatii*: testes exhibit lower global piRNA amounts (Figure 7B) and lower ping-pong signature levels than ovaries (Figure 7C). Accordingly, alignment rates of piRNAs to TEs are significantly higher in ovaries than in testes (Supplementary file S5, Student's t=-9.26, p=0.01586). Therefore, males tend to present higher TE expression and lower amounts of piRNAs than females.

Discussion

TE overexpression prevails over underexpression in *D. buzzatii-D. koepferae* hybrid ovaries (Tables 1, 2 and Supplementary table S2). This concurs with several studies focused on a single or few TEs, where higher transcription levels in hybrids than in parents were observed (Kawakami et al. 2011; Carnelossi et al. 2014; García Guerreiro 2015). At a whole-genome level, a few surveys also report cases of TE families underexpressed in hybrids, but these results are generally out of the main attention focus and consequently poorly discussed. For instance, in lake whitefish hybrids, approximately 38% of differentially expressed TEs are underexpressed in hybrids (Dion-Côté et al. 2014), a similar result to what we find in ovaries. Another well-studied case is that of hybrid sunflowers, where F1 hybrids present lower expression of the majority of TEs compared to parental species (Renaut et al. 2014). The presence of both overexpressed and underexpressed TEs suggests that hybrid TE deregulation is more complex than previously expected and may depend on the TE family.

Functional divergence between parental piRNA pathways can lead to hybrid incompatibilities

We demonstrate that TE families with important differences in their piRNA amounts between *D. buzzatii* and *D. koepferae* are not more commonly deregulated than families with similar levels (Figure 4). This shows that the maternal cytotype failure hypothesis cannot completely account for the observed pattern of TE deregulation, which is consistent with the similiarity of TE landscapes between our parental species (Supplementary figure S1). Thus, this explanation might be valid only for some particular TE families (Figure 4).

Sequence divergence between maternal piRNAs and paternal TE transcripts (and the reciprocal) could also lead to a decrease of silencing efficacy in hybrids, as suggested by piRNA alignment results on our TE library (Supplementary file S5). A genome-wide

comparison of sequences within a TE family between parental species cannot be performed because sequenced TEs in *D. koepferae* are scarce and its genome has not been sequenced yet. However, some TE families, such as *Helena*, have been shown to be highly conserved between these species (Romero-Soriano & García Guerreiro 2016). The presence of underexpressed TE families in hybrids also seems to rule out this explanation.

Therefore, our results point rather to the piRNA pathway global failure hypothesis, which states that accumulated divergence of piRNA pathway effector proteins is responsible for hybrid TE deregulation. In this way, we show that proteins involved in piRNA biogenesis and function are more divergent than expected between *D. buzzatii* and *D. koepferae* (Figure 6). Consistent with this observation, previous studies in other *Drosophila* species have demonstrated that some of these proteins are encoded by rapidly evolving genes with marks of adaptive selection (Simkin et al. 2013; Obbard et al. 2009). Furthermore, we find that almost all piRNA pathway genes present significant differences in expression between *D. buzzatii* and *D. koepferae* (Table 3). Such level of variability was also observed between different populations of a same species, *D. simulans* (Fablet et al. 2014).

D. koepferae seems to produce higher amounts of piRNAs compared to *D. buzzatii*, that exhibits higher levels of ping-pong signature (Figure 5). Those differences in global piRNA production strategies between parental species could be linked to the divergence and variability in expression between piRNA pathway genes. Indeed, the two main effectors of ping-pong amplification, *Aub* and *Ago3*, are more expressed in *D. buzzatii* than in *D. koepferae* (log2FC=2.62 and 0.80, Table 3), which is consistent with the important ping-pong fraction detected in this species. Furthermore, an excess of *Aub* expression relative to *Piwi* could lead to a decrease of piRNA production due to a less efficient phased piRNA biogenesis. After the cleavage of a piRNA cluster trancript by Ago3 in the ping pong cycle, the remnants of this transcript are loaded into Aub and processed to form the 3' end of an

antisense Aub-bound piRNA (Czech & Hannon 2016). The excised fragment of the piRNA cluster transcript is usually loaded into Piwi (and to a lesser extent, into Aub) and cut by Zucchini (Zuc) every 27-29 nucleotides, producing phased antisense piRNAs that allow sequence diversification (Han et al. 2015; Mohn et al. 2015). We can hypothesize that an excess of *Aub* expression leads to a more frequent loading of this protein for phased piRNA production; impairing the efficiency of phasing in *D. buzzatii*. This would lead to lower levels of piRNAs in *D. buzzatii*, that would mostly be produced by ping-pong amplification.

Contrary to *Aub*, *qin* is more expressed in *D. koepferae* than in *D. buzzatii* (log2FC=-1.30, Table 3), which can be at the origin of the observed lower amounts of antisense piRNAs in *D. buzzatii* (Supplementary file S5). Qin is known to enforce heterotypic ping-pong between Aub and Ago3 by preventing futile homotypic Aub:Aub cycles, which mainly produce sense piRNAs (Zhang et al. 2011). A recent study has demonstrated that homotypic Aub:Aub ping-pong also generates lower Piwi-bound antisense phased piRNAs, because qin ensures the correct loading of Piwi with antisense sequences (Wang et al. 2015). Therefore, a lower expression of *qin* (coupled with an excess of Aub) could lead to a less efficient production of antisense piRNAs (both secondary and phased) in *D. buzzatii* compared to *D. koepferae*. However, we must note that the remarkably higher expression levels of *krimper* in *D. buzzatii* (log2FC=5.0, Table 3) may diminish these effects, because krimper contributes to heterotypic ping-pong cycle formation by sequestering unloaded Ago3 proteins to prevent illegitimate access of other RNA sequences into them (Sato et al. 2015; Webster et al. 2015).

D. buzzatii and D. koepferae seem to present a functional divergence of the piRNA pathway, which could likely be at the origin of TE misregulation in hybrids. However, contrarily to the observed in *D. melanogaster-D. simulans* artificial hybrids, our hybrids do not exhibit deficient piRNA production (Kelleher et al. 2012). Indeed, global piRNA amounts in hybrids are higher than in *D. buzzatii* and resemble the amounts observed in *D. koepferae* (Figure 5B

and Supplementary file S5); and hybrid secondary piRNA biogenesis presents intermediate levels between parental species (Figure 5A). Thus, incompatibilities in our hybrids may entail piRNA-mediated silencing effectors rather than proteins involved in piRNA biogenesis, even though both kinds of protein are among those with the lowest identity percentages (Figure 6).

Misexpression of *SoYb*, *Hen1* and *Panoramix* can influence hybrid TE expression

Two of the piRNA pathway genes, *SoYb* and *Hen1*, are underexpressed in hybrids (Table 3). Hen1 is known to methylate piRNAs at their 3' ends in both follicle and germ cells (Horwich et al. 2007; Saito et al. 2007), but the impact of its mutation on TE expression may depend on the TE family. For instance, overexpression of *HeT-A* retrotransposon was observed in *Hen1* mutants due to a higher instability of piRNAs (Horwich et al. 2007), but other mutants exhibited an unchanged expression of retrotransposons (Saito et al. 2007). *SoYb* seems to be involved in primary piRNA biogenesis and has a partially redundant function with its paralog *BoYb* (Handler et al. 2011). Thus, even a complete gene loss of *SoYb* could be compensated by *BoYb* and would not lead to a widespread TE overexpression. Curiously, *BoYb* was underexpressed in *D. simulans-D. melanogaster* artificial hybrids (Kelleher et al. 2012). Although downregulation of *Hen1* and *SoYb* cannot explain the whole pattern of TE deregulation, we cannot dismiss it as a possible contributor to TE overexpression in some cases.

On the other hand, overexpression of *Panoramix*, known to be essential for TE transcriptional silencing (Yu et al. 2015; Czech et al. 2013; Handler et al. 2013; Sienski et al. 2015) may compensate silencing deficiencies (especially at a post-transcriptional level) and be at the origin of TE underexpression.

TE deregulation may involve other mechanisms

We have shown that TE deregulation in hybrid ovaries may be related to the piRNA pathway in terms of i) incompatibilities due to its divergence between parental species, ii) misregulation of some genes involved in TE silencing and iii) differences between parental piRNA pools (for a few TE families). However, changes in this pathway may not explain the whole set of alterations of TE expression observed in hybrids. Actually, an important fraction of overexpressed TE families does not present any associated piRNA (Figure 4B).

For instance, the endo-siRNA pathway is known to silence TEs in somatic and germinal tissues, with a partially redundant function with the piRNA pathway in gonads (Saito & Siomi 2010). Although our hybrids do not present lower global levels of 21 nucleotide reads than parental species (Figure 5A), we cannot completely reject the involvement of a putative endo-siRNA pathway dysfunction in TE deregulation, particularly for somatic elements. With our data, we cannot distinguish between somatic and germinal elements, and related bibliography in our species model is virtually nonexistent. However, the presence of *gypsy* elements among deregulated families (Tables 1 and 2) could indicate that some of them are indeed expressed in follicle somatic cells.

In wild wheat hybrids, two TE defence mechanisms have been proposed to be activated: deletion and methylation (Senerchia et al. 2015). In *Drosophila*, DNA methylation is not common, but internal or complete deletions of TE copies have been suggested to act as a TE prevention mechanism against genome invasions (Petrov & Hartl 1998; Romero-Soriano & García Guerreiro 2016; Lerat et al. 2011). In that case, suppression of active insertions could reduce the RNA amounts of some TE families, contributing to their underexpression. Furthermore, recombination between copies is known to control R1 elements expansion in *Drosophila*. These elements are specifically inserted in 28S rRNA genes and their copies are often deleted by recombination events (Eickbush & Eickbush 2014).

Finally, histone methylation marks linked with permissive or repressive chromatin states have frequently been associated with TE sequences and their surroundings (Klenov et al. 2007; Yasuhara & Wakimoto 2008; Riddle et al. 2011; Yin et al. 2011). We must note that this has been shown to be tightly connected with the piRNA pathway. For instance, expression of piRNA clusters depends (directly or indirectly) on methylation marks (Goriaux et al. 2014; Mohn et al. 2014; Rangan et al. 2011; Molla-Herman et al. 2015), and piRNA-mediated transcriptional silencing triggers the deposition of repressive H3K9me3 marks. However, other mechanisms (including endo-siRNAs) are also able to recruit this silencing machinery leading to heterochromatin formation. Failure in the deposition of histone modifications could hence result in abnormal TE expression.

TE deregulation across generations of hybridization

Interspecific gene flow between *D. buzzatii* and *D. koepferae* is a natural source of genetic diversity that can only be maintained through introgression of a parental genome in F1 females (F1 males are all sterile (Marin et al. 1993)). Therefore, the study of backcrossed hybrids delves into the understanding of the real impact of hybridization in nature. We show that differences in ovarian TE expression between hybrids and parents are concordant with the expected *D.buzzatii/D.koepferae* genome fraction at each generation: F1 is equally distant from both parental species, whereas BC1 drifts apart from *D. koepferae* (Figure 3A). Furthermore, the total amount of deregulated TE families is lower in BC1 (10.6% of the expressed TEs) than in F1 (15.2%): a generation of backcrossing seems to be sufficient to restore the regulatory mechanisms of some families, but not of the totality. A similar result was reported in inbred lines of *Oryza sativa* introgressed with genetic material from the wild species *Zizania latifolia*, where *copia* and *gypsy* retrotransposons were activated and then rapidly repressed within a few selfed generations (Liu & Wendel 2000). F1 and BC1 ovaries exhibit the lowest number of differentially expressed TEs within one-to-one sample

comparisons (Supplementary table S2) and present similar TE expression profiles (Figure 2B). This points to the hypothesis that more generations would be necessary to restore TE expression to the parental levels. Indeed, if TE activation in hybrids is caused by the failure of different epigenetic mechanisms (Michalak 2009), these are expected to be mitigated after several backcrosses thanks to the dominance of one of the parental genomes. In agreement to this hypothesis, we showed in a recent study that TE activation causes a genome expansion in *D. buzzatii-D. koepferae* hybrid females, but the C-value decreases after the first backcross (Romero-Soriano et al. 2016).

Tendency to TE repression in hybrid testes demonstrates that TE regulation is sex-biased

We show that TE expression presents different patterns between ovaries and testes, both at the quantitative and qualitative levels (Figure 2). Other studies have reported tissue-specific expression of transposons between male and female gonads. For instance, in *D. simulans* and *D. melanogaster*, transcripts of *412* are only found in testes (Borie et al. 2002), *I-like* elements are more expressed in testes than in ovaries of *D. mojavensis* and *D. arizonae* (Carnelossi et al. 2014), as well as are *Osvaldo* and *Helena* in *D. buzzatii* and *D. koepferae* (García Guerreiro 2015; Romero-Soriano & García Guerreiro 2016). All these studies show higher transcript abundances in male gonads, which is consistent with the bias we observe towards testes overexpression compared to ovaries (Supplementary table S2).

These findings point out a differential TE regulation between male and female gonads, which was previously suggested by studies in *Drosophila* testes demonstrating that male piRNA biogenesis is not always performed by the same mechanisms as in ovaries (Nagao et al. 2010; Siomi et al. 2010). Concordantly, we observe that testes have lower piRNA amounts and a less efficient ping-pong cycle than ovaries (Figure 7). It has indeed been shown that piRNAs

in testes are not only involved in TE repression but also in gene silencing, particularly of *Stellate* and *vasa* (Nishida et al. 2007).

Our results on TE deregulation in hybrids fully support the idea of sex-specificity in TE silencing. Contrarily to ovaries, hybrid testes exhibit a bias towards TE underexpression compared to *D. buzzatii* (Supplementary table S2). Accordingly, the retrotransposon *Helena* was shown to exhibit lower transcript abundances in F1 testes than in *D. buzzatii* and *D. koepferae* (Romero-Soriano & García Guerreiro 2016), as was the case for most TE families in a transcriptomic study in F1 sunflower hybrids (Renaut et al. 2014). Although two other studies in *Drosophila* hybrids, focused on individual TEs, displayed the opposite effect (García Guerreiro 2015; Carnelossi et al. 2014), we consider that disparity between specific studies fits in our global results.

TE underexpression prevalence in our hybrid testes can be explained by an increase of piRNA production and ping-pong signal in F1 testes (Figure 7B and C). Thus, activation of piRNA biogenesis, especially through the ping-pong cycle, seems to be responsible for TE repression in testes. Consistent with this tight repression of TE activity in males, the genome size increase observed in *D. buzzatii-D. koepferae* hybrids occurs only in females, whereas the hybridization impact in male genome size is undetectable (Romero-Soriano et al. 2016).

Conclusions

We suggest that TE deregulation in ovaries of *D. buzzatii-D. koepferae* hybrids might be the result of several interacting phenomena: a partial failure of the piRNA pathway due to a functional divergence between parental species, misexpression of some piRNA pathway genes, and differences in the amounts of TE-specific piRNAs between maternal cytoplasms (for some TE families). Furthermore, we cannot discard that other TE repression mechanisms might partially account for the observed set of deregulations. For instance, the endo-siRNA pahway function could also be affected, deletions could play a role in TE underexpession, and histone post-translational modifications may alter the chromatin state pattern of the hybrid genome and cause either overexpression or underexpression (depending on the TE insertion). The study of these mechanisms would be an interesting focus for future investigations, as it could shed light on other causes of hybrid TE deregulation.

On the other hand, comparison of ovaries and testes show that TE regulation is sex-biased. Surprisingly, piRNA biogenesis is enhanced in hybrid testes, which underlines that hybridization is a genomic stress that can activate response pathways to counteract TE deregulation. Further work in testes needs to be performed to elucidate the observed differences in TE silencing, which could be crucial to understand the molecular basis of hybrid breakdown and sterility.

Acknowledgements

The authors wish to thank Nuria Rius and Alfredo Ruiz for providing advanced access to the *D. buzzatii* genome and its TE list; Xavier Grau-Bové for his useful help with orthology methods; Esteban Hasson and Diego Nicolás de Panis for sharing sequencing data of *D. koepferae* genome; Clément Goubert for his valuable help with dnaPipeTE; and INRA – UMR AGAP (France) and Fasteris SA (Switzerland) for RNA sequencing services.

References

Antoniewski C. 2014. Computing siRNA and piRNA Overlap Signature. In: Animal Endo-SiRNAs: Methods and Protocols. Werner, A, editor. Springer New York pp. 135–146. doi: 10.1007/978-1-4939-0931-5_12.

Barbash DA. 2010. Ninety years of *Drosophila* melanogaster hybrids. Genetics. 186:1–8. doi: 10.1534/genetics.110.121459.

Bauer S, Grossmann S, Vingron M, Robinson PN. 2008. Ontologizer 2.0 - A multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics. 24:1650–1651. doi: 10.1093/bioinformatics/btn250.

Benson G. 1999. Tandem Repeats Finder: a program to analyse DNA sequences. Nucleic Acids Res. 27:573–578.

Borie N, Maisonhaute C, Sarrazin S, Loevenbruck C, Biémont C. 2002. Tissue-specificity of 412 retrotransposon expression in *Drosophila* simulans and D. melanogaster. Heredity. 89:247–52. doi: 10.1038/sj.hdy.6800135.

Brennecke J et al. 2008. An Epigenetic Role for Maternally Inherited piRNAs in Transposon Silencing. Science. 322:1387–1392.

Brennecke J et al. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. Cell. 128:1089–103. doi: 10.1016/j.cell.2007.01.043.

Brennecke J, Senti K-A. 2010. The piRNA pathway : a fly's perspective on the guardian of the genome. Trends Genet. 26:499–509. doi: 10.1016/j.tig.2010.08.007.

Carnelossi EAG et al. 2014. Specific activation of an I-like element in *Drosophila* interspecific hybrids. Genome Biol. Evol. 6:1806–17. doi: 10.1093/gbe/evu141.

Cunningham F et al. 2015. Ensembl 2015. Nucleic Acids Res. 43:D662–D669. doi: 10.1093/nar/gku1010.

Cutter AD. 2008. Divergence times in Caenorhabditis and *Drosophila* inferred from direct estimates of the neutral mutation rate. Mol. Biol. Evol. 25:778–786. doi: 10.1093/molbev/msn024.

Czech B, Hannon GJ. 2016. One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. Trends Biochem. Sci. 41:324–337. doi: 10.1016/j.tibs.2015.12.008.

Czech B, Preall JB, McGinn J, Hannon GJ. 2013. A transcriptome-wide RNAi screen in the

Drosophila ovary reveals factors of the germline piRNA pathway. Mol. Cell. 50:749–761. doi: 10.1016/j.molcel.2013.04.007.

Dion-Côté A-M, Renaut S, Normandeau E, Bernatchez L. 2014. RNA-seq Reveals Transcriptomic Shock Involving Transposable Elements Reactivation in Hybrids of Young Lake Whitefish Species. Mol. Biol. Evol. 31:1188–1199. doi: 10.1093/molbev/msu069.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature. 450:203–18. doi: 10.1038/nature06341.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797. doi: 10.1093/nar/gkh340.

Eickbush TH, Burke WD, Eickbush DG, Lathe WC. 1997. Evolution of R1 and R2 in the rDNA units of the genus *Drosophila*. Genetica. 100:49–61.

Eickbush TH, Eickbush DG. 2014. Integration, Regulation, and Long-Term Stability of R2 Retrotransposons. Microbiol. Spectr. 3:MDNA3–0011–2014. doi: 10.1128/microbiolspec.MDNA3-0011.

Evgen'ev MB, Yenikolopov GN, Peunova NI, Ilyin Y V. 1982. Transposition of Mobile Genetic Elements in Interspecific Hybrids of *Drosophila*. Chromosoma. 85:375–386.

Fablet M, Akkouche A, Braman V, Vieira C. 2014. Variable expression levels detected in the *Drosophila* effectors of piRNA biogenesis. Gene. 537:149–53. doi: 10.1016/j.gene.2013.11.095.

Fontdevila A. 2005. Hybrid genome evolution by transposition. Cytogenet. Genome Res. 110:49–55. doi: 10.1159/000084937.

Franco FF, Silva-Bernardi ECC, Sene FM, Hasson ER, Manfrin MH. 2010. Intra- and interspecific divergence in the nuclear sequences of the clock gene period in species of the *Drosophila buzzatii* cluster. J. Zool. Syst. Evol. Res. 48:322–331. doi: 10.1111/j.1439-0469.2010.00564.x.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics. 28:3150–3152. doi: 10.1093/bioinformatics/bts565.

García Guerreiro M. 2012. What makes transposable elements move in the *Drosophila* genome? Heredity. 108:461–468. doi: 10.1038/hdy.2011.89.

García Guerreiro MP. 2015. Changes of Osvaldo expression patterns in germline of male hybrids between the species *Drosophila buzzatii* and *Drosophila koepferae*. Mol. Genet. Genomics. 290:1471–1483. doi: 10.1007/s00438-015-1012-z.

García Guerreiro MP. 2014. Interspecific hybridization as a genomic stressor inducing mobilization of transposable elements in *Drosophila*. Mob. Genet. Elements. 4:e34394.

Ghildiyal M et al. 2008. Endogenous siRNAs derived from transposons and mRNAs in *Drosophila* somatic cells. Science. 320:1077–81. doi: 10.1126/science.1157396.

Gomez GA, Hasson E. 2003. Transpecific Polymorphisms in an Inversion Linked Esterase Locus in *Drosophila buzzatii*. Mol. Biol. Evol. 20:410–423. doi: 10.1093/molbev/msg051.

Goriaux C, Desset S, Renaud Y, Vaury C, Brasset E. 2014. Transcriptional properties and splicing of the flamenco piRNA cluster. EMBO Rep. 15:411–418. doi: 10.1002/embr.201337898.

Goubert C et al. 2015. De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). Genome Biol. Evol. 7:1192–1205. doi: 10.1093/gbe/evv050.

Grabherr MG et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–52. doi: 10.1038/nbt.1883.

Grentzinger T et al. 2013. A user-friendly chromatographic method to purify small regulatory RNAs. Methods. 67:91–101. doi: 10.1016/j.ymeth.2013.05.011.

Grentzinger T et al. 2012. piRNA-mediated transgenerational inheritance of an acquired trait piRNA-mediated transgenerational inheritance of an acquired trait. Genome Res. 22:1877–1888. doi: 10.1101/gr.136614.111.

Guillén Y et al. 2015. Genomics of Ecological Adaptation in Cactophilic *Drosophila*. Genome Biol. Evol. 7:349–366. doi: 10.1093/gbe/evu291.

Gunawardane LS et al. 2007. A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in *Drosophila*. Science. 315:1587–1590. doi: 10.1126/science.1140494.

Haas BJ et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8:1494–512. doi: 10.1038/nprot.2013.084.

Han BW, Wang W, Li C, Weng Z, Zamore PD. 2015. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. Science. 348:817–821.

Handler D et al. 2011. A systematic analysis of *Drosophila* TUDOR domain-containing proteins identifies Vreteno and the Tdrd12 family as essential primary piRNA pathway factors. EMBO J. 30:3977–3993. doi: 10.1038/emboj.2011.308.

Handler D et al. 2013. The genetic makeup of the *Drosophila* piRNA pathway. Mol. Cell. 50:762–777. doi: 10.1016/j.molcel.2013.04.031.

Hedges DJ, Deininger PL. 2007. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. Mutat. Res. 616:46–59. doi: 10.1016/j.mrfmmm.2006.11.021.

Horwich MD et al. 2007. The *Drosophila* RNA Methyltransferase, DmHen1, Modifies Germline piRNAs and Single-Stranded siRNAs in RISC. Curr. Biol. 17:1265–1272. doi: 10.1016/j.cub.2007.06.030.

Huerta-Cepas J et al. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 44:D286–93. doi: 10.1093/nar/gkv1248.

Iwasaki YW, Siomi MC, Siomi H. 2015. PIWI-Interacting RNA: Its Biogenesis and Functions. Annu. Rev. Biochem. 84:405–33. doi: 10.1146/annurev-biochem-060614-034258.

Jurka J et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110:462–467. doi: 10.1159/000084979.

Kawakami T, Dhakal P, Katterhenry AN, Heatherington CA, Ungerer MC. 2011. Transposable element proliferation and genome expansion are rare in contemporary sunflower hybrid populations despite widespread transcriptional activity of LTR retrotransposons. Genome Biol. Evol. 3:156–67. doi: 10.1093/gbe/evr005.

Keightley PD, Ness RW, Halligan DL, Haddrill PR. 2014. Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila* melanogaster full-sib family. Genetics. 196:313–320. doi: 10.1534/genetics.113.158758.

Kelleher ES, Edelman NB, Barbash DA. 2012. *Drosophila* Interspecific Hybrids Phenocopy piRNA-Pathway Mutants. PLoS Biol. 10:e1001428. doi: 10.1371/journal.pbio.1001428.

Kent WJ. 2002. BLAT---The BLAST-Like Alignment Tool. Genome Res. 12:656–664. doi:

10.1101/gr.229202.

Kidwell MG, Kidwell JF, Sved JA. 1977. Hybrid dysgenesis in *Drosophila* melanogaster: a syndrome of aberrant traits including mutation, sterility and male recombination. Genetics. 86:813–833.

Klattenhoff C, Theurkauf W. 2008. Biogenesis and germline functions of piRNAs. Development. 135:3–9. doi: 10.1242/dev.006486.

Klenov MS et al. 2007. Repeat-associated siRNAs cause chromatin silencing of retrotransposons in the *Drosophila* melanogaster germline. Nucleic Acids Res. 35:5430–5438. doi: 10.1093/nar/gkm576.

Kojima KK, Fujiwara H. 2003. Evolution of target specificity in R1 clade non-LTR retrotransposons. Mol. Biol. Evol. 20:351–361. doi: 10.1093/molbev/msg031.

Laayouni H, Hasson E, Santos M, Fontdevila A. 2003. The evolutionary history of *Drosophila buzzatii*. XXXV. Inversion polymorphism and nucleotide variability in different regions of the second chromosome. Mol. Biol. Evol. 20:931–944. doi: 10.1093/molbev/msg099.

Labrador M, Farré M, Utzet F, Fontdevila A. 1999. Interspecific hybridization increases transposition rates of Osvaldo. Mol. Biol. Evol. 16:931–7.

Lachaise D, Silvain J-F. 2004. How two Afrotropical endemics made two cosmopolitan human commensals: The *Drosophila* melanogaster-D. simulans palaeogeographic riddle. Genetica. 120:17–39. doi: 10.1023/B:GENE.0000017627.27537.ef.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 9:357–359. doi: 10.1038/nmeth.1923.

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. doi: 10.1186/gb-2009-10-3-r25.

Lerat E, Burlet N, Biémont C, Vieira C. 2011. Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. Gene. 473:100–109. doi: 10.1016/j.gene.2010.11.009.

Liu B, Wendel JF. 2000. Retrotransposon activation followed by rapid repression in introgressed rice plants. Genome. 43:874–880.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for

RNA-seq data with DESeq2. Genome Biol. 15:550. doi: 10.1186/s13059-014-0550-8.

Marín I, Fontdevila A. 1998. Stable *Drosophila buzzatii* - *Drosophila koepferae* Hybrids. J. Hered. 89:336–339.

Marin I, Ruiz A, Pla C, Fontdevila A. 1993. Reproductive Relationships among Ten Species of the *Drosophila* repleta Group from South America and the West Indies. Evolution (N. Y). 47:1616–1624. doi: 10.2307/2410173.

McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 32:W20–5. doi: 10.1093/nar/gkh435.

Metcalfe CJ et al. 2007. Genomic instability within centromeres of interspecific marsupial hybrids. Genetics. 177:2507–17. doi: 10.1534/genetics.107.082313.

Michalak P. 2009. Epigenetic, transposon and small RNA determinants of hybrid dysfunctions. Heredity. 102:45–50.

Modolo L, Lerat E. 2015. UrQt: an efficient software for the Unsupervised Quality trimming of NGS data. BMC Bioinformatics. 16:137. doi: 10.1186/s12859-015-0546-8.

Modolo L, Picard F, Lerat E. 2014. A new genome-wide method to track horizontally transferred sequences: Application to *Drosophila*. Genome Biol. Evol. 6:416–432. doi: 10.1093/gbe/evu026.

Mohn F, Handler D, Brennecke J. 2015. piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. Science. 348:812–817. doi: 10.1126/science.aaa1039.

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. 157:1364–1379. doi: 10.1016/j.cell.2014.04.031.

Molla-Herman A, Vallés AM, Ganem-Elbaz C, Antoniewski C, Huynh J-R. 2015. tRNA processing defects induce replication stress and Chk2-dependent disruption of piRNA transcription. EMBO J. 34:3009–27. doi: 10.15252/embj.201591006.

Nagao A, Mituyama T, Huang H, Chen D, Siomi MC. 2010. Biogenesis pathways of piRNAs loaded onto AGO3 in the *Drosophila* testis. RNA. 16:2503–2515. doi: 10.1261/rna.2270710.

Nishida KM, Saito K, Mori T. 2007. Gene silencing mechanisms mediated by Aubergine − piRNA complexes in *Drosophila* male gonad. RNA. 13:1911–1922. doi: 10.1261/rna.744307.

O'Neill RJW, O'Neill MJ, Marshall Graves JA. 1998. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. Nature. 393:68–73.

Obbard DJ, Gordon KHJ, Buck AH, Jiggins FM. 2009. The evolution of RNAi as a defence against viruses and transposable elements. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 364:99–115. doi: 10.1098/rstb.2008.0168.

Oliveira DCSG et al. 2012. Monophyly, divergence times, and evolution of host plant use inferred from a revised phylogeny of the *Drosophila* repleta species group. Mol. Phylogenet. Evol. 64:533–544. doi: 10.1016/j.ympev.2012.05.012.

Petrov DA, Hartl DL. 1998. High rate of DNA loss in the *Drosophila* melanogaster and *Drosophila* virilis species groups. Mol. Biol. Evol. 15:293–302. doi: 10.1093/oxfordjournals.molbev.a025926.

Picard G. 1976. Non mendelian female sterility in *Drosophila* melanogaster: hereditary transmission of I factor. Genetics. 83:107–123. doi: 10.1007/BF00123290.

Piccinali R, Aguadé M, Hasson E. 2004. Comparative molecular population genetics of the Xdh locus in the cactophilic sibling species *Drosophila buzzatii* and D. *koepferae*. Mol. Biol. Evol. 21:141–52. doi: 10.1093/molbev/msh006.

Punta M et al. 2012. The Pfam protein families databases. Nucleic Acids Res. 40:D290–D301. doi: 10.1093/nar/gkp985.

Rangan P et al. 2011. PiRNA production requires heterochromatin formation in *Drosophila*. Curr. Biol. 21:1373–1379. doi: 10.1016/j.cub.2011.06.057.

Renaut S, Rowe HC, Ungerer MC, Rieseberg LH. 2014. Genomics of homoploid hybrid speciation: diversity and transcriptional activity of long terminal repeat retrotransposons in hybrid sunflowers. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 369:20130345. doi: 10.1098/rstb.2013.0345.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16:276–277. doi: 10.1016/j.cocis.2008.07.002.

Riddle NC et al. 2011. Plasticity in patterns of histone modifications and chromosomal proteins in *Drosophila* heterochromatin. Genome Res. 21:147–163. doi: 10.1101/gr.110098.110.

Romero-Soriano V et al. 2016. *Drosophila* females undergo genome expansion after interspecific hybridization. Genome Biol. Evol. 8:556–561. doi: 10.1093/gbe/evw024.

Romero-Soriano V, García Guerreiro MP. 2016. Expression of the Retrotransposon Helena Reveals a Complex Pattern of TE Deregulation in *Drosophila* Hybrids. PLoS One. 11:e0147903. doi: 10.1371/journal.pone.0147903.

Rozhkov N V, Hammell M, Hannon GJ. 2013. Multiple roles for Piwi in silencing *Drosophila* transposons. Genes Dev. 27:400–412. doi: 10.1101/gad.209767.112.

Russo CAM, Takezaki N, Nei M. 1995. Molecular phylogeny and divergence times of drosophilid species. Mol. Biol. Evol. 12:391–404.

Saito K et al. 2007. Pimet, the *Drosophila* homolog of HEN1, mediates 2'-O-methylation of Piwi-interacting RNAs at their 3' ends. Genes Dev. 21:1603–1608. doi: 10.1101/gad.1563607.the.

Saito K, Siomi MC. 2010. Small RNA-mediated quiescence of transposable elements in animals. Dev. Cell. 19:687–97. doi: 10.1016/j.devcel.2010.10.011.

Sato K et al. 2015. Krimper Enforces an Antisense Bias on piRNA Pools by Binding AGO3 in the *Drosophila* Germline. Mol. Cell. 59:553–563. doi: 10.1016/j.molcel.2015.06.024.

Senerchia N, Parisod C, Parisod C. 2015. Genome reorganization in F1 hybrids uncovers the role of retrotransposons in reproductive isolation. Proc. R. Soc. B Biol. Sci. 282:20142874.

Sienski G et al. 2015. Silencio / CG9754 connects the Piwi – piRNA complex to the cellular heterochromatin machinery. Genes Dev. 29:1–14. doi: 10.1101/gad.271908.115.

Simkin A, Wong A, Poh Y-P, Theurkauf WE, Jensen JD. 2013. Recurrent and Recent Selective Sweeps in the piRNA Pathway. Evolution (N. Y). 67:1081–1090. doi: 10.1111/evo.12011.

Siomi MC, Miyoshi T, Siomi H. 2010. piRNA-mediated silencing in *Drosophila* germlines. Semin. Cell Dev. Biol. 21:754–9. doi: 10.1016/j.semcdb.2010.01.011.

The Gene Ontology Consortium. 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25:25–29. doi: 10.1038/75556.

Ungerer MC, Strakosh SC, Zhen Y. 2006. Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. Curr. Biol. 16:R872–3. doi: 10.1016/j.cub.2006.09.020.

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. doi: 10.1371/journal.pone.0088992.

Vela D, García Guerreiro MP, Fontdevila A. 2011. Adaptation of the AFLP technique as a new tool to detect genetic instability and transposition in interspecific hybrids. Biotechniques. 50:247–50. doi: 10.2144/000113655.

Wang N et al. 2010. Transpositional reactivation of the Dart transposon family in rice lines derived from introgressive hybridization with Zizania latifolia. BMC Plant Biol. 10:190. doi: 10.1186/1471-2229-10-190.

Wang W et al. 2015. Slicing and Binding by Ago3 or Aub Trigger Piwi-Bound piRNA Production by Distinct Mechanisms. Mol. Cell. 59:819–830. doi: 10.1016/j.molcel.2015.08.007.

Webster A et al. 2015. Aub and Ago3 Are Recruited to Nuage through Two Mechanisms to Form a Ping-Pong Complex Assembled by Krimper. Mol. Cell. 59:564–575. doi: 10.1016/j.molcel.2015.07.017.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591. doi: 10.1093/molbev/msm088.

Yang Z, Pillai RS. 2014. Fly piRNA biogenesis: tap dancing with Tej. BMC Biol. 12:77. doi: 10.1186/s12915-014-0077-1.

Yasuhara JC, Wakimoto BT. 2008. Molecular landscape of modified histones in *Drosophila* heterochromatic genes and euchromatin-heterochromatin transition zones. PLoS Genet. 4:0159–0172. doi: 10.1371/journal.pgen.0040016.

Yin H, Sweeney S, Raha D, Snyder M, Lin H. 2011. A High-Resolution Whole-Genome map of key chromatin modifications in the adult *Drosophila* melanogaster. PLoS Genet. 7:e1002380. doi: 10.1371/journal.pgen.1002380.

Yu Y et al. 2015. Panoramix enforces piRNA-dependent cotranscriptional silencing. Science. 350:339–342. doi: 10.1126/science.aab0700.

Zhang Z et al. 2011. Heterotypic piRNA Ping-Pong requires qin, a protein with both E3 ligase and Tudor domains. Mol. Cell. 44:572–84. doi: 10.1016/j.molcel.2011.10.011.

Figure captions

Figure 1. Crosses diagram. (A) is the first interspecific cross between *D. koepferae* (yellow) females and *D. buzzatii* (blue) males, and **(B)** is the backcross between F1 hybrid (green) females and *D. buzzatii* (blue) males, that gives rise to BC1 (turquoise). Colours have been assigned according to the *D. buzzatii/D.koepferae* genome content: yellow for *D. koepferae*, blue for *D. buzzatii*, green for F1 hybrids and turquoise for BC1 hybrids. Samples marked with a white background rectangle have not been sequenced.

Figure 2. Transposable element expression summary. Dbu= *D. buzzatii*; Dko= *D. koepferae*; ♂♂= testes; ♀♀= ovaries. **(A)** Mean proportion of reads aligning to the TE library. Bars represent standard deviation between replicates. ** Student's t=4.09, p=0.0035. **(B)** TE expression profiles following Repbase classification (Jurka et al. 2005): LTR and LINE (class I), DNA and RC/*Helitron* (class II), Unknown (unclassified). LTR= elements with Long Terminal Repeats; LINE= Long Interspersed Nuclear Element; RC= Rolling Circle elements (or *Helitrons*).

Figure 3. TE differential expression analyses in ovaries. (A) Differentially expressed TE families in hybrids compared separately to *D. buzzatii* (Dbu) and *D. koepferae* (Dko). The total number of differentially expressed TE families of each comparison is written in parenthesis. FC= fold change (hybrid *vs.* parent). **(B)** Expression of TE families in *D. koepferae* vs. *D. buzzatii*. In colour, deregulated TE families in hybrids (compared to both parental species). Dot lines represent 2-fold changes between parental expression and the solid line represents the same amount of expression between Dbu and Dko. Names of those TE families with differences of expression higher than 2-fold between parental species are indicated.

Figure 4. Parental piRNA populations and TE deregulation in ovaries. (A) Expression of TE-associated piRNA populations in *D. koepferae* (Dko) *vs. D. buzzatii* (Dbu). Dot lines represent 2-fold changes between parental piRNA amounts and the solid line represents the same piRNA levels between Dbu and Dko. Underlined TE names are examples of families that may be deregulated due to the maternal cytotype hypothesis (underexpressed with more piRNAs in *D. koepferae*, overexpressed with more piRNAs in *D. buzzatii*). Names of deregulated TE families with unexpected differences in piRNA amounts (underexpressed with more piRNAs in *D. buzzatii*, overexpressed with more piRNAs in *D. koepferae*) are also indicated, with an arrow in some cases. **(B)** Proportion of deregulated TE families of different categories, classified according to differences (of at least 2-fold) between parental piRNA populations: (i) more piRNAs in *D. buzzatii*, (ii) not differentially abundant between parental species, (iii) more piRNAs in *D. koepferae*, (iv) absence of piRNAs in both species.

Figure 5. Characterization of piRNA populations in parental and hybrid ovaries. Dbu= *D. buzzatii*; Dko= *D. koepferae*; ♀♀= ovaries. **(A)** Read length distribution of ovarian small RNAs. The vertical dot line separates miRNAs and siRNAs (left) from piRNAs (right). **(B)** piRNA ping-pong fraction for each TE family (grey lines) and for the whole piRNA population (upper number). Only families with detectable ping-pong signal (>0) for at least one ovarian sample are represented.

Figure 6. Distribution of identity percentages between *D. buzzatii* and *D. koepferae* proteomes (see Methods). A total of 30 proteins involved in the piRNA pathway were identified as reciprocal best BLAST hits of their *D. melanogaster* orthologs (represented by vertical bars, their identity in parenthesis). For Zucchini, four sequences were recognized as putative paralogs and named zucchini-A, B, C and D (only zucchini-A, B and C are shown because zucchini-D was only identified in *D. buzzatii*). At least in two other species of the

genus *Drosophila*, *D. melanogaster* and *D. grimshawi*, paralogs of Zucchini have been identified (*Drosophila* 12 Genomes Consortium 2007).

Figure 7. Differential expression analyses in testes. Dbu= *D. buzzatii;* ♂♂= testes; ♀♀= ovaries. **(A)** Differentially expressed TE families between F1 testes and Dbu (left) and between sexes of *D. buzzatii* (right). The total number of significant differences of each comparison is written in parenthesis. FC= fold change. **(B)** Read length distribution of *D. buzzatii* (testes and ovaries) and F1 testes small RNAs. The vertical dot line separates miRNAs and siRNAs (left) from piRNAs (right). **(C)** piRNA ping-pong fraction for each TE family (grey lines) and for the whole piRNA population (upper number). Only families with detectable ping-pong signal (>0) for at least one sample are represented.

Supporting information captions

Supplementary figure S1. *D. buzzatii* and *D. koepferae* present highly similar repeatomes. (A) TE abundance in parental species genome. **(B)** TE landscapes of our parental species: genomic reads are classified according to their identity against the TE contig assembled with dnaPipeTE.

Supplementary figure S2: Ping-pong fraction of ovarian piRNA populations associated to deregulated TE families. (A) Overexpressed in F1. **(B)** Underexpressed in F1. **(C)** Overexpressed in BC1. **(D)** Underexpressed in BC1.

Supplementary file S1: RNA-seq statistics summary. **(A)** Number of reads at each analysis step. **(B)** Raw read count per TE family after alignment to the TE library. **(C)** Read count per TE family after normalization by DESeq2.

Supplementary file S2: Deregulated genes in ovaries. FC= Fold Change; BH= Bonferroni-Hochberg**. (A)** Overexpressed genes in F1. **(B)** Overexpressed genes in BC1. **(C)** Underexpressed genes in F1. **(D)** Underexpressed genes in BC1.

Supplementary file S3: Summary of codeml results. Rate of substitution per non-synonymous site (dN) and per synonymous site (dS) for each *D. buzzati-D. koepferaee* contig pair.

Supplementary file S4: Summary of dnaPipeTE results. Read count and proportion (%) of each class of repetitive sequences for *D. buzzatii* and *D. koepferae* genomic reads.

Supplementary file S5: small RNA population sequencing statistics summary. (A) Number of reads at each analysis step. **(B)** Raw piRNA read count per TE family after alignment to the TE library. **(C)** piRNA read count per TE family after normalization by DESeq2.

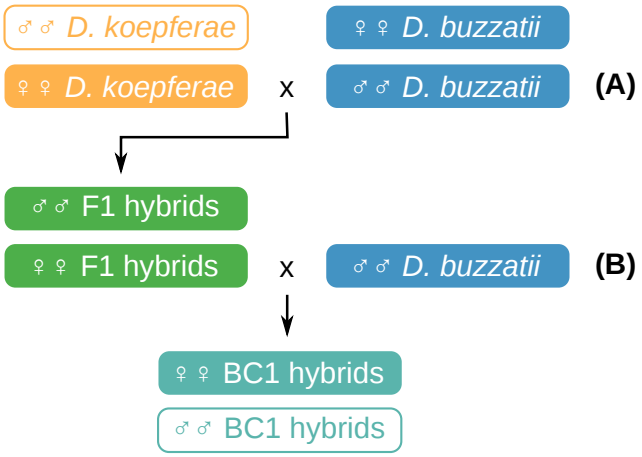
Supplementary file S6: TE families with notable differences (≥2-fold) in their piRNA populations in hybrid ovaries (F1 or BC1) compared to both parental species. FC= Fold Change. **(A)** Lower piRNA levels in parents. **(B)** Lower piRNA levels in hybrids.

Supplementary file S7: Differential expression of TEs in F1 testes compared to *D. buzzatii*. FC= Fold Change; BH= Bonferroni-Hochberg**. (A)** Overexpressed TE families in F1. **(B)** Underexpressed TE families in F1. **(C)** TE families with lower piRNA abundance in F1. **(D)** TE families with higher piRNA abundance in F1.

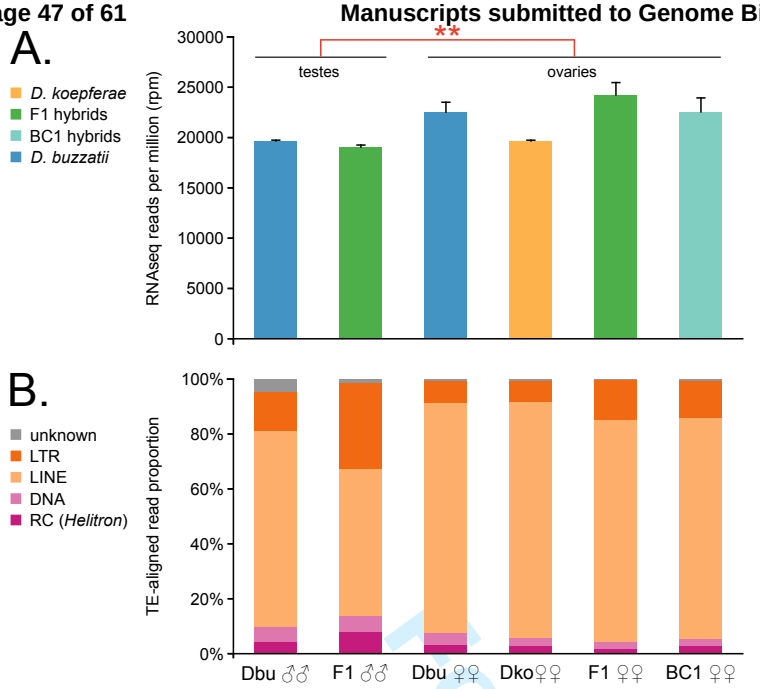
Supplementary table S1. Summary of assemblies and annotation. NA= not annotated. [a] clustering step with CD-HIT.

Supplementary table S2: Differential expression summary. Dbu= *D. buzzatii*, Dko= *D. koepferae*. Above the main diagonal (grey), number of TE families with significant differential expression for each comparison. In parenthesis, fraction (%) of differentially expressed TE families of *column* sample showing overexpression (green) or underexpression (red) compared to the sample in *row*. Below the main diagonal, fraction of the differentially expressed families which present 1.5 fold or higher differences.

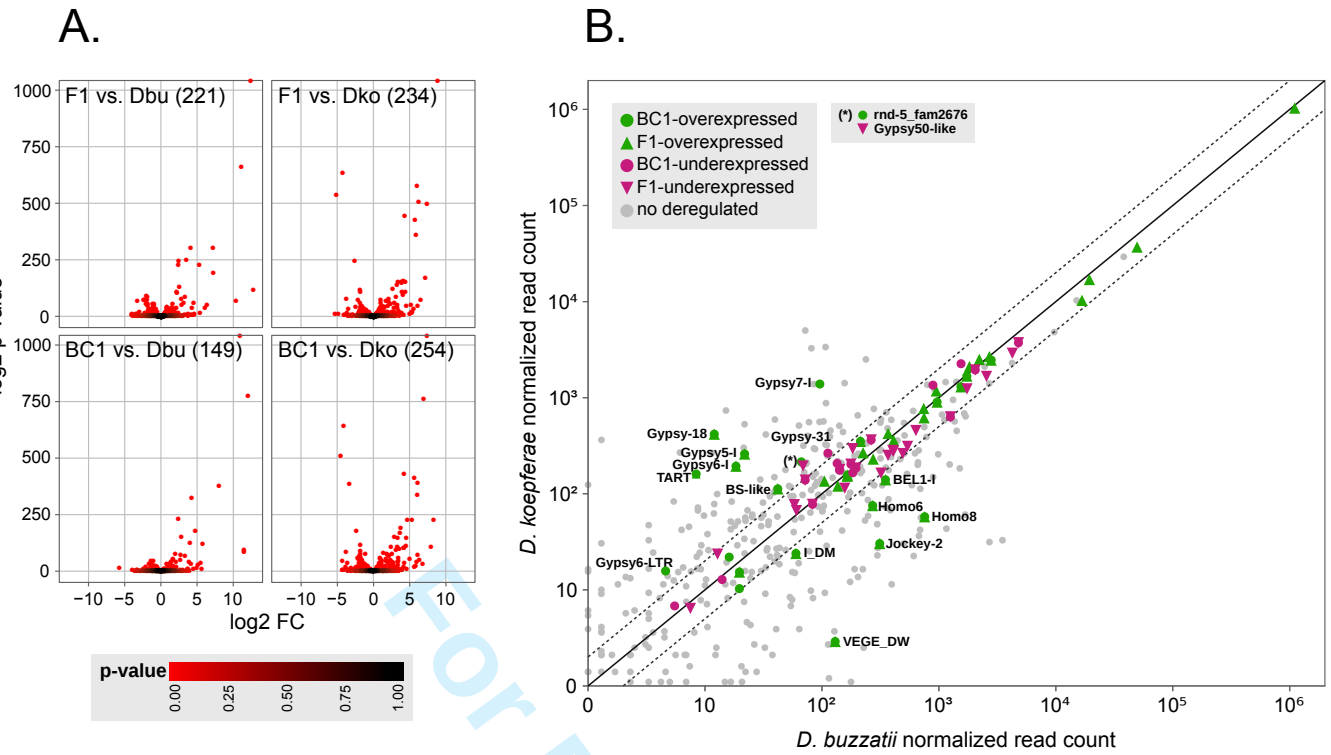
Supplementary table S3. Gene Ontology terms with significant enrichment in overexpressed and underexpressed genes of hybrid ovaries. Only GO terms common in F1 and BC1 are shown.



For Peer Review

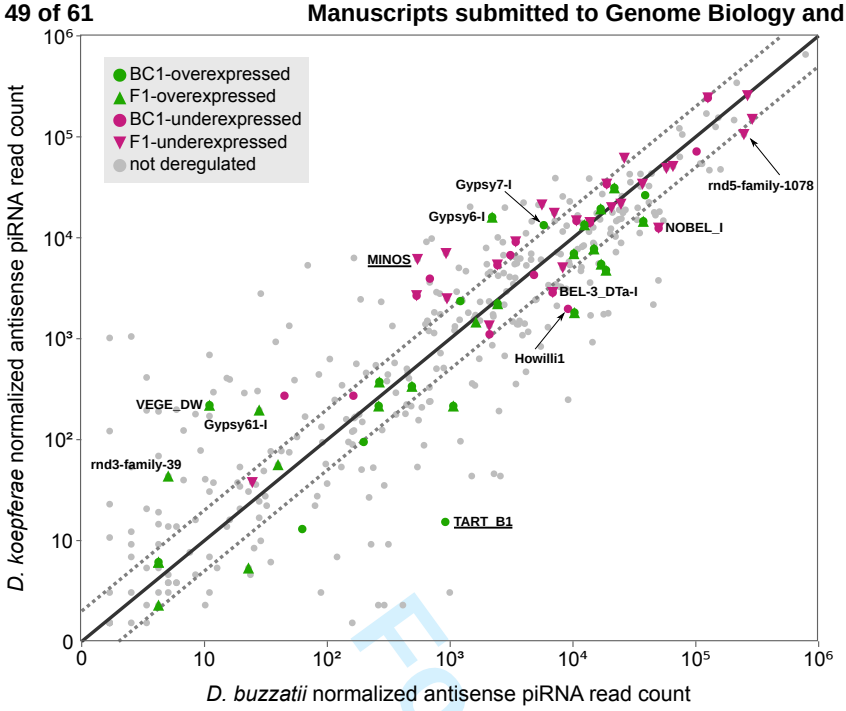


For Peer Review

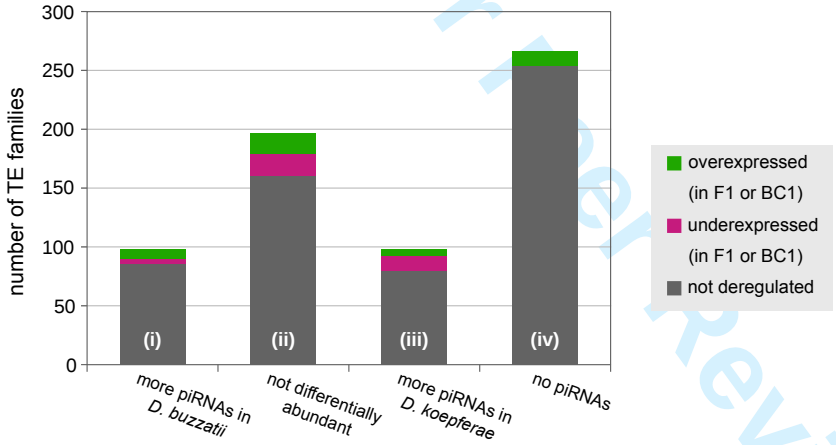


For Peer Review

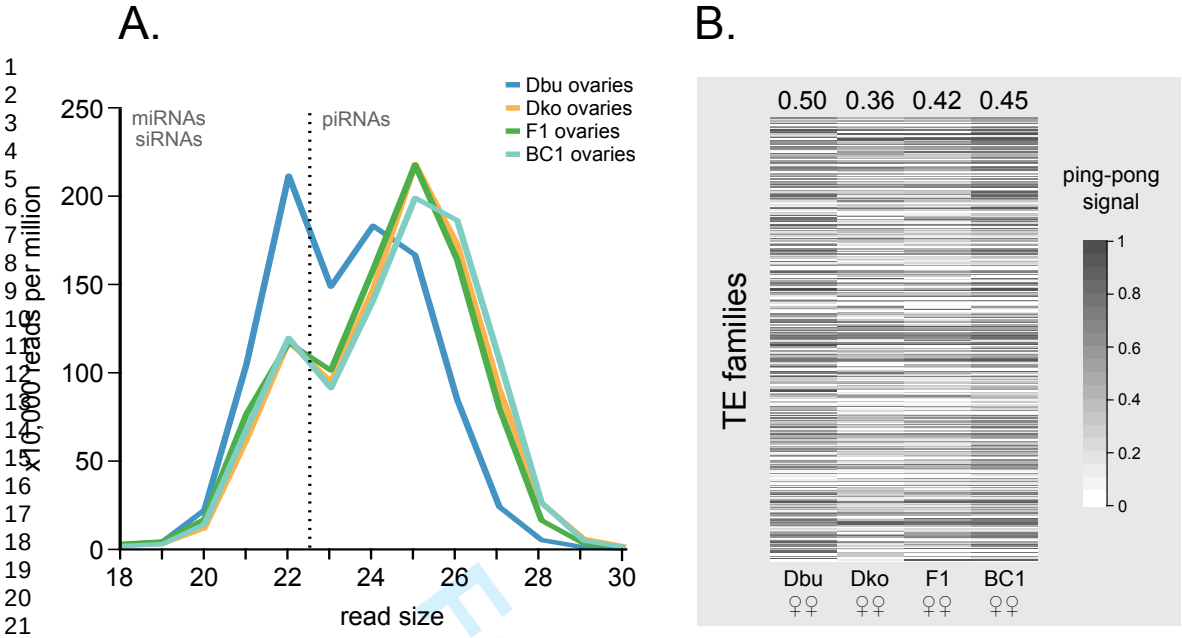
A.



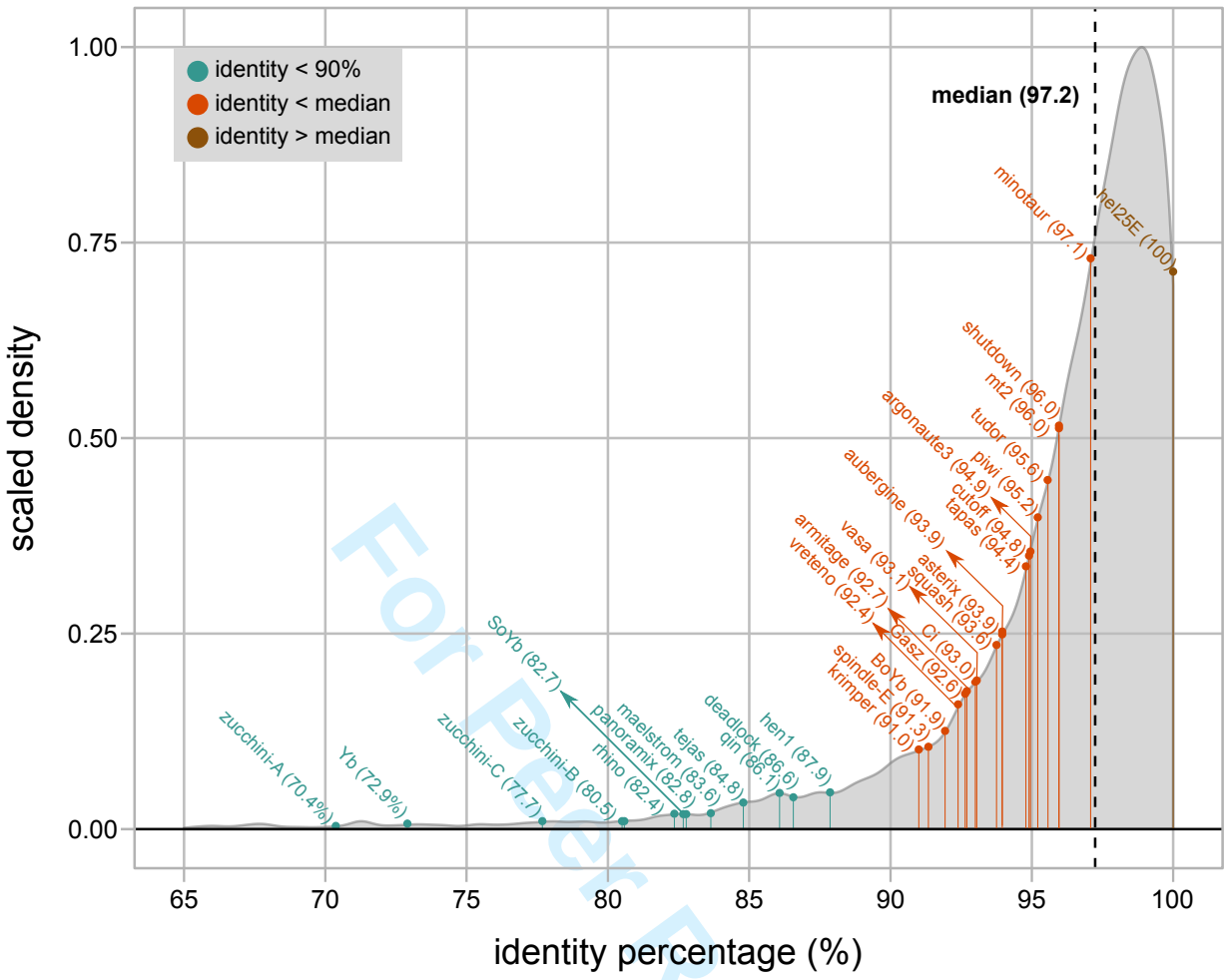
B.



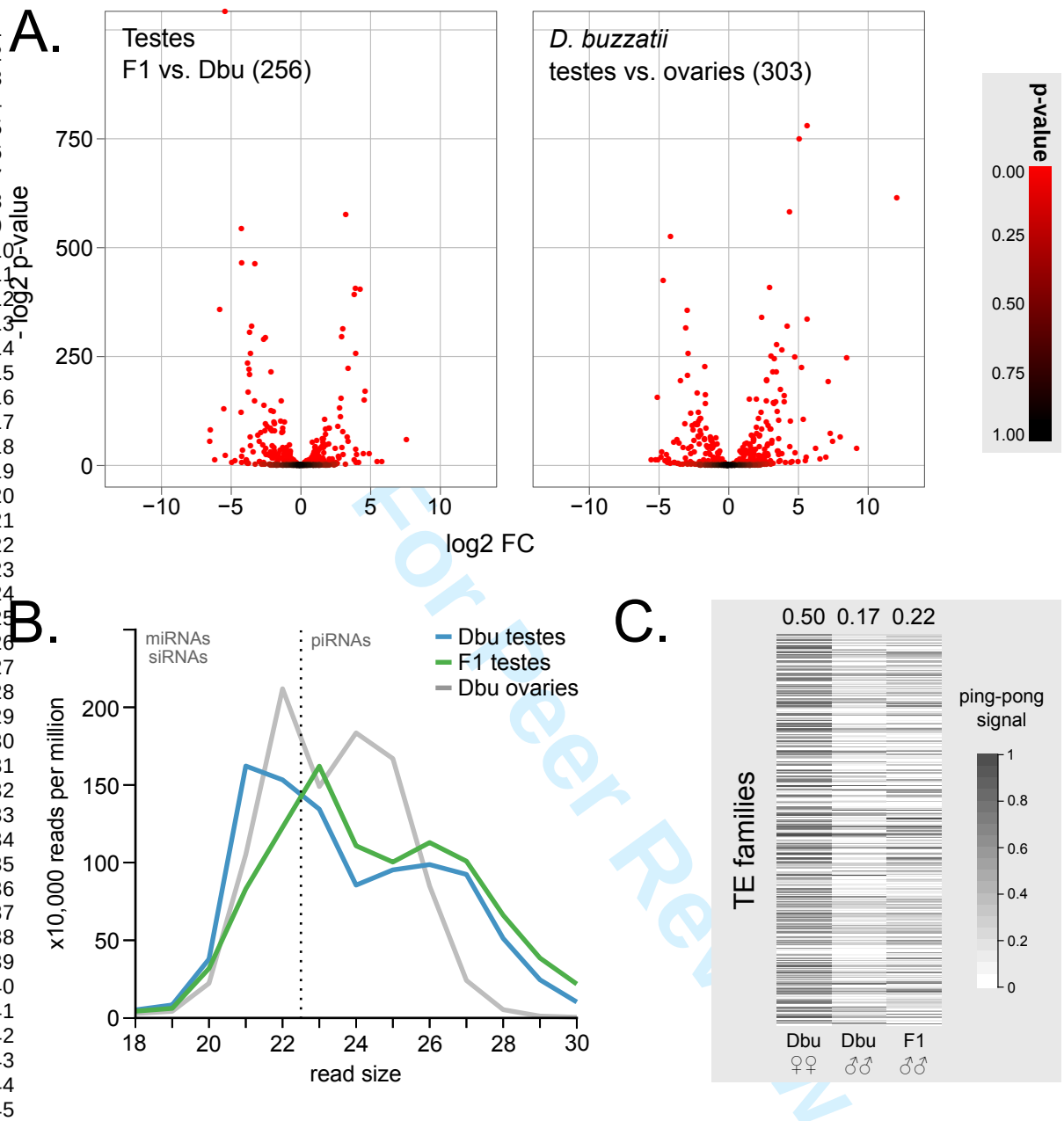
For Peer Review



For Peer Review



For Peer Review



For Peer Review

Table 1. Overexpressed TE families in hybrid ovaries. Dbu= *D. buzzatii*; Dko= *D. koepferae*; FC= fold change; BH= Benjamini–Hochberg correction. [a] overexpressed only in BC1; [b] FC increases after BC.

TE family	Order	Superfamily	F1 ovaries				BC1 ovaries			
			log2(FC) vs.		BH adjusted p-value		log2(FC) vs.		BH adjusted p-value	
			Dbu	Dko	Dbu	Dko	Dbu	Dko	Dbu	Dko
Homo6	DNA	hAT	2.46	4.32	5.47E-75	7.81E-135	2.38	4.25	2.26E-70	5.04E-130
Homo8	DNA	hAT	2.55	6.26	3.35E-40	5.01E-153	1.97	5.68	8.03E-24	1.77E-125
R=81	DNA	hAT	0.68	0.79	1.23E-03	1.44E-04	0.62	0.73	5.92E-03	4.50E-04
rnd-5_family-1117	DNA	hAT	0.63	0.37	1.44E-03	7.44E-02	-	-	-	-
VEGE_DW [b]	DNA	hAT	1.26	6.53	3.28E-04	2.02E-22	2.69	7.96	1.64E-16	3.04E-33
Rehavkus-2_Nvi	DNA	MULE-MuDR	0.77	0.46	8.12E-08	2.00E-03	-	-	-	-
rnd-5_family-4211	DNA	MULE-MuDR	0.37	0.56	7.16E-02	3.61E-03	-	-	-	-
DNA8-7_CQ	DNA	OtherDNA	0.61	0.65	9.85E-06	1.51E-06	0.38	0.43	1.49E-02	2.51E-03
rnd-4_family-786	DNA	Transib	0.41	0.67	5.59E-02	9.17E-04	-	-	-	-
rnd-5_family-1551	DNA	Transib	0.69	0.48	4.49E-04	1.76E-02	-	-	-	-
CR1-1_CQ	LINE	CR1	1.16	0.80	2.25E-04	1.31E-02	-	-	-	-
CR1-2_CQ	LINE	CR1	0.52	0.53	2.94E-02	2.24E-02	-	-	-	-
I_DM	LINE	I	1.28	2.58	1.07E-02	2.61E-07	1.27	2.57	1.82E-02	2.27E-07
rnd-5_family-156	LINE	I	1.68	0.96	1.65E-08	1.81E-03	1.36	0.64	1.28E-05	4.89E-02
BS-like	LINE	Jockey	5.33	3.90	5.91E-69	1.82E-45	4.73	3.31	4.52E-54	1.02E-32
Jockey-2_Dya	LINE	Jockey	2.39	5.77	5.28E-69	1.98E-129	0.32	3.70	9.10E-02	2.50E-51
rnd-3_family-39	LINE	Jockey	0.39	0.58	4.60E-03	7.14E-06	-	-	-	-
TART_B1 [a]	LINE	Jockey	-	-	-	-	1.46	2.30	3.53E-02	3.45E-04
TART	LINE	Jockey	7.24	3.14	1.13E-58	2.60E-26	5.74	1.64	1.43E-36	1.11E-07

For Peer Review

rnd-4_family-338	LINE	L2	0.57	0.40	4.36E-04	1.83E-02	-	-	-	-
rnd-5_family-2046	LINE	L2	0.71	0.65	1.84E-04	6.54E-04	-	-	-	-
Bilbo	LINE	LOA	0.83	1.02	8.33E-13	8.82E-19	0.78	0.97	4.22E-11	4.64E-17
R1_Dps	LINE	R1	0.56	0.81	3.23E-05	5.52E-10	0.53	0.78	1.57E-04	1.91E-09
rnd-5_family-1630	LINE	R1	0.53	0.63	1.03E-04	2.48E-06	0.30	0.40	7.15E-02	4.93E-03
RT2	LINE	R1	0.74	0.53	1.21E-08	5.45E-05	-	-	-	-
RTAg3	LINE	R1	0.93	1.02	3.33E-05	5.48E-06	0.54	0.63	4.22E-02	7.98E-03
RTAg4	LINE	R1	0.51	0.60	2.20E-04	6.74E-06	-	-	-	-
BEL1-I_Dmoj	LTR	BelPao	2.81	4.13	5.42E-24	1.03E-47	1.02	2.34	1.33E-03	1.15E-15
BEL1-LTR	LTR	BelPao	1.53	1.92	3.80E-03	3.25E-04	1.05	1.45	9.10E-02	9.24E-03
Gypsy-14_Dwil-I [a]	LTR	Gypsy	-	-	-	-	3.94	3.91	7.45E-02	4.72E-02
Gypsy-151_AA-I	LTR	Gypsy	0.43	0.71	4.33E-03	8.58E-07	-	-	-	-
Gypsy16-I_Dpse	LTR	Gypsy	12.76	7.39	2.88E-36	5.41E-150	11.47	6.09	2.94E-29	5.80E-102
Gypsy-172_AA-I	LTR	Gypsy	0.64	0.81	4.66E-02	7.87E-03	-	-	-	-
Gypsy-18_Dwil-I [b]	LTR	Gypsy	11.10	6.04	1.49E-199	8.22E-174	12.01	6.95	8.02E-234	2.40E-230
Gypsy-18_Dwil-LTR [b]	LTR	Gypsy	10.35	7.19	2.00E-21	9.12E-52	11.48	8.32	5.49E-26	2.18E-69
Gypsy5-I_Dya	LTR	Gypsy	12.40	8.88	0.00E+00	0.00E+00	10.94	7.41	0.00E+00	0.00E+00
Gypsy61-I_AG	LTR	Gypsy	0.31	1.00	5.90E-02	7.47E-13	-	-	-	-
Gypsy6-I_Dya [b]	LTR	Gypsy	7.21	3.87	1.15E-91	6.99E-47	8.03	4.69	5.22E-114	3.81E-69
Gypsy6-LTR_Dya [a]	LTR	Gypsy	-	-	-	-	4.17	2.48	5.89E-11	5.30E-07
Gypsy7-I_Dmoj [a]	LTR	Gypsy	-	-	-	-	4.23	0.38	5.37E-98	5.37E-02
Gypsy8-I_Dpse	LTR	Gypsy	0.42	0.84	2.23E-03	3.08E-11	-	-	-	-
R=961 [a]	LTR	Gypsy	-	-	-	-	1.71	1.28	6.75E-03	3.08E-02
rnd-5_family-2676 [a]	LTR	Gypsy	-	-	-	-	2.72	1.04	1.74E-22	8.93E-05
mean			2.48		6.16E-03		3.34		1.22E-02	

For Peer Review

For Peer Review

Table 2. Underexpressed TE families in hybrid ovaries. Dbu= *D. buzzatii*; Dko= *D. koepferae*; FC= fold change; BH= Benjamini–Hochberg correction. [a] underexpressed only in BC1; [b] FC increases after BC.

TE family	Order	Superfamily	F1 ovaries				BC1 ovaries			
			log2(FC) vs.		BH adjusted p-value		log2(FC) vs.		BH adjusted p-value	
			Dbu	Dko	Dbu	Dko	Dbu	Dko	Dbu	Dko
Howilli1 [a]	DNA	hAT	-	-	-	-	-1.70	-1.59	8.09E-02	7.33E-02
MINOS	DNA	Tc1Mariner	-1.32	-0.53	8.12E-08	6.02E-02	-	-	-	-
rnd-5_family-1477 [a]	DNA	Tc1Mariner	-	-	-	-	-0.59	-1.13	1.21E-06	6.24E-24
rnd-5_family-3658 [a]	DNA	Tc1Mariner	-	-	-	-	-0.66	-0.97	2.23E-02	8.48E-05
Transib1_DP [b]	DNA	Transib	-0.57	-0.90	8.58E-02	2.44E-03	-0.64	-0.97	6.76E-02	8.76E-04
Transib3_DP	DNA	Transib	-2.01	-2.86	9.45E-02	8.46E-03	-	-	-	-
HELITRON1_DM	RC	Helitron	-3.37	-3.11	1.34E-02	2.37E-02	-	-	-	-
Helitron-1_Dvir	RC	Helitron	-0.81	-0.32	4.66E-08	5.73E-02	-	-	-	-
rnd-3_family-48	RC	Helitron	-0.95	-0.59	1.29E-16	7.62E-07	-0.60	-0.23	6.44E-07	7.37E-02
rnd-4_family-133	RC	Helitron	-1.08	-0.53	1.50E-06	3.50E-02	-	-	-	-
DMCR1A-like	LINE	CR1	-1.21	-0.65	8.95E-11	1.27E-03	-	-	-	-
DPSEMINIME-like	LINE	CR1	-0.76	-0.26	2.38E-08	9.53E-02	-	-	-	-
DMRER1DM-like	LINE	R1	-1.55	-1.08	4.39E-09	1.08E-04	-	-	-	-
BEL-11_Dta-I	LTR	BelPao	-1.91	-1.29	7.37E-18	1.24E-08	-	-	-	-
BEL-20_AA-I [a]	LTR	BelPao	-	-	-	-	-0.67	-0.52	2.23E-02	6.39E-02
BEL-3_Dta-I	LTR	BelPao	-0.70	-0.61	8.23E-03	2.24E-02	-0.57	-0.48	5.13E-02	7.61E-02
BEL-6_Dwil-I	LTR	BelPao	-1.08	-1.47	1.10E-02	2.05E-04	-	-	-	-
BEL-8_Dwil-I	LTR	BelPao	-2.08	-1.10	5.93E-17	3.88E-05	-	-	-	-
Nobel_I [b]	LTR	BelPao	-0.81	-0.73	9.17E-06	6.08E-05	-0.82	-0.74	9.24E-06	3.64E-05

For Peer Review

rnd-4_family-529 [b]	LTR	BelPao	-0.45	-0.91	9.41E-02	1.06E-04	-0.70	-1.16	8.53E-03	4.98E-07
rnd-5_family-1078	LTR	BelPao	-1.00	-0.44	2.92E-12	3.79E-03	-	-	-	-
rnd-5_family-2670	LTR	BelPao	-2.02	-1.11	2.35E-28	1.50E-08	-	-	-	-
Copia-3-like [a]	LTR	Copia	-	-	-	-	-0.45	-1.04	6.63E-02	8.92E-08
rnd-5_family-4686	LTR	Copia	-0.92	-1.08	1.24E-02	2.22E-03	-	-	-	-
Beagle-like	LTR	Gypsy	-0.59	-1.27	1.58E-02	5.00E-09	-	-	-	-
Gypsy1-I_Dmoj	LTR	Gypsy	-0.85	-1.05	8.73E-04	2.01E-05	-0.53	-0.73	6.52E-02	2.80E-03
Gypsy-22_Dya-I [b]	LTR	Gypsy	-1.74	-1.63	1.23E-04	3.51E-04	-2.13	-2.02	5.53E-06	9.98E-06
Gypsy2-I_DM	LTR	Gypsy	-1.17	-0.65	3.86E-10	1.20E-03	-	-	-	-
Gypsy-31_Dwil-I [a]	LTR	Gypsy	-	-	-	-	-1.11	-2.33	5.27E-02	5.69E-07
Gypsy4-I_Dpse	LTR	Gypsy	-1.90	-0.90	1.40E-26	1.62E-06	-1.37	-0.38	8.49E-15	6.15E-02
Gypsy50-like	LTR	Gypsy	-0.98	-2.47	1.34E-02	4.85E-13	-	-	-	-
QUASIMODO-like [a]	LTR	Gypsy	-	-	-	-	-0.58	-1.20	1.62E-02	1.38E-09
rnd-5_family-1084	LTR	Gypsy	-0.91	-1.85	8.70E-03	1.66E-09	-0.67	-1.61	7.57E-02	2.96E-08
TABOR_DA-LTR [a]	LTR	Gypsy	-	-	-	-	-3.27	-3.46	5.43E-02	2.13E-02
	mean		-1.19		1.29E-02		-1.11		2.81E-02	

For Peer Review

For Peer Review

Table 3. Summary of differential expression analyses of piRNA pathway genes: comparisons between parental species and between parents and hybrids. Dbu= *D. buzzatii*; Dko= *D. koepferae*; FC= fold change; BH= Benjamini–Hochberg correction. * significant p-value.

Gene name	Gene symbol	*D. buzzatii* vs. *D. koepferae*			F1 vs. parental species				BC1 vs. parental species			
		% id	log2(FC)	BH adjusted p-value	log2(FC)		BH adjusted p-value		log2(FC)		BH adjusted p-value	
					Dbu	Dko	Dbu	Dko	Dbu	Dko	Dbu	Dko
Argonaute3	Ago3	94.90	0.80	3.60E-29*	-0.77	0.02	1.69E-27*	7.68E-01	-0.76	0.04	6.66E-26*	6.41E-01
Armitage	armi	92.70	-0.59	1.51E-18*	0.43	-0.16	2.77E-10*	2.24E-02*	0.27	-0.33	1.86E-04*	1.43E-06*
asterix	arx	93.89	1.73	4.67E-65*	-0.30	1.43	3.21E-03*	2.45E-44*	-0.02	1.71	8.72E-01	2.43E-63*
aubergine	aub	93.92	2.62	3.45E-183*	-0.98	1.64	1.24E-26*	1.56E-72*	-0.46	2.16	1.08E-06*	1.40E-124*
Brother of Yb	BoYb	91.93	-0.42	9.63E-09*	0.52	0.10	1.79E-12*	1.83E-01	0.49	0.07	7.25E-11*	3.39E-01
cubitus interruptus	Ci_tf	92.97	-1.52	2.73E-18*	0.34	-1.18	6.40E-02*	1.66E-11*	0.24	-1.28	2.55E-01	2.78E-13*
cutoff	cuff	94.79	1.85	1.62E-78*	-0.64	1.22	2.77E-10*	2.16E-34*	-0.07	1.78	5.77E-01	1.68E-72*
deadlock	del	86.56	-0.88	7.51E-14*	0.32	-0.57	8.98E-03*	2.57E-06*	-0.03	-0.91	8.72E-01	1.82E-14*
GASZ ortholog	Gasz	92.64	0.65	1.00E-21*	0.07	0.72	3.05E-01	3.98E-26*	0.37	1.02	1.01E-07*	8.22E-52*
helicase at 25E	Hel25E	100	-0.41	1.36E-17*	0.25	-0.16	2.97E-07*	1.29E-03*	0.07	-0.34	2.51E-01	1.40E-12*
Hen1	Hen1	87.86	-0.02	9.13E-01	-0.44	-0.46	2.50E-06*	1.87E-06*	-0.50	-0.51	2.48E-07*	7.01E-08*
krimper	krimp	91.00	5.04	0.00E+00*	-0.62	4.41	3.02E-32*	0.00E+00*	-0.07	4.97	2.59E-01	0.00E+00*
maelstrom	mael	83.64	-1.20	8.48E-66*	0.77	-0.43	1.69E-27*	8.37E-10*	0.39	-0.81	1.13E-07*	6.11E-31*
minotaur	mino	97.08	-0.30	1.11E-04*	0.31	0.01	9.79E-05*	9.17E-01	0.03	-0.27	7.79E-01	5.30E-04*
Methyltransferase2	Mt2	95.95	0.74	9.90E-18*	-0.07	0.67	3.65E-01	2.95E-14*	-0.06	0.68	5.77E-01	6.58E-15*
Panoramix	Panx	95.95	0.01	9.20E-01	0.48	0.50	3.89E-09*	1.81E-09*	0.32	0.33	1.86E-04*	5.27E-05*
piwi	piwi	95.21	0.13	4.58E-02*	-0.23	-0.11	2.51E-04*	1.03E-01	-0.20	-0.07	2.49E-03*	2.63E-01
qin	qin	86.07	-1.30	9.28E-14*	0.47	-0.83	8.98E-03*	2.85E-06*	0.02	-1.29	9.23E-01	2.94E-13*
rhino	rhi	82.35	-1.03	7.85E-27*	0.34	-0.69	6.93E-04*	5.76E-13*	-0.06	-1.09	6.61E-01	1.13E-29*

For Peer Review

shutdown	shu	95.97	2.26	0.00E+00*	-0.64	1.63	1.09E-53*	4.43E-302*	-0.17	2.10	1.37E-04*	0.00E+00*
Sister of Yb	SoYb	82.65	-0.32	4.11E-02*	-1.30	-1.62	1.43E-16*	4.20E-25*	-0.50	-0.82	2.11E-03*	9.76E-08*
spindle E	spn-E	91.34	-0.85	3.11E-17*	0.52	-0.33	5.13E-07*	1.29E-03*	0.23	-0.62	3.73E-02*	1.27E-09*
squash	squ	93.55	1.34	8.63E-23*	-0.72	0.62	1.10E-07*	9.35E-06*	-0.73	0.61	1.45E-07*	1.09E-05*
tapas	tapas	94.42	-0.94	3.03E-19*	0.63	-0.31	3.74E-09*	3.97E-03*	0.17	-0.77	1.67E-01	3.09E-13*
tejas	tej	84.79	0.01	9.62E-01	0.15	0.15	1.95E-01	1.83E-01	0.02	0.02	8.90E-01	8.52E-01
tudor	tud	95.56	-0.50	7.43E-04*	0.32	-0.19	3.89E-02*	2.26E-01	0.14	-0.37	4.80E-01	1.50E-02*
vasa	vas	93.05	0.67	1.41E-43*	-0.16	0.51	1.56E-03*	5.27E-26*	-0.11	0.56	4.90E-02*	3.57E-31*
vret	vreteno	92.39	0.68	7.64E-21*	-0.29	0.39	9.79E-05*	1.09E-07*	-0.26	0.42	7.92E-04*	6.71E-09*
Yb	Yb	72.89	1.05	4.22E-43*	-0.09	0.96	2.23E-01	1.11E-35*	-0.37	0.68	5.50E-07*	2.91E-18*
zucchini (A)	zucA	70.37	-1.55	4.19E-62*	1.21	-0.34	8.74E-38*	3.07E-04*	0.87	-0.67	5.21E-20*	4.03E-13*
zucchini (B)	zucB	80.50	-2.17	2.02E-04*	1.02	-1.15	1.10E-01	2.24E-02*	0.71	-1.45	3.57E-01	4.31E-03*
zucchini (C)	zucC	77.68	1.16	8.18E-53*	-0.28	0.88	1.65E-04*	2.05E-30*	-0.22	0.95	5.11E-03*	4.67E-35*
zucchini (D)	zucD	-	-0.43	6.87E-01	0.04	-0.39	9.62E-01	7.01E-01	0.48	0.05	6.61E-01	9.55E-01

For Peer Review

For Peer Review

2 TEtools : quantification des éléments transposables et des piRNA dans des données RNA-seq

Nucleic Acids Research Advance Access published October 19, 2016

Nucleic Acids Research, 2016 **1**
doi: 10.1093/nar/gkw953

TETOOLS facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes

Emmanuelle Lerat†, Marie Fablet†, Laurent Modolo†, Hélène Lopez-Maestre and Cristina Vieira*

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

Received March 11, 2016; Revised September 29, 2016; Editorial Decision October 06, 2016; Accepted October 11, 2016

ABSTRACT

Over recent decades, substantial efforts have been made to understand the interactions between host genomes and transposable elements (TEs). The impact of TEs on the regulation of host genes is well known, with TEs acting as platforms of regulatory sequences. Nevertheless, due to their repetitive nature it is considerably hard to integrate TE analysis into genome-wide studies. Here, we developed a specific tool for the analysis of TE expression: TEtools. This tool takes into account the TE sequence diversity of the genome, it can be applied to unannotated or unassembled genomes and is freely available under the GPL3 (https://github.com/l-modolo/TEtools). TEtools performs the mapping of RNA-seq data obtained from classical mRNAs or small RNAs onto a list of TE sequences and performs differential expression analyses with statistical relevance. Using this tool, we analyzed TE expression from five *Drosophila* wild-type strains. Our data show for the first time that the activity of TEs is strictly linked to the activity of the genes implicated in the piwi-interacting RNA biogenesis and therefore fits an arms race scenario between TE sequences and host control genes.

INTRODUCTION

Transposable elements (TEs) are mobile sequences that can be highly abundant in genomes (1). First described by B. McClintock in the 1950s (2), TEs have a high impact on genome dynamics, and are undoubtedly major players in genome evolution (1,3). Despite the increasing amount of transcriptomic data being produced for many species, very few studies have performed genome-wide analyses of the transcription levels of TEs (4–7). Such knowledge gap is partly due to the low levels of transcription of TEs in normal conditions, but also to the fact that one given TE family may be represented by several sequences, making more difficult to have an accurate idea of TE transcription levels.

In Drosophila, a category of small RNAs called piwi-interacting RNAs (piRNAs) are involved in the control of TEs in germline and somatic cells (8–11) and participate in transcriptional and post-transcriptional control of TEs (12). The disruption of the piRNA biogenesis pathway leads to TE mobilization (transcription and transposition), DNA breaks and sterility (13). Understanding the way TE activity is regulated thus requires to have an accurate knowledge of piRNA abundances which could then be associated with TE mRNA levels. Currently, no available method is dedicated to both the analysis of TE expression and piRNA production, associated with differential expression analysis with statistical relevance, for both model and non-model species with non-annotated genomes.

Presently, one tool is available to analyze piRNAs that is based on the approach proposed by Brennecke (10,14). This tool is suited for the analysis of well annotated genomes. However, the methodology that is applied may lead to a loss of information. The first step consisting in a strict mapping at a unique position on the reference genome makes two strong assumptions. Firstly, retaining only reads mapping with no mismatch implies that the corresponding small RNA displays a perfect match with the regulated TE sequences. Secondly, retaining only reads mapping at unique positions when they are supposed to target repeated sequences assumes that only particular small RNA can be generated by only one given position. Other major problems are that this step completely relies on the quality of the genome sequence and assembly, and that it cannot be directly applied when a TE family is absent from the reference genome but exists in the genomes of other strains. Moreover, the association between piRNAs and the TE family is

*To whom correspondence should be addressed. Tel: +33 4 72 43 29 18; Fax: +33 4 72 44 88 98; Email: cristina.vieira@univ-lyon1.fr
†These authors contributed equally to the paper as first authors.

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on October 20, 2016

made by comparing the reads to TE consensus sequences and allowing up to three mismatches, which corresponds to a divergence of approximately 10%. The consensus sequence in itself represents an average sequence of a given family and may result in a sequence that is not present in the genome. A consensus will be representative of the family only if the copies used to build it are very similar, which is the case for the majority of the *Drosophila melanogaster* families, but it is not the case in other Drosophila genomes, such as the sister species *Drosophila simulans* (15). The same is true when determining TE expression from mRNA reads.

In this article we propose a different approach implemented in the pipeline TETOOLS which is dedicated to the analysis of the TE transcriptome, and takes into account the sequence diversity at the TE copy level, using a complete list of all available TE copies from an organism. This pipeline provides quantitative information for both small and messenger RNAs, performing differential expression analyses among different samples using the DESeq2 program (16). It can be used for non-model organisms with no annotated reference genome but for which a list of TE copies is available. When this list is not available, TETOOLS can be jointly used with a dedicated tool for TE identification from raw reads, such as DnaPipeTE (17), RepeatExplorer (18) or other TE identification tools if the genome is assembled (see as a review (19)). The pipeline is user friendly and is available for use in Galaxy (20).

We applied TETOOLS to explore TE regulation in *D. simulans* wild-type strains. In this species, TE sequences belonging to the same family are very diverse and the activity of TEs depends on the strain studied (21–24). Several hypotheses have been proposed to understand the origin and evolution of the intra-specific variability of TEs (25–28), but none has integrated in a satisfying way the high variability uncovered in **g**enes **i**nvolved in the **p**iRNA **p**athway (GIPPs) (both at the DNA sequence (29,30) and transcription levels (31)). Indeed, we propose that the natural variation of TEs is due to variability in the piRNA pathway, which evolves very rapidly and constitutes a genomic immune pathway (29–31). We sequenced mRNAs and small RNAs in several wild-type strains of *D. simulans* and used TETOOLS to analyze TE expression levels and the production of corresponding piRNAs. Our results show, for the first time, a negative relationship between TE and GIPP activities and provide insights into the dynamics of TEs in their natural context.

MATERIALS AND METHODS

Biological material

Four wild-type strains of *D. simulans* were used; these strains originated from various regions around the world: Chicharo (Portugal), Makindu (Kenya), Mayotte (Indian Ocean island) and Zimbabwe. We also included the main source of the reference genome sequence (w501). This last strain originated from the USA and was obtained from the UC San Diego Drosophila Stock Center. Flies were kept in the lab at 24°C in regular fruit fly medium.

Thirty pairs of ovaries were dissected in phosphate buffered saline. Total RNA was extracted using the RNeasy kit (Qiagen) followed by RNase treatment (DNA free kit, Ambion). Two replicates were performed for each strain and the overall qualities were assessed using the Bioanalyzer 2100 (Agilent).

Illumina library production and mRNA sequencing

The TruSeq RNA sample Preparation v2 kit (Illumina Inc., California, USA) was used according to the manufacturer's protocol with the following modifications. Poly-A-containing mRNA molecules were purified from 1 μg of total RNA using poly-T oligo-attached magnetic beads. The purified mRNA was fragmented by the addition of the fragmentation buffer and heated to 94°C in a thermocycler for 4 min. A fragmentation time of 4 min was used to yield library fragments of 250–500 bp. First-strand cDNA was synthesized using random primers to eliminate the general bias towards the 3′ end of the transcript. Second-strand cDNA synthesis, end repair, A-tailing and adapter ligation were performed in accordance with the manufacturer's supplied protocols. Purified cDNA templates were enriched by 15 cycles of polymerase chain reaction (PCR) for 10 s at 98°C, 30 s at 65°C and 30 s at 72°C using the PE1.0 and PE2.0 primers and the Phusion DNA polymerase (NEB, USA). Each indexed cDNA library was verified and quantified using a DNA 100 Chip on a Bioanalyzer 2100 and then mixed equally with six different samples. The final library was quantified by real-time PCR with the KAPA Library Quantification Kit for Illumina Sequencing Platforms (Kapa Biosystems Ltd, South Africa), adjusted to 10 nM in water and provided to the Get-PlaGe core facility (GenoToul platform, INRA Toulouse, France http://www.genotoul.fr) for sequencing. The final mixed cDNA library was sequenced using the Illumina mRNA-Seq paired-end protocol on a HiSeq2000 sequencer for 2 × 100 cycles. Each sample provided between 30 and 55 million reads (SRX1287831, SRX1287832, SRX1287833, SRX1287834 and SRX1287843).

Small RNA extraction and sequencing

Small RNAs from *D. simulans* ovaries were manually isolated in HiTrap Q HP anion exchange columns (GE Healthcare) as described in Grentzinger and Chambeyron (32). Library construction and 50 nt read sequencing were performed by Fasteris SA (Switzerland) on an Illumina HiSeq 2500 instrument. Libraries from the Makindu and Chicharo strains were previously published (33). The small RNA library of the Mayotte strain is available under the accession number SRX1287860. The poly-A tails attached to the sequence before sequencing to obtain 50nt RNA were removed using UrQt (–N A) before other analysis (34).

Gene transcript analysis

D. simulans gene sequences were obtained from FlyBase (ftp.flybase.net/genomes/Drosophila_simulans/dsim_r1.4_FB2014_03/fasta/dsim-all-gene-r1.4.fasta.gz). RNA-seq reads were trimmed to remove poor quality nucleotides using UrQt (–t 25) (34) and then aligned against *D. simulans* genes using Tophat2 (35). Alignment counts were performed on sorted bam files using eXpress (36), and differential expression was assessed using DESeq2 (16).

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on October 20, 2016

We used a 0.05 FDR threshold value for significance. All subsequent calculations were performed on the DESeq2 normalized read counts. Genetic Euclidian distance matrices were computed on the 10 samples using the R dist() function with default parameters on normalized read counts. We retrieved *D. melanogaster* orthologs using the gene_orthologs_fb_2014_06.tsv.gz file from FlyBase and used the corresponding gene IDs to obtain gene ontology data from FlyBase.

To test whether genes of the piRNA pathway (GIPPs) are more frequently differentially expressed than other genes, we randomly sampled 10 000 sets of 19 genes in the complete list of genes (because our list of GIPPs is made of 19 genes) and determined the proportion of differentially expressed genes for each set. We then compared this empirical distribution of the proportion of differentially expressed genes to the value observed for GIPPs.

TE transcript analyses

Fasta sequences of TE copies and rosette file construction. To be as exhaustive as possible concerning the identification of TE copies in the *D. simulans* genome, we retrieved the copies from the two *D. simulans* sequenced genomes. The first genome was produced in 2007 (37) and corresponded to a hybrid assembly of sequences from five different strains. The second genome was produced in 2013 (38) and corresponded to the sequencing of the majority strain (w501) present in the 2007 version. We used the RepeatMasker program (39) using a custom library of TE references to identify the hits in the genome. The sequences of each copy were obtained using the tool 'One code to find them all' (40) (sequences available upon request). The rosette file (available as Supplementary Data) was generated using the sequence names of each copy by adding a column corresponding to the TE (sub)family and a column corresponding to the TE class, which represented 36 046 copies associated with 793 (sub)families.

*The TE*TOOLS *pipeline.* To determine the read count corresponding to each TE family, we used the first module of TETOOLS (TECOUNT) with the TE (sub)family column in the rosette file as the variable (Figure 1A). The output table from this module was used in the second module (TEDIFF) to perform the differential expression analyses (Figure 1B). The module TEDIFF outputs a table of TE families (or any other variables specified in the rosette file) that are differentially expressed among the various conditions/strains, as well as various graphics on the quality of the analysis and the results corresponding to DESeq2 analyses. As an example, we put on Figure 1(C to H) the graphics corresponding to an mRNA analysis of three of our strains. Figure 1C corresponds to the model goodness of fit of the data that takes into account the within-group variability and that corresponds to the dispersion plot of the data estimates (black), the fit to a trend curve to the maximum likelihood estimates to capture the dependence of these estimates on average expression strength (red) and the maximum *a posteriori* estimates used in testing (blue). Figure 1D and E show the principal component analysis (PCA) of the different samples and the heatmap of the sample-to-sample distances, respectively. These two figures allow to verify that the replicates of a given sample are congruent and may also provide information concerning the grouping of the samples based on the divergence of the variable (TE family expression for example). The heatmap gives additional information over similarities and dissimilarities between samples concerning the variation of TE expression, which do not appear on the PCA. Figure 1F shows a MA plot of all samples, which displays the log2 fold changes of all TEs between all samples according to the mean normalized read counts. The TEs with an adjusted P-value < 0.1 are shown in red and correspond to the differentially expressed TEs. A heatmap corresponding to the expression levels of each variable (TE families for example) for the various samples and replicates is provided (Figure 1G). This allows to visualize the differences between samples and which variables are implicated. The volcano plots of all pairwise sample comparisons are provided with red dots corresponding to differentially expressed variables (TE families for example) between the two considered samples (Figure 1H).

Identification of ping-pong signatures. The identification of ping-pong signatures was performed using the tool Small RNA Signatures (41) after mapping the piRNA reads from each strain onto all TE reference sequences using bowtie (42).

RESULTS

A new approach for the automatic transcriptomic analysis of TEs

We developed TETOOLS, which is a new pipeline to perform analyses of the differential amounts of mRNAs and piRNAs from TE copies across different samples. This tool can be used to analyze factors such as different strains, conditions and tissues. This pipeline is implemented in two different modules.

The first module (TECOUNT, Figure 1A) is a python script that performs the mapping of all reads from the RNA-seq dataset to a large list of TE sequences representing different copies, and produces a list of read counts. The use of a list of TE copies provided by the user rather than a sequenced genome or TE consensus sequences has two advantages. First, we can work with TE families not present in the sequenced genome and with non-annotated genomes. Second, the reads are more likely to map with fewer mismatches onto the TE copy than onto the TE consensus sequence (43). This second point can be critical for piRNA analysis for which the read size is small, and a few mismatches can make a difference between mapped and unmapped reads. In contrast to other analytical pipelines, we set the mapper bowtie (42) to its most sensitive option (–best) to position the maximum number of reads along the TE copies. The parameters of the mapper are set to randomly choose a position for a read mapping at multiple positions with the same score. With these settings and a list of TE copies, we can include more reads than other approaches as they discard reads mapping at multiple positions and reads with non-perfect mapping along the genome. The higher number of reads obtained gives

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on October 20, 2016

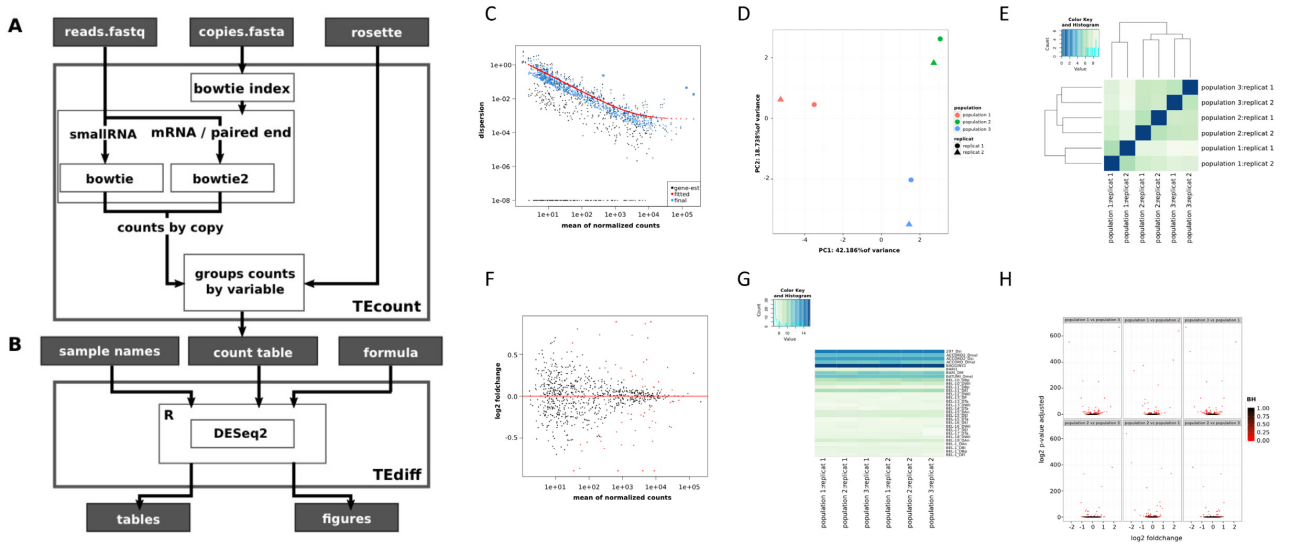


Figure 1. Workflow of the TETOOLS pipeline and the different outputs that can be obtained. (**A**) Details of the TECOUNT module, which uses reads in fastq format, TE sequences in fasta format and a rosette file (see text) as input. (**B**) Details of the TEDIFF module, which uses DESeq2 to perform the differential analysis of expression and produces result files in tables and figures. Examples of the various figures produced by the TEDIFF module are presented from C to H. (**C**) Model goodness of fit of the data. (**D**) Principal component analysis of the different samples with their replicates. (**E**) Heatmap of the various samples. (**F**) MA plot of all samples. The red dots correspond to significant differences. (**G**) Heatmap corresponding to the expression levels of each variable for the various samples and replicates. (**H**) Volcano plots of all pairwise sample comparisons. The figures were obtained with three strains from our mRNA data.

more power for subsequent differential expression analyses. The third input of the TECOUNT module is a rosette file that contains the names of each TE copy. This simple tabular text file can be easily built to group the TE copies by family or any other criteria (i.e. super-family, or even according to other features, such as germline or somatic cell specificity). TECOUNT produces a list of read counts corresponding to the chosen criteria in the rosette file. We stress the fact that TETOOLS uses raw counts in contrast to other piRNA analysis pipelines, which allows the system to avoid biased normalization and to lower the number of false positives for the subsequent differential analyses (44). An option is also available to filter by size and place read counts that could correspond to siRNAs (21 nt-long reads) into a separate file. The novelty of TEtools is that it intends to integrate the TE intra-family sequence diversity that was observed in some genomes. Thus, the expected outcome is a higher number of aligned reads compared to the use of only consensus sequences, as already existing software do. However, in genomes that show low intra-family sequence diversity for TEs—such as *D. melanogaster*—we expect the outcomes of both tools not to be significantly different. We used TETOOLS on our dataset using a list of consensus sequences instead of the full set of TE insertions. The total number of TE aligned reads was then 20% lower to what we got using the full set of TE insertions (2 175 381 versus 1 780 985), reinforcing the relevance of our procedure.

The second module of the TETOOLS pipeline (TEDIFF) is an R script (45) that performs a differential analysis of the read counts using DESeq2 (46) (Figure 1B). TEDIFF requires only the list of counts computed by TECOUNT, a description of each sample (i.e. names and replicates) and a formula specifying the conditions under which to perform the differential analyses. Then, TEDIFF outputs a table of TE families (or any other variables specified in the rosette file) that are differentially expressed among the various conditions/strains. Our tool also uses a logarithmic transformation of read counts (using the Rlog function of DESeq2) to output various graphics on the quality of the analysis and the results (i.e. volcano plots and expression heatmaps) that are ready for interpretation (Figure 1C–H).

TETOOLS was first intended to study small RNA data. However, this tool can also be used to study any type of RNA-seq data, with the possibility of using bowtie2 (47) instead of bowtie for better mapping of mid-length or long reads and paired-end reads (Figure 1A). To use bowtie2 on paired-end reads, the user must specify the size of the insert and the mapper is set to its most sensitive option (–very-sensitive).

To facilitate the use of TETOOLS and its adoption, the pipeline has been implemented as a Galaxy package (20). All the modules of TETOOLS, which are distributed under the GNU General Public License version 3 (https://github.com/l-modolo/TEtools), can also be used with a command line interface.

Gene transcription reflects the geographical distribution of strains

Our dataset was generated from five wild-type strains of *D. simulans*. Four strains of natural origin (Chicharo, Makindu, Mayotte and Zimbabwe) were chosen because they were known to present variable proportions of some TEs, different levels of TE transcripts and different amounts of piRNAs (22,24,33,48,49). We also included w501, which

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on October 20, 2016

is the most represented strain in the 2007 *D. simulans* sequenced genome (37).

Hierarchical clustering on the sample-to-sample distances from normalized gene counts (Figure 2) first clusters samples per replicate of the same strain and then groups them together with two strains from the ancestral area (Mayotte and Makindu) and strains from the derived area (w501 and Chicharo) (50). This geographical pattern is reinforced by the significant correlation between the geographical distance (in km) and genetic distance calculated from the read counts (see 'Materials and Methods' section, Mantel test, $r = 0.434$, P-value $= 0.016$).

Globally, we found that 7416 genes out of a total of 16 169 genes were differentially expressed between the five strains. When we considered the geographical structure (derived versus ancestral areas), we found 3188 differentially expressed genes between the two groups. The top 20 differentially expressed genes belonged to biological categories such as antennal morphogenesis, DNA repair, epigenetic modifications and eye morphogenesis (Supplementary Table S1).

TE expression is variable across *D. simulans* wild-type strains

As previously mentioned, most of the analyses performed to date on TE and gene expression were performed on *D. melanogaster* strains. In this species, copies of TEs are mostly identical (15,51,52), which is not the case for most genomes and especially for other Drosophila genomes (15). For instance, *D. simulans* harbors a majority of degraded and deleted copies (15,48). Thus, the use of the latter organism as a model requires access to all the TE sequence diversity data and hence to use TETOOLS. All figures and the complete tables produced by the TETOOLS pipeline are available as supplemental data (Supplementary Tables S2, 3 and 4; Supplementary File 1).

The PCA discriminates the different strains and the positions of the replicates are consistent in this system (Supplementary Figure S1), indicating that we can globally discriminate between the five different strains based on TE variability. This finding supports previous observations using other experimental approaches concerning the variability in TE expression between natural strains on a global scale (22,25,53,54). According to the normalized read counts, we observe that the most highly expressed TE (sub)families are the same in all strains (Supplementary Table S2). These (sub)families correspond to the Long Terminal Repeat (LTR) retrotransposons Gypsy-28_DAn, and Gypsy-12_DVir and to the non-LTR retrotransposon Jockey3_DSim, which together represent more than 20% of the total TE reads for the different strains (20.48% in w501, 23.49% in Chicharo, 24.04% in Makindu, 25.03% in Mayotte and 23.88% in Zimbabwe).

Pairwise differential analyses allowed us to identify several significant TE (sub)families as differentially expressed (Figure 3). The numbers of these TE (sub)families are indicated in Figure 3A. For example, we can observe that many TE (sub)families are differentially expressed between Makindu and three other strains w501, Chicharo and Zimbabwe (62, 73 and 63 TE (sub)families, respectively). Conversely, only 23 TE (sub)families are differentially expressed between Makindu and Mayotte. In Figure 3B, the log2-fold changes for each differentially expressed TE family for these pairwise comparisons is represented. Clearly, the expression of some TE (sub)families is specific for a given strain compared to the other strains. For example, DM412_Dmel is always more highly expressed in Makindu than in the other strains. The same is true for BLASTOPIA_Dmel in Chicharo and R1_DMo in Zimbabwe.

These data show that the TE transcript levels are significantly different between strains. However, the correlation between genetic distances calculated on TE read counts and geographic distances is weaker than when considering genes (Mantel test, $r = 0.385$, P-value $= 0.036$) (see Results previous section).

piRNA amounts are positively correlated to TE transcript amounts

To deepen our study of TE dynamics, we used piRNA sequencing data previously obtained for three of our wild-type strains (see (33) for Chicharo and Makindu) and we performed small RNA sequencing in one additional strain, Mayotte. These data were analyzed using TETOOLS and all figures and complete tables produced are available as supplemental data (Supplementary Tables S5, 6 and Supplementary File 2). Because the piRNA data were not produced with replicates, DESeq2 could not provide a statistical result on the differential expression analysis. We compared the expression of the piRNAs based on their normalized read counts and observed that the most targeted TEs by piRNAs were the same for all strains (Supplementary Table S7). These TEs correspond to the LTR retrotransposons MAX_Dsi and Gypsy-13_DSim and to the non-LTR retrotransposons R1_Dsi and DMCR1A. The piRNAs of these four elements correspond to 18.54, 15.77 and 23.26% of all piRNA reads in Makindu, Chicharo and Mayotte, respectively (Figure 4A).

The pairwise comparison of the piRNA normalized read counts for each TE family is depicted on Figure 4B. This approach allows us to analyze the piRNA production of specific TEs that display differential mRNA expression levels across the three strains (i.e. the LTR retrotransposons DM412_Dmel, TirantC and BLASTOPIA_Dmel as highlighted in Figure 4B). In these cases, the log2-fold changes in the piRNAs corresponding to these elements are higher than 1.5 (output from TEDIFF). For example, in the comparison between Chicharo and Mayotte, the piRNAs targeting the TirantC element exhibit a log2-fold change of 1.84, with more piRNAs targeting TirantC in the Mayotte strain than in the Chicharo strain. The same is true for this element in the comparison between Chicharo and Makindu, which is in agreement with our experimental knowledge of this TE (33).

The silencing of TEs depends on two distinct piRNA pathways that specifically trigger either somatic or germline-expressed TEs. Primary piRNAs are produced from genomic clusters and are implicated in the somatic regulation of TEs. Secondary piRNAs are either produced from TE transcripts that participate in the ping-pong amplification loop or are maternally transmitted from the mother to the embryo. One way to distinguish primary from

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on October 20, 2016

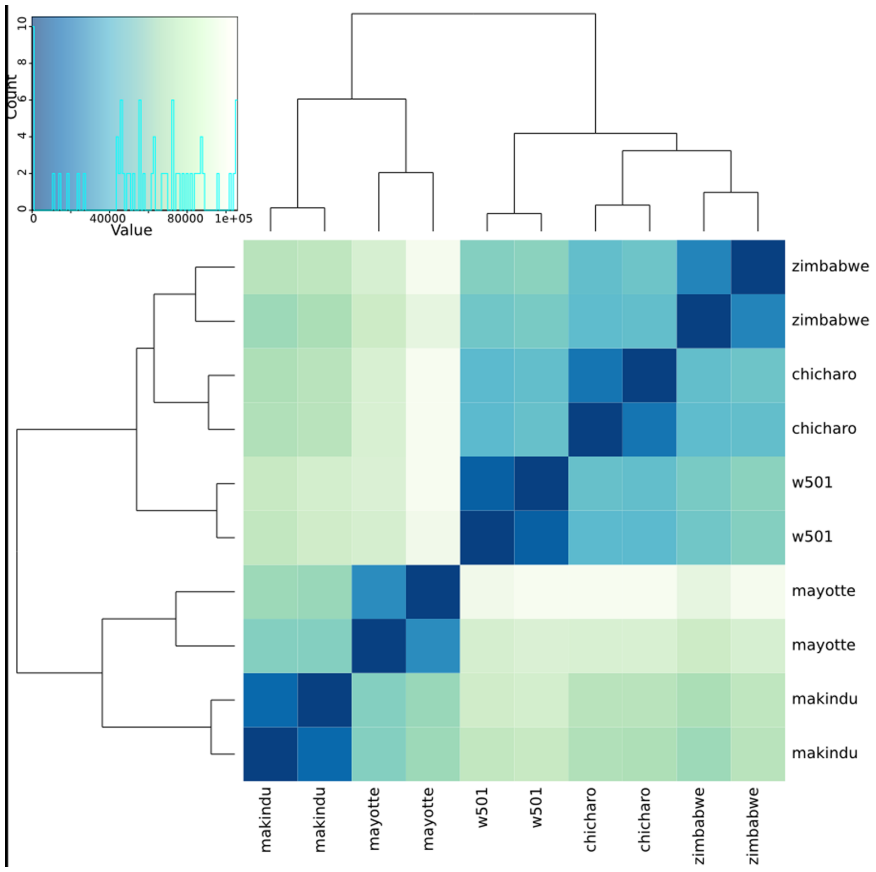


Figure 2. Heatmap of sample-to-sample distances. This heatmap was built using DESeq2 on normalized gene read counts. Strains are clustered by replicate and the analysis separates strains from derived (w501 (USA) and Chicharo (Portugal)) and ancestral (Mayotte and Makindu (Kenya)) areas.

secondary piRNAs is to identify the ping-pong signature. We estimated the proportion of piRNAs implicated in the ping-pong loop for 10 representative TEs with high piRNA production log fold changes (>|1|) (Supplementary Figure S2). We observe that a ping-pong signal is detectable for most of the considered TEs. Additionally, the ping-pong signature is dependent not only on the TEs but also on the strain. For example, no ping-pong signal is detectable in the Chicharo strain when considering the LTR retrotransposon TirantC as is expected from previous experimental work (33). Moreover, a ping-pong signature for this element is detected for the Mayotte strain, which we previously described as having only somatic transcripts (49). The TirantS, which is a structural variant specific to *D. simulans* that was previously described as non-transcribed (22,55), has a very weak ping-pong signature, which is expected for non-active TEs. DOC and Gypsy-13_Dsim present the highest proportion of piRNAs with ping-pong signatures, suggesting that these TEs are probably highly transcribed in the germ line.

One hypothesis to explain the variability in copy numbers between different natural strains links the expression of TEs to the amount of piRNAs (27). Kelleher and Barbash tested this model in two strains of *D. melanogaster*. In the present study, using three strains of *D. simulans*, we found a significant positive correlation between TE read counts and piRNA read counts for each strain (Pearson correlation tests on log transformed read counts: Chicharo: $r = 0.857$, P-value $< 2.10^{-16}$, Makindu, $r = 0.866$, P-value $< 2.10^{-16}$ and Mayotte: $r = 0.860$, P-value $< 2.10^{-16}$, Figure 4C). This finding illustrates a general trend for which an increase in TE transcripts is associated with an increase in piRNA production. This result is expected because secondary piRNAs are implicated in the regulation due to the ping-pong amplification loop. Thus, we searched for ping-pong signatures in the most highly expressed elements. In Supplementary Figure S3, we show that the signature is strong for most of the TEs that have the highest amount of total piRNAs. Moreover, this analysis also reveals TE families that have no associated piRNAs but have reads in the RNA-seq data (197 (sub)families in Chicharo, 186 in Makindu and 222 in Mayotte). This result could indicate that these TEs are absent from piRNA clusters in these specific strains.

TE expression is negatively correlated with piRNA pathway gene activity

The analysis of our dataset provides a demonstration of the huge natural variability in TE expression. Indeed, we find significant variation in the levels of TE transcripts between strains and this is correlated with the corresponding

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on October 20, 2016

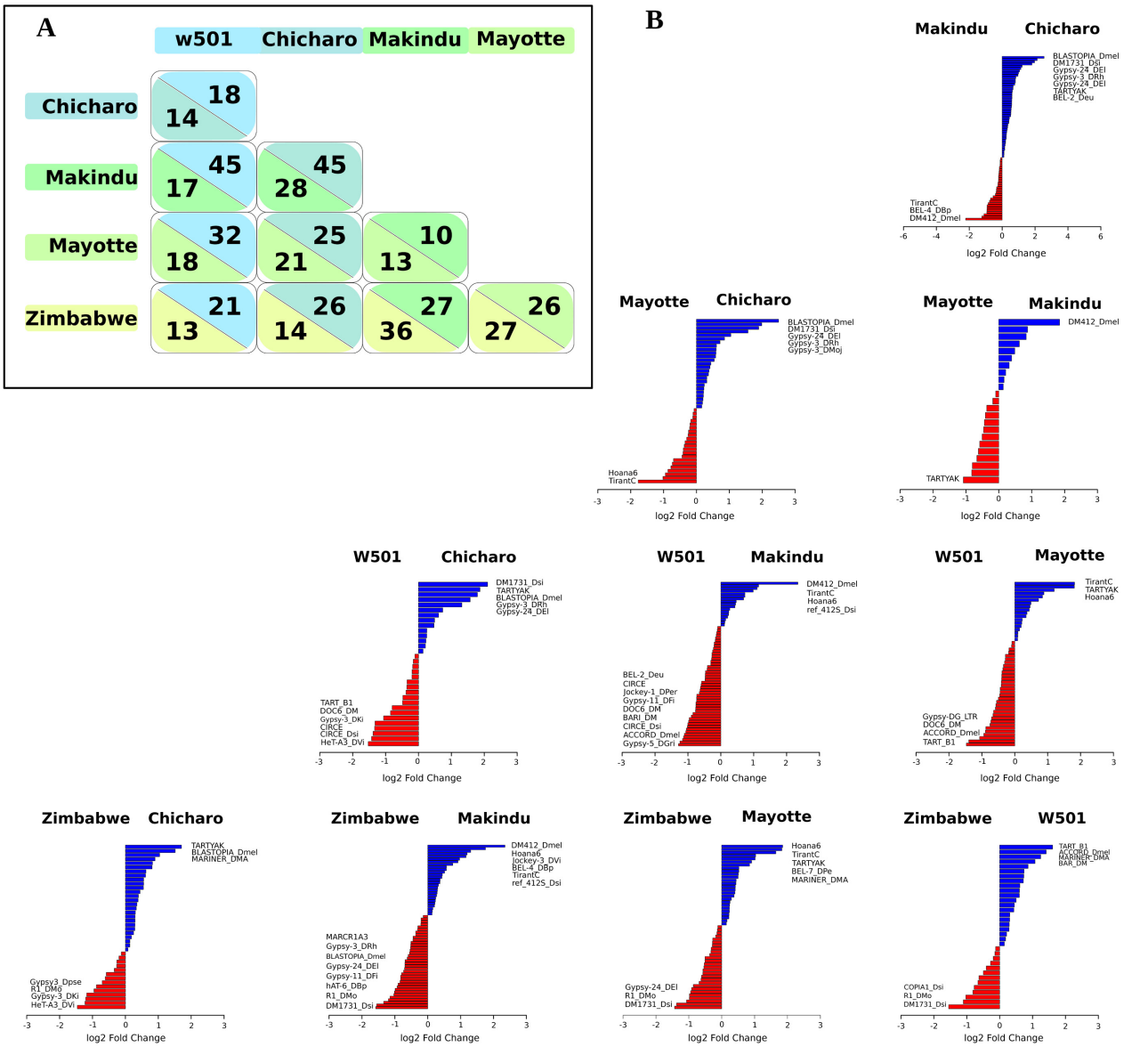


Figure 3. Differentially expressed TEs between strain pairs. (**A**) Numbers of differentially expressed TE (sub)families between strains. The comparisons were performed between pairs of strains. Numbers above the diagonal indicate the numbers of more highly expressed TEs for the strains in columns, numbers above the diagonal indicate the numbers of more highly expressed TEs for the strains in rows. Each color corresponds to a different wild-type strain. (**B**) Pairwise log2-fold change for each differentially expressed TE family. The names of the most differentially expressed TEs are indicated. Blue and red indicate the sense of the comparison.

piRNA production levels. In a previous study, we showed that GIPPs also displayed high transcription and sequence variability (31). Therefore, we sought to confirm the GIPP variability in the present dataset and explore its relationship with TE expression variability.

We focused on subsets of genes involved in the piRNA pathway and used other genes involved in the siRNA and immune pathways for comparison (see Supplementary Table S8 for the complete lists of genes). We find that the piRNA pathway genes are more frequently differentially expressed than other random sets of genes (piRNA pathway 19/19 versus total dataset 7416/16 169, P-value = 0, see 'Materials and Methods' section). Therefore, the analysis of the present dataset confirms the existence of high intra-specific variability for GIPPs.

Subsequently, we tested whether the variability in TE expression was related to GIPP activity estimated by the amount of transcripts. Based on the sum of the read counts for each category of sequences, we find a strong negative correlation between the activity of GIPPs and the global TE expression (Pearson correlation test, $r = -0.93$, P-value = 0.022, Figure 5). No significant correlations are found be-

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on October 20, 2016

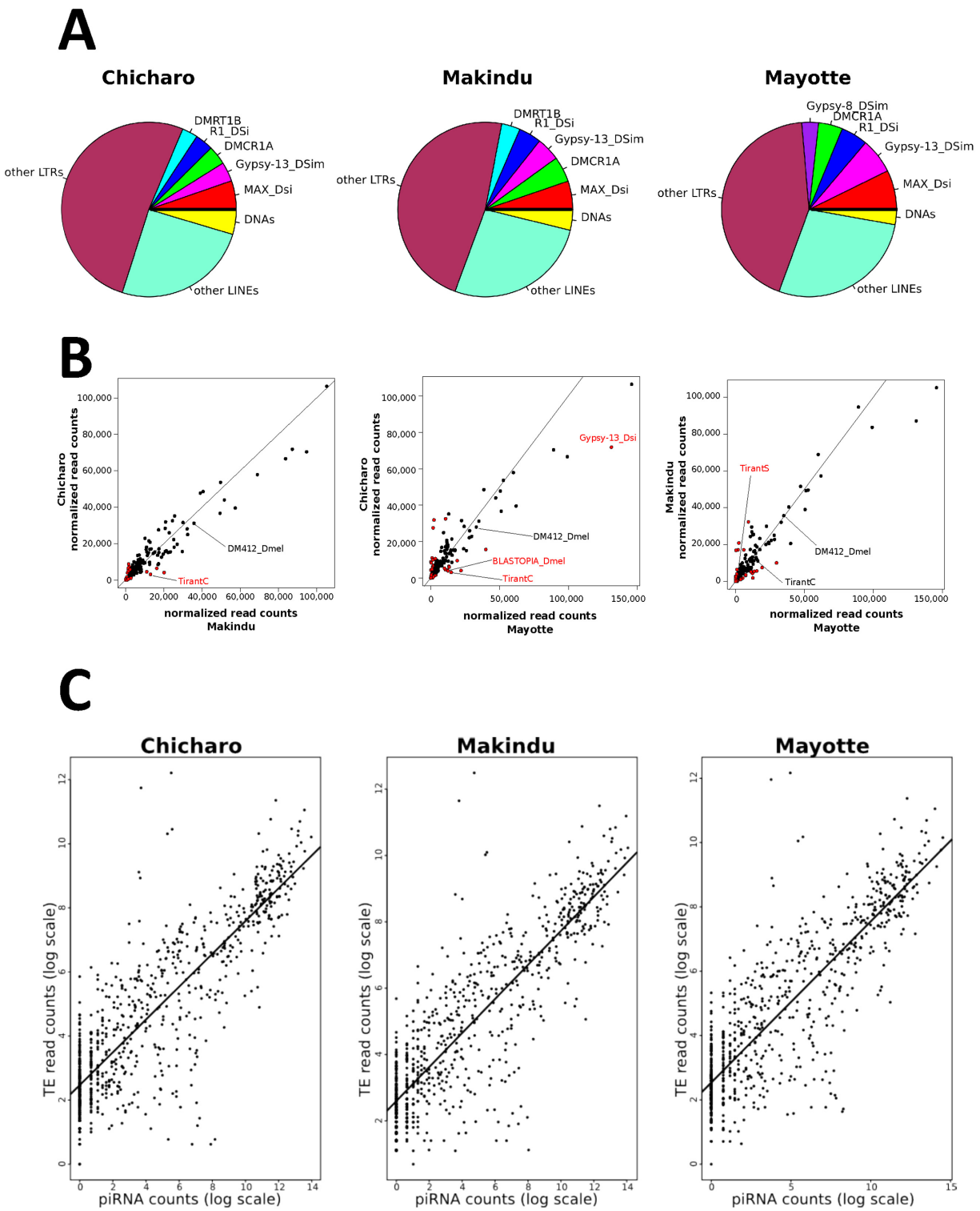


Figure 4. Normalized piRNA read count analysis. (**A**) piRNA production in the different strains. The more abundant piRNAs are identified in the picture and are the same in all the strains. (**B**) Comparison of the normalized piRNA read counts for each pair of strains. Red dots indicate piRNAs with a log2-fold change >1. The black line corresponds to the 1:1 ratio line. As an example we indicate some TEs that display differential mRNA expression levels (see Figure 3). (**C**) Positive correlation between TE read counts and piRNA read counts for the different three strains. Pearson correlation tests on log transformed read counts: Chicharo: $r = 0.857$, P-value $< 2.10^{-16}$, Makindu, $r = 0.866$, P-value $< 2.10^{-16}$ and Mayotte: $r = 0.860$, P-value $< 2.10^{-16}$

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on October 20, 2016

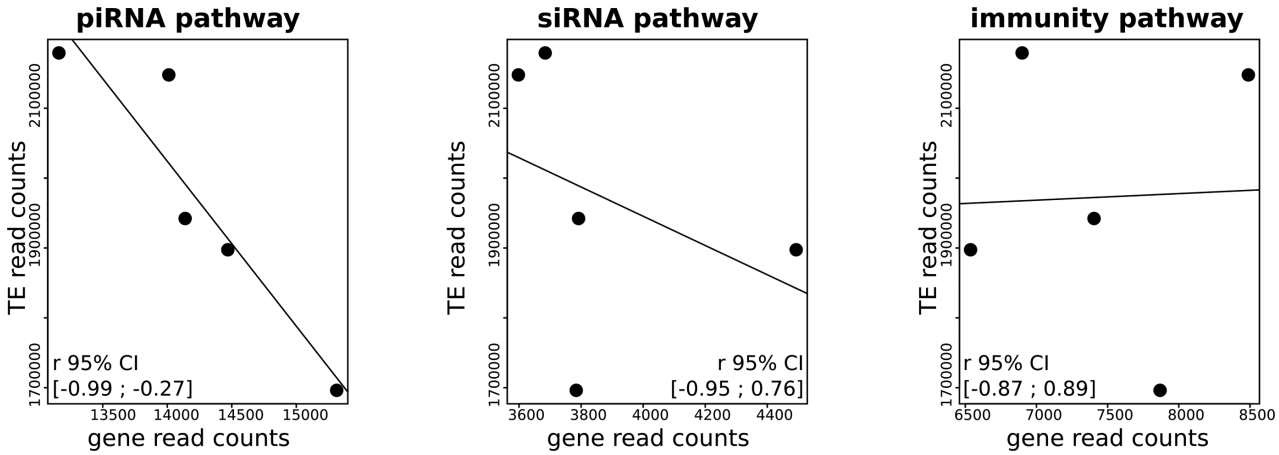


Figure 5. Negative correlation between the sum of TE read counts and the sum of GIPP read counts. No significant correlations are observed when considering genes of the siRNA pathway or genes of the immunity. Confidence intervals (95%) for Pearson correlation coefficients are mentioned at the bottom of each graph.

tween TE expression and the activity of the siRNA pathway genes (Pearson correlation test, $r = -0.38$, P-value $= 0.530$) or between TE expression and the activity of immune genes (Pearson correlation test, $r = 0.04$, P-value $= 0.953$).

DISCUSSION

Advantages of TETOOLS

In this manuscript, we present a new analysis pipeline dedicated to the analysis of TE expression for both messenger and small RNAs. Contrary to previous approaches, this method places emphasis on the TE copies rather than on consensus sequences. This approach allows us to consider more reads and thus to reduce the loss of information because we take into account reads mapping at several positions on the genome and the individual copy variability. Moreover, this pipeline uses raw counts as proposed by Anders and Huber (16), which is a less biased approach than other normalization methods used for RNA-seq data. The pipeline also allows the use of various types of mapper and expression analysis software. In the current version we use bowtie/bowtie2 and DESeq2, but the use of alternative programs is also possible.

TETOOLS relies on DESeq2 for the differential expression analysis, which works well when the differentially expressed sequences account for a small amount of the total number of reads. All other differential expression programs available to date behave the same way. DESeq2 first adjusts the geometric means of the read counts across samples. This approach is valid if the potential differences reflect differences in the sample sizes that are not biologically relevant. Therefore, our procedure is valuable for the majority of transcriptome studies in which a few TE families are differentially expressed. However, in very specific cases in which one sample could be expected to display higher expression levels of all TE families (and thus increased total numbers of TE reads), the DESeq2 approach will not be relevant because differences in the geometric means of the read counts will be expected to be biologically different. In such cases, we advise pooling the count files obtained for genes and TEs separately (we recommend using TECOUNT to obtain the read counts) and performing the differential expression analysis on the pooled count file. When we applied the latter procedure to the present data, the results were comparable to those obtained using TEDIFF on the TE reads alone (data not shown).

TE and gene expression exhibit strain differentiation but with specific dynamics

Gene transcription variation among species and populations has been previously described in *D. melanogaster* and *D. simulans* (56–58). Our study on *D. simulans* wild-type strains shows that variation in gene transcription is important and is sufficient to separate strains from the ancestral area (50) from strains from the derived areas.

Our data also suggest that genes that are differentially expressed between the ancestral and derived areas belong to functional categories linked to antennal morphogenesis, DNA repair, epigenetic modifications and eye morphogenesis. Some of these genes could be associated with specific different environments and could be linked to local adaptations, but further experiments are necessary to link expression levels to phenotypic features.

Previous works on TE dynamics showed that *D. simulans* strains harbored different numbers of TEs and different TE activities, suggesting that strains could be well distinguished based on TE dynamics (22,24,53,59). However, these previous studies were performed on a small scale. The present analysis allowed a genome-wide confirmation of these results. We find that the variability uncovered for TEs does not follow geographical patterns as strongly as genes. We propose that the regulation of TE expression evolves faster than the regulation of expression of the rest of the genome, thereby starting to erase more rapidly the geographical structures inherited from the worldwide colonization process. This faster evolution of TE expression regulation is consistent with the work by Song *et al.* (28), which showed that piRNA cluster expression was more variable

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on October 20, 2016

than protein-coding gene expression in 16 inbred lines of *D. melanogaster*.

These data also raise the question of the interaction between TEs and gene expression. Several decades ago, McClintock (2) and Britten (60) proposed that TEs participated in gene regulatory networks and provided regulatory regions; this finding was recently confirmed (61–63). More recently, TE insertions were shown to affect the chromatin structure of nearby genes via the spread of chromatin silencing marks (i.e. H3K9me3) that may affect gene expression (6,33,64). Considering that TE expression evolves faster than protein-coding gene expression and that TEs can contribute to the modulation of gene expression through epigenetic processes, then TEs appear to be potential fundamental actors of genome expression diversification and thus adaptation (65). Further studies are necessary to elucidate the interactions between TEs and gene expression in different genetic backgrounds in a genome-wide manner.

piRNA production is positively correlated with TE expression

Previous works on TE dynamics attempted to relate piRNA production to TE copy numbers (26,28,66) but found no significant correlation. A previous analysis of wild-type strains of *D. simulans* showed that TE copy numbers were not correlated with GIPP expression (31). Song *et al*. (28) found the same result for *D. melanogaster* inbred lines. Taking advantage of the present dataset, we tested whether piRNA production was related to TE expression instead of TE copy numbers. Indeed, only active (expressed) TE copies are the targets of piRNA inhibition. We find a significant positive correlation between piRNA production and TE expression. The most highly expressed TE families display the highest quantity of piRNAs and *vice versa*. This result is consistent with the work of Kelleher and Barbash (27), which was performed on two strains of *D. melanogaster*. However, this result concerns only TE families controlled in the germline by secondary piRNAs.

GIPP activity can explain TE activity

We found a strong negative correlation between GIPP activity and TE expression. This result indicates that TE expression is higher in strains in which effectors of the piRNA pathway are weakly transcribed and *vice versa*. This is a characteristics of the genome of each given strain. We have also shown in this work a positive correlation between TE transcription and piRNA production. This result reflects a property of TE families. Thus, the two above mentioned correlations are not incompatible but deal with different levels of variability. TE global activity varies between strains, inversely to the activity of the piRNA pathway. In addition, within the genome of each strain, at the TE family level, the production of piRNAs is positively correlated to the transcription level of TEs (Figure 6). This model can conciliate differences in copy numbers between strains that are not associated with piRNA pathway activity or piRNA production, since it considers the same evolutionary scale.

The negative correlation that we find between GIPP activity and TE expression fits perfectly with the Red Queen hypothesis (67): the pathogen/host relationship is embodied

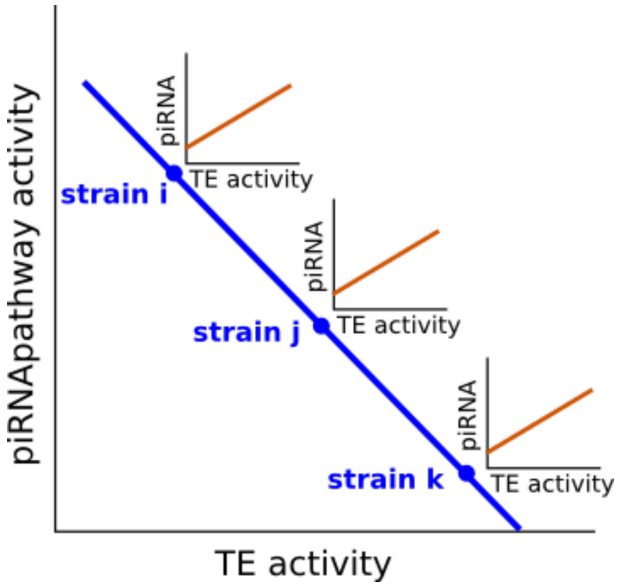


Figure 6. Proposed model to integrate the inside genome regulation of TEs and the strain differences in the TE transcript amounts. Each strain has a specific activity of TEs that is negatively associated with the piRNA pathway efficiency. At a different level, inside each genome strain the activity of TEs is positively associated with the production of piRNAs.

by the 'pathogenic' TEs and the piRNA pathway which acts as a genomic defense against them. We previously explored this issue, using TE copy number data and this did not allow us to find any correlation between TEs and GIPP activity (31). At that time, we proposed that the evolutionary time scales were not compatible because TE copy number includes recent as well as very ancient TE insertion events, whereas GIPP activity is highly dynamic on a short time scale. The transcriptomes that we analyzed here provide us with data from compatible evolutionary time scales and reveal a relationship between TEs and GIPPs. Therefore, TEs and GIPPs do appear to follow the same evolutionary dynamics and are involved in an antagonistic, rapidly evolving relationship. Natural variability in the GIPPs (31) may be envisioned as tightly linked to natural variability in TEs and their dynamics in natural strains (25,49). We believe this is a very strong result, which has to be considered in future evolutionary studies of TEs. We propose that this arms race may drive strain divergence and be implicated in the beginning of speciation.

ACCESSION NUMBERS

SRX1287831, SRX1287832, SRX1287833, SRX1287834, SRX1287843 and SRX1287860

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The work was performed using the computing facilities of the CC LBBE/PRABI and the galaxy.prabi.fr web service.

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on October 20, 2016

We thank P. Veber, S. Chambeyron and R. Rebollo for useful discussions, and A. Gibert, C. Goubert, N. Burlet and S. Martinez for technical assistance.

FUNDING

Agence Nationale de la Recherche [Exhyb ANR-14-CE19-0016-01 to C.V.]; Fondation pour la Recherche Médicale [DEP20131128536 to C.V.]; CNRS; Institut Universitaire de France (to C.V.). Funding for open access charge: Agence National de la Recherche.
Conflict of interest statement. None declared.

REFERENCES

1. Biémont,C. and Vieira,C. (2006) Genetics: junk DNA as an evolutionary force. *Nature*, **443**, 521–524.
2. McClintock,B. (1953) Induction of instability at selected loci in maize. *Genetics*, **38**, 579–599.
3. Kidwell,M.G. and Lisch,D.R. (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. *Evol. Int. J. Org. Evol.*, **55**, 1–24.
4. Lipatov,M., Lenkov,K., Petrov,D.A. and Bergman,C.M. (2005) Paucity of chimeric gene-transposable element transcripts in the Drosophila melanogaster genome. *BMC Biol.*, **3**, 24.
5. Deloger,M., Cavalli,F.M.G., Lerat,E., Biémont,C., Sagot,M.-F. and Vieira,C. (2009) Identification of expressed transposable element insertions in the sequenced genome of Drosophila melanogaster. *Gene*, **439**, 55–62.
6. Sienski,G., Dönertas,D. and Brennecke,J. (2012) Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell*, **151**, 964–980.
7. Graveley,B.R., Brooks,A.N., Carlson,J.W., Duff,M.O., Landolin,J.M., Yang,L., Artieri,C.G., van Baren,M.J., Boley,N., Booth,B.W. *et al.* (2011) The developmental transcriptome of Drosophila melanogaster. *Nature*, **471**, 473–479.
8. Aravin,A.A. and Hannon,G.J. (2008) Small RNA silencing pathways in germ and stem cells. *Cold Spring Harb. Symp. Quant. Biol.*, **73**, 283–290.
9. Vagin,V.V., Klenov,M.S., Kalmykova,A.I., Stolyarenko,A.D., Kotelnikov,R.N. and Gvozdev,V.A. (2004) The RNA interference proteins and vasa locus are involved in the silencing of retrotransposons in the female germline of Drosophila melanogaster. *RNA Biol.*, **1**, 54–58.
10. 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.
11. Saito,K. and Siomi,M.C. (2010) Small RNA-mediated quiescence of transposable elements in animals. *Dev. Cell*, **19**, 687–697.
12. Sienski,G., Batki,J., Senti,K.-A., Dönertas,D., Tirian,L., Meixner,K. and Brennecke,J. (2015) Silencio/CG9754 connects the Piwi-piRNA complex to the cellular heterochromatin machinery. *Genes Dev.*, **29**, 2258–2271.
13. Le Thomas,A., Rogers,A.K., Webster,A., Marinov,G.K., Liao,S.E., Perkins,E.M., Hur,J.K., Aravin,A.A. and Tóth,K.F. (2013) Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes Dev.*, **27**, 390–399.
14. Han,B.W., Wang,W., Zamore,P.D. and Weng,Z. (2015) piPipes: a set of pipelines for piRNA and transposon analysis via small RNA-seq, RNA-seq, degradome- and CAGE-seq, ChIP-seq and genomic DNA sequencing. *Bioinformatics*, **31**, 593–595.
15. Lerat,E., Burlet,N., Biémont,C. and Vieira,C. (2011) Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. *Gene*, **473**, 100–109.
16. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
17. Goubert,C., Modolo,L., Vieira,C., ValienteMoro,C., Mavingui,P. and Boulesteix,M. (2015) De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). *Genome Biol. Evol.*, **7**, 1192–1205.
18. Novák,P., Neumann,P., Pech,J., Steinhaisl,J. and Macas,J. (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792–793.
19. 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.
20. Giardine,B., Riemer,C., Hardison,R.C., Burhans,R., Elnitski,L., Shah,P., Zhang,Y., Blankenberg,D., Albert,I., Taylor,J. *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.
21. Vieira,C., Fablet,M., Lerat,E., Boulesteix,M., Rebollo,R., Burlet,N., Akkouche,A., Hubert,B., Mortada,H. and Biémont,C. (2012) A comparative analysis of the amounts and dynamics of transposable elements in natural populations of Drosophila melanogaster and Drosophila simulans. *J. Environ. Radioact.*, **113**, 83–86.
22. Fablet,M., McDonald,J.F., Biémont,C. and Vieira,C. (2006) Ongoing loss of the tirant transposable element in natural populations of Drosophila simulans. *Gene*, **375**, 54–62.
23. Mugnier,N., Biémont,C. and Vieira,C. (2005) New regulatory regions of Drosophila 412 retrotransposable element generated by recombination. *Mol. Biol. Evol.*, **22**, 747–757.
24. Rebollo,R., Horard,B., Begeot,F., Delattre,M., Gilson,E. and Vieira,C. (2012) A snapshot of histone modifications within transposable elements in Drosophila wild type strains. *PLoS ONE*, **7**, e44253.
25. 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.
26. Lu,J. and Clark,A.G. (2010) Population dynamics of PIWI-interacting RNAs (piRNAs) and their targets in Drosophila. *Genome Res.*, **20**, 212–227.
27. Kelleher,E.S. and Barbash,D.A. (2013) Analysis of piRNA-mediated silencing of active TEs in Drosophila melanogaster suggests limits on the evolution of host genome defense. *Mol. Biol. Evol.*, **30**, 1816–1829.
28. Song,J., Liu,J., Schnakenberg,S.L., Ha,H., Xing,J. and Chen,K.C. (2014) Variation in piRNA and transposable element content in strains of Drosophila melanogaster. *Genome Biol. Evol.*, **6**, 2786–2798.
29. Kolaczkowski,B., Hupalo,D.N. and Kern,A.D. (2011) Recurrent adaptation in RNA interference genes across the Drosophila phylogeny. *Mol. Biol. Evol.*, **28**, 1033–1042.
30. Obbard,D.J., Gordon,K.H.J., Buck,A.H. and Jiggins,F.M. (2009) The evolution of RNAi as a defence against viruses and transposable elements. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **364**, 99–115.
31. Fablet,M., Akkouche,A., Braman,V. and Vieira,C. (2014) Variable expression levels detected in the Drosophila effectors of piRNA biogenesis. *Gene*, **537**, 149–153.
32. Grentzinger,T. and Chambeyron,S. (2014) Fast and accurate method to purify small noncoding RNAs from Drosophila ovaries. *Methods Mol. Biol.*, **1093**, 171–182.
33. Akkouche,A., Grentzinger,T., Fablet,M., Armenise,C., Burlet,N., Braman,V., Chambeyron,S. and Vieira,C. (2013) Maternally deposited germline piRNAs silence the tirant retrotransposon in somatic cells. *EMBO Rep.*, **14**, 458–464.
34. Modolo,L. and Lerat,E. (2015) UrQt: an efficient software for the Unsupervised Quality trimming of NGS data. *BMC Bioinformatics*, **16**, 137.
35. Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.
36. Roberts,A., Trapnell,C., Donaghey,J., Rinn,J.L. and Pachter,L. (2011) Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.*, **12**, R22.
37. Drosophila 12 Genomes Consortium, Clark,A.G., Eisen,M.B., Smith,D.R., Bergman,C.M., Oliver,B., Markow,T.A., Kaufman,T.C., Kellis,M., Gelbart,W. *et al.* (2007) Evolution of genes and genomes on the drosophila phylogeny. *Nature*, **450**, 203–218.
38. Hu,T.T., Eisen,M.B., Thornton,K.R. and 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.
39. Smit,A., Hubley,R. and Green,P. (2013) RepeatMasker Open-4.0.

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on October 20, 2016

40. Bailly-Bechet,M., Haudry,A. and Lerat,E. (2014) 'One code to find them all': a perl tool to conveniently parse RepeatMasker output files. *Mob. DNA*, **5**, 13.
41. Antoniewski,C. (2014) Computing siRNA and piRNA overlap signatures. *Methods Mol. Biol.*, **1173**, 135–146.
42. 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.
43. Caboche,S., Audebert,C., Lemoine,Y. and Hot,D. (2014) Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. *BMC Genomics*, **15**, 264.
44. Dillies,M.-A., Rau,A., Aubert,J., Hennequet-Antier,C., Jeanmougin,M., Servant,N., Keime,C., Marot,G., Castel,D., Estelle,J. *et al.* (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.*, **14**, 671–683.
45. R Core Team. (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
46. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
47. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
48. Rebollo,R., Lerat,E., Kleine,L.L., Biémont,C. and Vieira,C. (2008) Losing helena: the extinction of a drosophila line-like element. *BMC Genomics*, **9**, 149.
49. Akkouche,A., Rebollo,R., Burlet,N., Esnault,C., Martinez,S., Viginier,B., Terzian,C., Vieira,C. and Fablet,M. (2012) tirant, a newly discovered active endogenous retrovirus in Drosophila simulans. *J. Virol.*, **86**, 3675–3681.
50. Lachaise,D. and Silvain,J.-F. (2004) How two Afrotropical endemics made two cosmopolitan human commensals: the Drosophila melanogaster-D. simulans palaeogeographic riddle. *Genetica*, **120**, 17–39.
51. Kaminker,J.S., Bergman,C.M., Kronmiller,B., Carlson,J., Svirskas,R., Patel,S., Frise,E., Wheeler,D.A., Lewis,S.E., Rubin,G.M. *et al.* (2002) The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol.*, **3**, RESEARCH0084.
52. Lerat,E., Rizzon,C. and Biémont,C. (2003) Sequence divergence within transposable element families in the Drosophila melanogaster genome. *Genome Res.*, **13**, 1889–1896.
53. Vieira,C. and Biémont,C. (1996) Geographical variation in insertion site number of retrotransposon 412 in Drosophila simulans. *J. Mol. Evol.*, **42**, 443–451.
54. Vieira,C., Piganeau,G. and Biémont,C. (2000) High copy numbers of multiple transposable element families in an Australian population of Drosophila simulans. *Genet. Res.*, **76**, 117–119.
55. Fablet,M., Lerat,E., Rebollo,R., Horard,B., Burlet,N., Martinez,S., Brasset,E., Gilson,E., Vaury,C. and Vieira,C. (2009) Genomic environment influences the dynamics of the tirant LTR retrotransposon in Drosophila. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.*, **23**, 1482–1489.
56. Zhao,L., Wit,J., Svetec,N. and Begun,D.J. (2015) Parallel Gene Expression Differences between Low and High Latitude Populations of Drosophila melanogaster and D. simulans. *PLoS Genet.*, **11**, e1005184.
57. Müller,L., Hutter,S., Stamboliyska,R., Saminadin-Peter,S.S., Stephan,W. and Parsch,J. (2011) Population transcriptomics of Drosophila melanogaster females. *BMC Genomics*, **12**, 81.
58. Lee,Y.C.G. (2015) The role of piRNA-mediated epigenetic silencing in the population dynamics of transposable elements in Drosophila melanogaster. *PLoS Genet.*, **11**, e1005269.
59. Biémont,C., Nardon,C., Deceliere,G., Lepetit,D., Loevenbruck,C. and Vieira,C. (2003) Worldwide distribution of transposable element copy number in natural populations of Drosophila simulans. *Evol. Int. J. Org. Evol.*, **57**, 159–167.
60. Britten,R.J. (1996) DNA sequence insertion and evolutionary variation in gene regulation. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 9374–9377.
61. Casacuberta,E. and González,J. (2013) The impact of transposable elements in environmental adaptation. *Mol. Ecol.*, **22**, 1503–1517.
62. Rebollo,R., Romanish,M.T. and Mager,D.L. (2012) Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu. Rev. Genet.*, **46**, 21–42.
63. Feschotte,C. (2008) Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.*, **9**, 397–405.
64. Shpiz,S., Ryazansky,S., Olovnikov,I., Abramov,Y. and Kalmykova,A. (2014) Euchromatic transposon insertions trigger production of novel Pi- and endo-siRNAs at the target sites in the drosophila germline. *PLoS Genet.*, **10**, e1004138.
65. Fablet,M. and Vieira,C. (2011) Evolvability, epigenetics and transposable elements. *Biomol. Concepts*, **2**, 333–341.
66. Castillo,D.M., Mell,J.C., Box,K.S. and Blumenstiel,J.P. (2011) Molecular evolution under increasing transposable element burden in Drosophila: a speed limit on the evolutionary arms race. *BMC Evol. Biol.*, **11**, 258.
67. Liow,L.H., Van Valen,L. and Stenseth,NC. (2011) Red Queen: from populations to taxa and communities. *Trends Ecol. Evol.*, **26**, 349–358.

Downloaded from http://nar.oxfordjournals.org/ at UCBL SCD Lyon 1 on October 20, 2016