L'annotation des éléments transposables par la compréhension de leur diversification

Timothée Flutre

THÈSE DE DOCTORAT DE L'UNIVERSITÉ PARIS DIDEROT

École doctorale « Interdisciplinaire Européenne Frontières du Vivant »

Présentée par

TIMOTHÉE FLUTRE

pour l'obtention du grade de

Docteur de l'Université Paris Diderot

L'ANNOTATION DES ÉLÉMENTS TRANSPOSABLES PAR LA COMPRÉHENSION DE LEUR DIVERSIFICATION

Soutenue le 28 octobre 2010 devant le jury composé de :

Rapporteur	OLIVIER PANAUD	Professeur, Université de Perpignan
Rapporteur	PIERRE ROUZÉ	Professeur, Université de Gand / VIB
Examinateur	THOMAS WICKER	Researcher, University of Zurich
Examinateur	KHASHAYAR PAKDAMAN	Professeur, Université Paris Diderot / IJM
Directeur de thèse	HADI QUESNEVILLE	Directeur de recherche INRA, Versailles
Directeur de thèse	CATHERINE FEUILLET	Directeur de recherche INRA, Clermont-Ferrand

Unité de Recherche en Génomique et Informatique
Centre INRA de Versailles-Grignon, Route de Saint-Cyr, 78026 Versailles

Comprendre, c'est unifier.

Albert Camus

REMERCIEMENTS

Je remercie vivement Olivier Panaud, Pierre Rouzé, Thomas Wicker et Khashayar Pakdaman pour avoir accepté de faire partie du jury de cette thèse. Je leur en suis très reconnaissant.

Je remercie profondément Hadi Quesneville pour avoir ouvert sa porte à un étudiant, somme toute très inexpérimenté, venu le voir à l'improviste au beau milieu du mois de juillet 2006. J'ai eu l'opportunité grâce à lui de me plonger dans ce qui m'attirait alors le plus en biologie, l'étude de l'évolution des génomes, et ceci par le biais d'entités singulièrement attirantes, les éléments transposables.

Je remercie sincèrement Catherine Feuillet pour avoir accepté de m'encadrer pendant ma thèse et pour ses conseils toujours si avisés. Avoir en toile de fond l'exemple du blé avec son génome six fois plus grand que celui de l'homme m'a permis de garder en tête le lien fort qui, à l'INRA, unit la recherche fondamentale et ses applications agronomiques.

Je remercie également Hugues Roest Crollius et Sébastien Aubourg pour avoir accepté de faire partie de mon comité de thèse, et donc pour leurs conseils et leur écoute tout au long de ces trois ans.

Je remercie également Dmitri Petrov pour m'avoir accueilli un mois dans son laboratoire et m'avoir permis ainsi de découvrir la bouillonnante vie scientifique californienne.

Tout au long de mes trois années de thèse, j'ai eu l'occasion de travailler avec de nombreuses personnes. Je ne saurai trop les remercier pour leur aide, leurs encouragements, leurs rires et leur patience à mon égard. Ce sont elles surtout qui m'ont fait apprécier le quotidien de la recherche. Un grand merci donc à Anna-Sophie Fiston-Lavier, Emmanuelle Permal, Olivier Inizan, Michael Alaux, Isabelle Luyten, Philippe Leroy, Joëlle Amselem, Nacer Mohellibi et Matthias Zytnicki, ainsi qu'à tous les membres du laboratoire qui ont contribué à son atmosphère chaleureuse.

Tardivement et petit à petit, par des chemins détournés, au gré de rencontres et d'expériences en France et à l'étranger, j'en suis arrivé à m'intéresser sérieusement à la recherche scientifique. J'ai

alors eu la chance de découvrir un lieu improbable peuplé d'enthousiastes chercheurs aux idées plein la tête. Je remercie donc tous les membres du Centre de Recherche Interdisciplinaire, en particulier François Taddéi et Ariel Lindner, ainsi que tous ceux qui contribuent activement à l'existence d'un tel lieu. Ils ont souvent rendu possible la réalisation de certains projets un peu fous, notamment le premier symposium interdisciplinaire et international de Paris organisé par et pour les doctorants.

Dans le même esprit et souvent avec les mêmes personnes, mon activité de recherche n'aurait pas eu la même saveur sans l'association Paris Montagne et son indéfinissable dynamisme. Nous y avons appris ensemble à rechercher, à concilier, exigence et plaisir, innovation et solidarité.

Enfin, puissent mes proches, famille et amis, ressentir ici mon émotion au souvenir des discussions et regards, repas et voyages, que nous avons eu en commun. En partageant réflexions et révoltes, en jouissant côte-à-côte de l'inattendu et du paisible, ils m'ont offert plus que je ne saurai leur rendre.

RÉSUMÉ

Tout organisme vivant est le produit d'interactions complexes entre son génome et son environnement, interactions caractérisées par des échanges de matière et d'énergie indispensables à la survie de l'organisme et la transmission de son génome. Depuis la découverte dans les années 1910 que le chromosome est le support de l'information génétique, les biologistes étudient les génomes afin de décrypter les mécanismes et processus à l'oeuvre dans le développement des organismes et l'évolution des populations. Grâce aux améliorations technologiques des dernières décennies, plusieurs génomes ont été entièrement séquencés, leur nombre s'accroissant rapidement, mais ils sont loin d'être décryptés pour autant. En effet, certains de leurs composants, les éléments transposables, sont encore mal compris, bien qu'ils aient été détectés chez quasiment toutes les espèces étudiées, et qu'ils puissent représenter jusqu'à 90% du contenu total de leurs génomes.

Les éléments transposables sont des fragments du génome possédant la particularité d'être mobiles. Ils ont donc un impact majeur sur la structure des génomes mais également sur l'expression des gènes avoisinants, notamment via des mécanismes épigénétiques. Leur évolution est aussi particulière étant donné qu'ils ont une transmission verticale non-mendélienne et que de nombreux cas de transferts horizontaux ont été mis en évidence. Mais, à part dans le cas de certains organismes modèles pour lesquels nous disposons de séquences de référence, l'annotation des éléments transposables représente souvent un goulot d'étranglement dans l'analyse des séquences génomiques. A cela s'ajoute le fait que les études de génomique comparée montrent que les génomes sont bien plus dynamiques qu'on ne le croyait, en particulier ceux des plantes, ce qui complique d'autant l'annotation précise des éléments transposables.

Pendant mes travaux de thèse, j'ai commencé par comparer les programmes informatiques existants utilisés dans les approches d'annotation *de novo* des éléments transposables. Pour cela, j'ai mis au point un protocole de test sur les génomes de *Drosophila melanogaster* et *Arabidopsis thaliana*. Ceci m'a permis de proposer une approche *de novo* combinant plusieurs outils, capable ainsi de reconstruire automatiquement un grand nombre de sé-

quences de référence. De plus, j'ai pu montrer que notre approche mettait en évidence les variations structurales au sein de familles bien connues, notamment en distinguant des variants structuraux appartenant à une même famille d'éléments transposables, reflétant ainsi la diversification de ces familles au cours de leur évolution.

Cette approche a été implémentée dans une suite d'outils (REPET) rendant possible l'analyse des éléments transposables de nombreux génomes de plantes, insectes, champignons et autres. Ces travaux ont abouti à une feuille de route décrivant de manière pratique comment annoter le contenu en éléments transposables de tout génome nouvellement séquencé. Par conséquent, de nombreuses questions concernant l'impact de ces éléments sur l'évolution de la structure des génomes peuvent maintenant être abordées chez différents génomes plus ou moins proches. Je propose également plusieurs pistes de recherche, notamment la simulation des données nécessaires à l'amélioration des algorithmes de détection, démarche complémentaire de la modélisation de la dynamique des éléments transposables.

Mots clés : génomique, bioinformatique, élément transposable, annotation, variation structurale

ABSTRACT

Any living organism is the result of complex interactions between its genome and its environment, interactions characterized by transfers of matter and energy required for the survival of the organism and the transmission of its genome. Since the discovery in the years 1910 that the chromosome is the mechanical basis of the genetic information, the biologists study genomes in order to decipher the mecanisms and processes operating in the development of organisms and the evolution of populations. Thanks to the technological improvements of the last decades, several genomes were fully sequenced, their number increasing quickly, but they are far from being deciphered. Indeed, some of their components, the transposable elements, are still not well understood, although they were detected in almost every species studied so far, and they can account for up to 90% of their genome.

Transposable elements are DNA sequences that can move and duplicate within genomes. They hence have a major impact on genome structure but also on the expression of neighbouring genes, notably via epigenetic mechanisms. Their evolution is also peculiar as they have a non-mendelian vertical transmission and as numerous cases of horizontal transfers were highlighted. However, except for some model organisms for which reference sequences are available, the annotation of transposable elements often corresponds to a bottleneck in the analysis of genomic sequences. Moreover, comparative genomics studies have shown that genomes are much more dynamic than previously expected, particularly in plants, thus making even more difficult the precise annotation of transposable elements.

During my PhD work, I started by comparing existing computer programs used in *de novo* approaches of transposable element identification. In this aim, I designed a test protocol on the genomes of *Drosophila melanogaster* and *Arabidopsis thaliana*. As a result, I proposed a *de novo* approach combining several tools, thus enabling the automatic recovery of a great number of reference sequences. Moreover, I showed that our approach highlighted the structural variations present within well-known families, notably by distinguishing structural variants belonging

to a same family of transposable elements, thus reflecting the diversification of such families during their evolution.

This approach was implemented in a package (REPET) making possible the analysis of transposable elements in numerous genomes from plants, insects and fungi among others. This work lead to a roadmap describing, from a practical point of view, how to annotate the transposable element content of any newly sequenced genome. As a consequence, many questions about the impact of these elements on the evolution of genome structure can now be tackled using several genomes more or less related withe ach other. I also propose several perspectives, notably the simulation of the data required for the improvement of the tools, a way complementary to the modeling of transposable element dynamics.

Keywords : genomics, bioinformatics, transposable element, annotation, structural variation

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

ABRÉVIATIONS

- ADN : acide désoxyribonucléique
- ARN : acide ribonucléique
- BAC : bacterial artifical chromosome
- BDGP : Berkeley Drosophila Genome Project
- ET : élément transposable
- indel : insertion-délétion
- ISBP : insertion site-based polymorphism
- LINE : long interspersed nuclear element
- LTR : long terminal repeat
- MITE : miniature inverted element
- ORF : open reading frame
- pb/kb/Gb : paire de bases, kilo-bases, giga-bases
- PBS : primer binding site
- PPT : poly-purine tract
- RIP : repeat-induced point mutation
- SINE : short interspersed nuclear element
- TIR : terminal inverted repeat
- TSD : target site duplication

GLOSSAIRE

- Un fragment d'ET est une région du génome issue de la transposition d'un ET. Il peut être identifié par un alignement entre une séquence de référence représentant une famille d'ET et une séquence génomique.
- Une copie d'ET peut être composée d'un ou plusieurs fragments. Une telle configuration est fréquente étant donné que les ET s'insèrent souvent les uns dans les autres et peuvent subir de courtes délétions. Le chainage, ou *défragmentation*, consiste à reconstruire une copie en connectant les fragments la composant.
- Une famille d'ET est un ensemble de copies ayant une copie ancestrale en commun. Elle est traditionnellement représentée par une seule séquence qui peut être une copie complète ou bien une séquence consensus formée à partir de plusieurs copies.

Première partie

INTRODUCTION

1

INTRODUCTION

Un élément transposable (ET) est traditionnellement décrit comme étant un fragment d'ADN possédant la particularité d'être mobile. Un tel élément est donc capable de s'insérer en d'autres endroits du génome, lui permettant ainsi de se multiplier. Ces caractéristiques en font des composants atypiques des génomes. Leur découverte a été à l'origine de fructueuses réflexions théoriques en biologie, de la régulation de l'expression des gènes jusqu'au niveau auquel la sélection naturelle est supposée agir. Pour autant, leur dynamique et leurs impacts sur les génomes ne sont toujours pas bien compris. Je tente, dans cette introduction, de décrire les développements historiques de la recherche sur les ET en mettant en relief leurs relations parfois conflictuelles avec d'autres domaines de la biologie. Ceci me permettra de poser le cadre conceptuel sur lequel mon travail repose, et finalement d'introduire mes recherches dans le contexte actuel.

1.1 DE LA DÉCOUVERTE DES ÉLÉMENTS TRANSPOSABLES À LEUR DESCRIPTION COMME PARASITES INTRAGÉNOMIQUES

1.1.1 *Découverte des éléments transposables*

Au début de l'année 1944, alors que les forces alliées s'enlisaient en Italie dans la bataille du Monte Cassino, des résultats expérimentaux furent publiés, prouvant que la substance chimique induisant la transformation spécifique de pneumocoques correspond à de l'acide désoxyribonucléique (Avery et coll. 1944). Comme l'ont noté les auteurs :

> *It is particularly significant in the case of pneumococci that the experimentally induced alterations are definitely correlated with the development of a new morphological structure and the consequent acquisition of new antigenic and invasive properties. Equally if not more significant is the fact that these changes are predictable, type-specific, and heritable.*

Ils apportaient la première preuve expérimentale que l'ADN est le support matériel de l'hérédité. Il a cependant fallu attendre 1952 et l'expérience de Hershey et Chase pour convaincre la communauté scientifique de cette conclusion. L'année suivante, en 1953, Watson et Crick publiaient la structure de l'ADN sous forme de double hélice (Watson et Crick 1953) et faisaient négligemment remarquer l'importance de leur découverte en précisant :

> *It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.*

Ce résultat entraîna le développement de la biologie moléculaire avec le succès que l'on sait (Morange 2000).

En 1944 également, lors d'un croisement visant à révéler la composition génique du bras court du chromosome 9 du maïs, Barbara McClintock déclencha ce que certains appelleront un tremblement de terre génétique (Jones 2005). Dans les années qui suivirent, elle décrivit précisément les réarrangements chromosomiques observés à l'issue de ce fameux croisement. Plus intriguant, elle découvrit des locus mutables ne restant pas toujours à la même position le long des chromosomes. Elle parvint à les caractériser, distinguant deux classes (McClintock 1950) :

> *(1) those that require a separate activator factor for instability to be expressed, and (2) those that are autonomous with respect to the factor that controls the onset of mutability.*

Elle découvrit ainsi les premiers ET qu'elle nomma Ac pour Activator et Ds pour Dissociation. Poursuivant ses investigations, McClintock proposa alors dès 1956 de décrire ces locus mutables comme étant des éléments contrôlant l'action des gènes (McClintock 1956) :

> *It is now known that controlling elements may modify gene action in a number of different ways. They may influence the time of gene action in the development of a tissue, and also determine the cells in which it will occur. Again, they may influence the type of action, with regards to either degree (quantitative aspects) or kind (qualitative aspects).*

Dans cette dernière publication, elle mentionne également l'impact des éléments contrôlants sur des gènes localisés à quelque distance de chaque côté *via* un effet de diffusion (*spreading*), ce

qui serait maintenant interprété comme étant de l'hétérochromatinisation.

Au milieu du vingtième siècle, alors que la stucture de l'ADN venait tout juste d'être découverte, des éléments mobiles ont été mis en évidence dans les génomes. Comment interpréter la présence de tels éléments ? Jouent-ils un rôle dans la façon dont l'information génétique est exprimée ?

1.1.2 *Régulation de l'expression des gènes*

Au début des années 1960, lorsque que Jacob et Monod proposèrent leur modèle de régulation de l'expression des gènes à partir de leurs travaux sur la bactérie *Escherichia coli* (Jacob et Monod 1961), McClintock y vit immédiatement un lien avec ce qu'elle avait découvert chez le maïs (McClintock 1961). Cependant, il s'avéra par la suite que les deux modèles impliquaient des composants bien différents, notamment des facteurs de transcription d'une part et des ET de l'autre. De plus, alors que les bactériologistes de l'Institut Pasteur décrivaient un véritable modèle approfondi, la généticienne du Cold Spring Harbor Laboratory cherchait plus à faire un parallèle entre ses observations et ce modèle. Pour autant, ces deux contributions furent mentionnées par Britten et Davidson à la fin des années 1960 lorsqu'ils proposèrent un modèle théorique visant à donner sens aux nombreuses données expérimentales accumulées chez les eucaryotes multicellulaires (Britten et Davidson 1969). Pour la première fois de façon aussi construite depuis leur découverte, un rôle majeur était donné aux séquences répétées dans la régulation de l'expression des gènes :

> *For bovine DNA (and probably that of other organisms) repeated sequences are intimately interspersed with non-repeated sequences, throughout the length of the genome. This is precisely the pattern required in our model if repeated sequences are usually or often regulatory in function.*

De plus, Britten et Davidson justifiaient la présence d'autant de séquences répétées dans les génomes de la manière suivante (Britten et Davidson 1971) :

> *Natural selection might promote the survival of populations capable of testing new regulatory relationships, thus maintaining the potentiality of producing new repetitive DNA sequences and translocating them.*

Bien qu'à cette époque, les idées sur les mécanismes d'amplification des ET et autres séquences répétées étaient encore trop peu claires, l'idée selon laquelle la sélection naturelle peut agir en vue d'avantages futurs, a longtemps posé problème aux évolutionnistes. C'est dans les années 1970-1980 que de nouveaux résultats permirent d'affiner notre compréhension de ces phénomènes. Par exemple, en croisant certaines souches de *D. melanogaster* maintenues en laboratoire depuis des décennies avec des souches sauvages, des chercheurs se sont aperçus que certains croisements n'étaient pas viables et présentaient de nombreuses aberrations (Kidwell et coll. 1977). De nombreuses études ont alors collectivement montré que ce phénomène était dû à une famille d'ET qui avait progressivement envahi les génomes des populations de *D. melanogaster* du monde entier en quelques décennies (Anxolabéhère et coll. 1988).

Parmi les premiers modèles proposées visant à expliquer la régulation de l'expression des gènes, plusieurs donnaient un rôle important aux ET, du fait de leur mobilité et de leur caractère répété. Mais la dynamique de ces éléments dans les génomes ne semble pas toujours être en accord avec le rôle qu'on leur prête.

1.1.3 *Théorie du gène égoïste*

De nombreuses études théoriques en génétique des populations ont été entreprises, modélisant l'action des différentes forces évolutives sur la dynamique des ET. Elles recherchaient les conditions dans lesquelles ceux-ci pouvaient se maintenir dans les génomes, voire les envahir (Charlesworth et Charlesworth 1983). Cela faisait alors quelques années que certains évolutionnistes revendiquaient une interprétation de l'évolution centrée sur les gènes : la théorie du *gène égoïste* (Hamilton 1963, Williams 1966). Celle-ci propose que la sélection naturelle agit à travers la survie différentielle de gènes en compétition, augmentant la fréquence des allèles dont les effets phénotypiques assurent avec succès leur propagation (Dawkins 1982). Dans un tel cadre, les ET représentent l'archétype du gène égoïste, pouvant même être décrits comme étant les parasites ultimes (Orgel et Crick 1980) :

> *The existence of selfish DNA is possible because DNA is a molecule which is replicated very easily and because selfish DNA occurs in an environment in which DNA replication*

> *is a necessity. It thus has the opportunity of subverting these essential mechanisms to its own purpose.*

A cette époque, le consensus n'était donc plus de savoir si les ET étaient fonctionnels comme McClintock ou Britten et Davidson l'avaient imaginé, mais plutôt de savoir comment le reste du génome arrivait à s'en débarrasser. Le terme de *junk DNA* est ainsi devenu habituel pour parler des ET (Ohno 1972).

La question du rôle possible de certains ET dans les génomes était restée en suspens. Certains évolutionnistes ont alors critiqué une utilisation excessive de cette théorie, les phénotypes d'un organisme étant tous vus comme des adaptations (Gould et Lewontin 1979). De cette intervention émergea un nouveau concept, celui d'*exaptation*. Il qualifie les traits ayant initialement été sélectionnés pour une fonction donnée, voire ayant évolué sans être soumis à une quelconque pression de sélection, mais ayant ensuite été cooptés pour une autre fonction, l'actuelle (Gould et Vrba 1982). Le terme adaptation est donc maintenant réservé aux phénotypes qui, eux, ont été sélectionnés au fil du temps pour une fonction qui est restée la même. Les auteurs à l'origine de ce concept d'exaptation ont ainsi ré-interprété le cas des séquences répétées dans les génomes :

> *Such selfish DNA may be playing its own Darwinian game at the genic level, but it represents a true nonaptation at the level of the phenotype. When used to great advantage in that future, these repeated copies are exaptations.*

En 50 ans, les ET ont amené les biologistes à reconsidérer profondément la théorie de l'évolution et à les associer à plusieurs reprises à des modèles spéculatifs sur la régulation de l'expression des gènes. Dans les années 1990, le séquençage de génomes complets est devenu une réalité, permettant ainsi d'analyser l'ensemble des ET dans les génomes. Ces études ont contribué à améliorer notre compréhension de leur biologie, et ainsi de l'évolution et du fonctionnement des génomes.

1.2 ANNOTATION ET ANALYSES DES ÉLÉMENTS TRANSPOSABLES DANS LES GÉNOMES MODÈLES

1.2.1 *Caractérisation des propriétés moléculaires des éléments transposables*

Depuis la découverte des ET et leur reconnaissance comme composants majeurs des génomes, de nombreuses expériences ont été menées afin de caractériser leurs mécanismes de transposition et leur structure moléculaire. Elles ont principalement été entreprises sur des organismes modèles comme la levure *Saccharomyces cerevisiae* (Boeke et coll. 1985) et la mouche *Drosophila melanogaster* (O'Hare et Rubin 1983). Elles ont notamment permis de distinguer deux classes d'éléments transposables, les premiers nommés rétrotransposons se multipliant *via* un intermédiaire à ARN par un mécanisme « copier-coller », et les deuxièmes nommés transposons se déplaçant par excision d'ADN par un mécanisme « couper-coller » (Finnegan 1989).

Grâce à leur capacité de reconnaître des motifs bien précis, de générer des coupures double-brins dans l'ADN à l'endroit même de ces motifs et d'insérer un fragment d'ADN dans un autre, les transposons ont été très tôt utilisés comme vecteurs de transformation génétique (Rubin et Spradling 1982). Les biologistes en ont fait des outils moléculaires d'ingéniérie génétique leur permettant d'intégrer un fragment d'ADN, comme un gène d'intérêt, dans le génome de l'espèce qu'ils étudient. En pratique, ils construisent deux plasmides, l'un portant le fragment à insérer, entouré de TIR (*terminal inverted repeats*) ainsi qu'un gène rapporteur, et l'autre ne portant que la transposase censée reconnaître les TIR en question (voir la figure 1 pour plus de détails).

Plusieurs familles d'ET sont utilisées dans un but de transformation génétique, toutes ne fonctionnant pas forcément dans les mêmes espèces, ou n'ayant pas les mêmes biais d'insertion (Thibault et coll. 2004). Une famille en particulier, *Sleeping Beauty*, a été reconstruite synthétiquement à partir de copies inactives (Ivics et coll. 1997), et est maintenant particulièrement utilisée chez les mammifères. Chaque étape du processus a été aujourd'hui considérablement étudié, permettant ainsi d'envisager des tests cliniques de thérapie génique à moyen terme (Izsvák et coll. 2010).

L'étude des propriétés biochimiques des ET a permis de mieux comprendre les mécanismes de transposition, fournis-

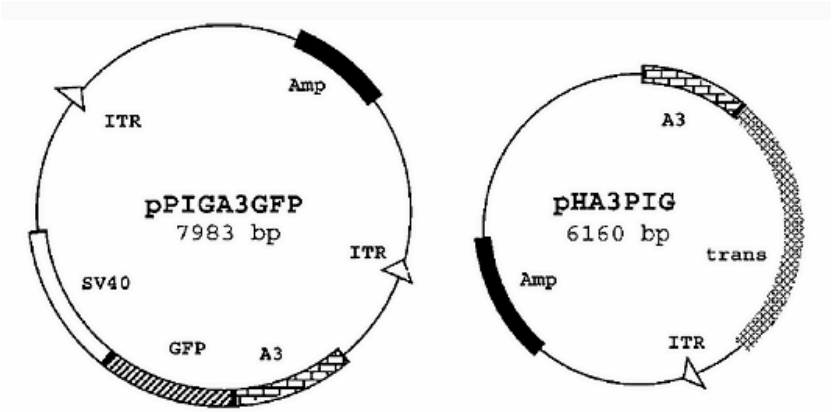


FIG. 1: Organisation de plasmides dérivés de l'ET piggyBac. Le plasmide « *marker* » pPIGA3GFP contient le promoteur du gène de l'actine A3 fusionné à la séquence codant pour la GFP (*green fluorescent protein*) et à la séquence non-traduite 3′ du SV40 (*simian virus 40*). Le plasmide « *helper* » pHA3PIG est la source de la transposase (*trans*) de l'ET piggyBac. ITR, répétitions terminales inversées ; Amp, gène de résistance à l'ampicilline. (Tiré de Tamura et coll. 2000.)

sant ainsi de puissants outils de biologie moléculaire. En même temps, d'autres études ont montré la présence des ET dans quasiment tous les génomes étudiés, et se sont donc attachées, autant que faire se peut, à en décrire la diversité de manière exhaustive, en premier lieu chez *Homo sapiens*.

1.2.2 *Cas du génome humain*

C'est en 2001 avec la séquence d'un génome humain que, pour la première fois, une étude de l'ensemble du contenu en ET d'un génome d'eucaryote supérieur a été réalisée (the International Human Genome Sequencing Consortium 2001). Cette étude a nécessité pour l'annotation des ET, la construction d'une banque de séquences connues d'ET, ainsi qu'une méthode de recherche adaptée basée sur l'alignement de séquences. Les auteurs de l'étude ont mis à profit la banque de séquences RepBase qui tentait de rassembler les séquences d'ET découvertes jusqu'alors. Certaines d'entre elles correspondaient à de vraies copies d'ET identifées expérimentalement, par exemple *via* l'analyse de gènes inactivés (Dombroski et coll. 1991). D'autres correspondaient à des séquences consensus construites à partir de l'analyse informatique de fragments

génomiques (Smit 1993). Pour la recherche des ET dans les séquences génomiques, plusieurs méthodes ont été développées, utilisant des algorithmes d'alignement permettant d'aligner rapidement et efficacement des séquences d'ET avec de grandes séquences génomiques. Ainsi, le programme REPEATMASKER (`http://www.repeatmasker.org/`) utilise des matrices de scores différentes selon le taux de GC du fragment génomique considéré. Suivant une démarche similaire, l'annotation du génome de *D. melanogaster* a bénéficié de la banque du *Berkeley Drosophila Genome Project* (BDGP) rassemblant toutes les séquences d'ET appartenant à cet organisme et découvertes jusqu'alors (Kaminker et coll. 2002).

Dans le cas du génome humain, les auteurs ont pu confirmer qu'une grande partie du génome était composé d'ET, environ 45% . En effet, des analyses de dissociation-réassociation thermique d'ADN avaient déjà montré que le génome humain est composé pour une grande partie de répétitions (Britten et Davidson 1971). De plus, ils ont montré qu'une grande majorité des ET correspondent à des éléments non-autonomes ne pouvant se déplacer par leurs propres moyens. Cette annotation a également rendu possible la reconstruction de la dynamique des ET dans ce génome (figure 2).

En un petit peu plus d'une décennie, un profond renversement s'est ainsi opéré. Alors qu'auparavant les moyens technologiques ne permettaient d'étudier principalement que certains composants du génome pris séparément les uns des autres, l'essor de la génomique a offert la possibilité d'étudier les génomes dans leur globalité. C'est bien l'architecture des génomes, ceux-ci étant considérés en tant qu'entités à part entière, qui devient investigable. Les idées imaginées au cours du développement de la génétique moléculaire peuvent maintenant être précisées et les hypothèses testées à l'échelle du génome entier.

Ainsi, il devient clair que la dynamique de la grande majorité des ET dans les génomes eucaryotes est le résultat d'un conflit intra-génomique avec le reste du génome comme décrit par la théorie du gène égoïste (Bestor 1999). La répression épigénétique par méthylation (Tsukahara et coll. 2009) ainsi que la découverte de mécanismes d'interférence à ARN servant à empêcher les ET de transposer (Aravin et coll. 2007) sont proposés comme les mécanismes moléculaires émergeant de ce conflit. Bien que la théorie de Britten et Davidson soit loin d'être vérifiée à la lettre, l'idée selon laquelle une même séquence, répétée à différents en-

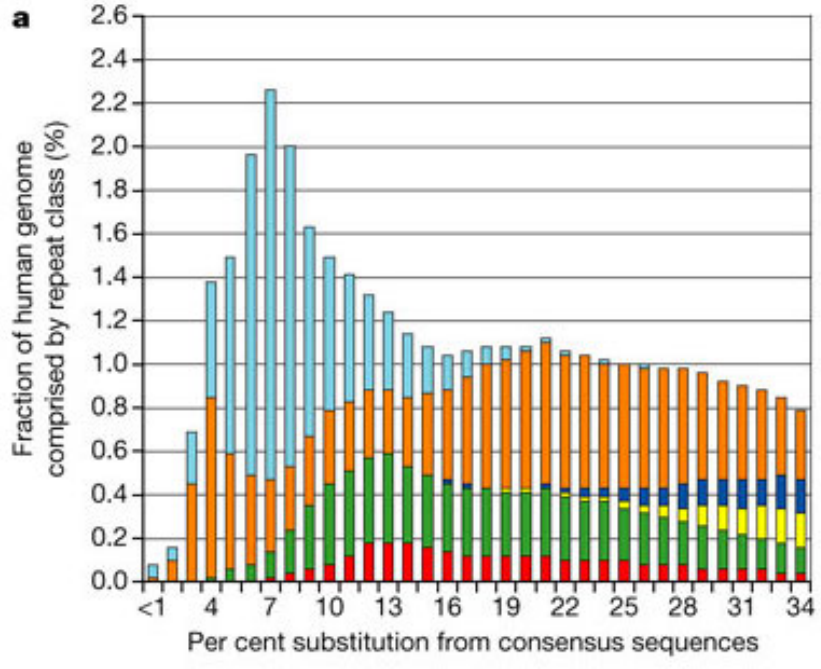


FIG. 2: Distribution de l'âge des ET dans le génome humain. Chaque couleur représente un type d'ET : les Alu en bleu clair, les L1 en orange, les MIR en bleu foncé, les L2 en jaune, les rétrotransposons à LTR en vert et les transposons à ADN en rouge. (Tiré de the International Human Genome Sequencing Consortium 2001.)

droits du génome, puisse participer à la régulation d'un réseau de gènes, se trouve elle aussi remise sur le devant de la scène par de récents résultats (Bourque et coll. 2008). L'idée de François Jacob de décrire l'évolution comme étant du bricolage moléculaire, trouve plusieurs exemples parmi les gènes dérivés d'ET (Jacob 1977) :

> *Evolution does not produce novelties from scratch. It works on what already exists, either transforming a system to give it new functions or combining several systems to produce a more elaborate one.*

Cette idée, ainsi que le concept d'exaptation proposé par Gould et Vrba, se voient élégamment illustrés par certains ET ayant été utilisés, sélectionnés, et donc domestiqués au cours de l'évolution pour des fonctions participant à la survie (Zhou et coll. 2004) et la reproduction de l'hôte (Mallet et coll. 2004).

L'annotation des ET dans les génomes modèles, notamment chez l'homme, a permis d'étudier en détail la contribution de

ces éléments à la biologie des génomes. Mais comment aborder l'analyse des génomes d'autres espèces ?

1.2.3 *Besoin d'approches de novo pour l'annotation des éléments transposables*

Il est vite apparu que les ET, du fait de leur nature répétée, posaient des problèmes spécifiques, tant au niveau du séquençage des génomes, qu'au niveau de leur assemblage et de leur annotation. Depuis les années 1980, un nouveau champ disciplinaire s'était développé parallèlement à l'amélioration des technologies de séquençage : la bioinformatique. Ainsi, différentes approches bioinformatiques pour l'annotation des ET ont été proposées pour aider à l'assemblage et à l'annotation des génomes (Bergman et Quesneville 2007). Dans le cas simple où l'on dispose de séquences déjà connues (comme par exemple pour *H. sapiens* décrit précédemment), une étape importante a été franchie avec la mise au point d'un outil combinant efficacement plusieurs algorithmes afin de détecter les fragments d'ET, filtrer les faux-positifs et connecter les fragments appartenant à une même copie ancestrale (Quesneville et coll. 2005). Cet outil, nommé TEannot, a permis de ré-annoter le contenu en ET du génome des deux organismes modèles *D. melanogaster* (Quesneville et coll. 2005, Bergman et coll. 2006) et *A. thaliana* (Buisine et coll. 2008), et d'en faire ainsi des annotations de référence. Mais pour la grande majorité des autres génomes séquencés, nous ne disposons pas ou peu de séquences connues d'ET spécifiques de ces génomes. Des approches *de novo* d'annotation sont donc indispensables.

L'expression « approche *de novo* » signifie que l'identification des éléments transposables ne requiert, comme données initiales, que les séquences génomiques brutes, permettant ainsi de trouver des éléments spécifiques du génome étudié, souvent absents des banques de données. Ce type de procédure peut également être qualifié d'approche *ab initio* (Saha et coll. 2008). C'est pour la mise au point de l'algorithme lui-même que sont mises à profit les caractéristiques des ET. Certaines approches, par exemple, recherchent dans les génomes des structures proches de celles des rétrotransposons ayant des répétitions terminales de même orientation appelées LTR pour *long terminal repeat* (McCarthy et McDonald 2003). Ces approches basées sur la structure seront discutées dans le dernier chapitre (voir 3.2.2) mais il est important de noter dès à présent qu'elles visent en premier lieu à iden-

tifier des copies complètes d'ET satisfaisant les critères de l'algorithme, et ne sont performantes que si les éléments concernés ont une structure bien marquée. Le séquençage des premiers génomes d'organismes modèles a donc permis de vérifier et d'affiner les théories proposées dans les décennies précédentes. A présent, le séquençage de nombreux génomes pour lesquels le contenu en ET est loin d'être aussi bien connu requiert l'utilisation d'approches *de novo*.

Avec le séquençage de nombreux génomes complets, des approches *de novo* d'annotation d'ET deviennent de plus en plus nécessaires. Les méthodes doivent pouvoir identifier les familles présentes et construire automatiquement des séquences les représentant.

1.3 ÉTUDE DE LA DIVERSIFICATION DES ÉLÉMENTS TRANSPOSABLES ET DE LA DYNAMIQUE DES GÉNOMES PAR UNE APPROCHE *de novo*

1.3.1 *Difficultées posées par les réarrangements chromosomiques*

L'approche canonique d'annotation *de novo* des ET commence par une comparaison de la séquence du génome avec elle-même afin de repérer par alignement les répétitions plus ou moins dégénérées, celles-ci étant ensuite regroupées en fonction de leur similarité et de leur pourcentage de recouvrement. Finalement une séquence consensus est construite pour chaque groupe de répétitions et l'ensemble des séquences consensus est utilisé comme banque d'ET avec des outils tel TEannot pour annoter leurs copies dans le génome en question (Quesneville et coll. 2005). L'étape de regroupement des répétitions doit être spécifique des ET. Ceux-ci correspondent à des répétitions dispersées présentes généralement en un grand nombre de copies, par opposition à d'autres types de répétitions, telles les répétitions en tandem ou les duplications segmentaires. Plusieurs algorithmes ont été proposés pour tenter de regrouper correctement les répétitions dispersées dans les génomes (Bao et Eddy 2002, Quesneville et coll. 2003, Edgar et Myers 2005). La plupart de ces algorithmes et les programmes qui les implémentent avaient à l'origine surtout pour but d'identifier les fragments du génome appartenant à des ET, en particulier pour les masquer et faciliter l'annotation des gènes codants. Or ce but peut être atteint même si les séquences consensus correspondent à des ET tronqués, ce

qui est souvent le cas, les ET étant rapidement fragmentés après leur insertion dans un génome. Mais disposer de séquences d'ET tronquées empêche l'étude approfondie de la dynamique de ces éléments, et freine donc notre compréhension de l'évolution de la structure des génomes.

Depuis longtemps (Dobzhansky et Sturtevant 1938), mais de manière frappante avec l'essor de la génomique, de nombreuses analyses ont montré à quel point le contenu des génomes est dynamique. Comme l'ont fait remarquer certains auteurs après le séquençage du génome de l'homme et de la souris, les analyses comparées ont apporté de nouveaux éléments permettant une compréhension systémique de l'évolution des génomes à plusieurs échelles (Eichler et Sankoff 2003) :

> *Comparative analyses of complete genomes can provide a comprehensive view of large-scale changes in synteny, gene order, and regions of nonconservation while simultaneously affording exquisite molecular resolution at the level of single-base pair differences.*

Concernant les réarrangements chromosomiques de type inversion, translocation, fusion et fission de chromosomes, il devient par exemple possible de reconstruire de manière parcimonieuse l'ordre dans lequel un chromosome de souris pourrait s'être réarrangé par rapport à un chromosome humain (Pevzner et Tesler 2003). En termes de duplications segmentaires, l'étude du génome humain a été, et continue d'être, féconde autant sur le plan inter-espèce en le comparant avec des génomes de chimpanzés et macaques (Jiang et coll. 2007) que sur le plan intra-espèce en comparant le génome de plusieurs individus (Kidd et coll. 2008). Les réarrangements chromosomiques sont donc fréquents et sont souvent à l'origine de duplications de gènes perturbant leur fonctionnement. Mais tout cela n'explique pas pour autant leur origine, ce qui les déclenche. Pourtant, les points de cassure sont fréquemment associés à des ET, ce qui a permis de proposer des modèles pour leur formation (Fiston-Lavier et coll. 2007). Il n'est pas toujours clair si les ET sont véritablement à l'origine des réarrangements ou bien s'ils se sont simplement accumulés autour des points de cassure, a posteriori, par exemple à cause d'une instabilité physique de ces régions (Zhou et Mishra 2005). Toujours est-il que les ET ont la capacité de profondément transformer le paysage génomique (Gray 2000), du fait de pouvoir transposer, de s'accumuler dans certaines régions, ou simplement du fait d'être présents en de nombreuses copies dans le génome, favorisant ainsi les recombinaisons ectopiques.

Les études de génomique comparée ont montré à quel point les génomes étaient dynamiques. Or, les séquences d'ET elles-mêmes sont également très dynamiques.

1.3.2 *Classification et variations structurales*

Au fur et à mesure de l'analyse des génomes eucaryotes, de nouveaux types d'ET sont encore découverts comme notamment les Hélitrons qui transposent *via* un mécanisme de type *rolling circle* (Kapitonov et Jurka 2001), et les Mavericks/Polintons supposés transposer par excision et réplication (Feschotte et Pritham 2005, Kapitonov et Jurka 2006). Ceci a donc amené à reproposer une classification des différents types d'ET (Wicker et coll. 2007). Dans celle-ci, les deux classes identifiées par Finnegan en 1989 sont conservées mais elles se subdivisent maintenant en plusieurs ordres incorporant les nouveaux types d'ET découverts, chaque ordre se subdivisant en superfamilles (figure 3). Cependant, l'analyse de génomes de plus en plus nombreux et divers a surtout montré à quel point les ET étaient structurellement variables et a confirmé l'importance des éléments non-autonomes (Sabot et Schulman 2006, Yang et coll. 2009). Tout ET subit à un moment ou à un autre des mutations, substitutions et indels (insertions/délétions), les rendant incapables de produire eux-mêmes la machinerie protéique nécessaire à leur transposition. Pour autant, il peut continuer à transposer, en recrutant la machinerie protéique d'autres éléments autonomes. Continuant à transposer, il est dit *actif*. A cause de leur similarité de séquence, il a été proposé que les éléments autonomes et leurs parasites non-autonomes gardent la même classification, différant seulement au niveau de la sous-famille (Wicker et coll. 2007).

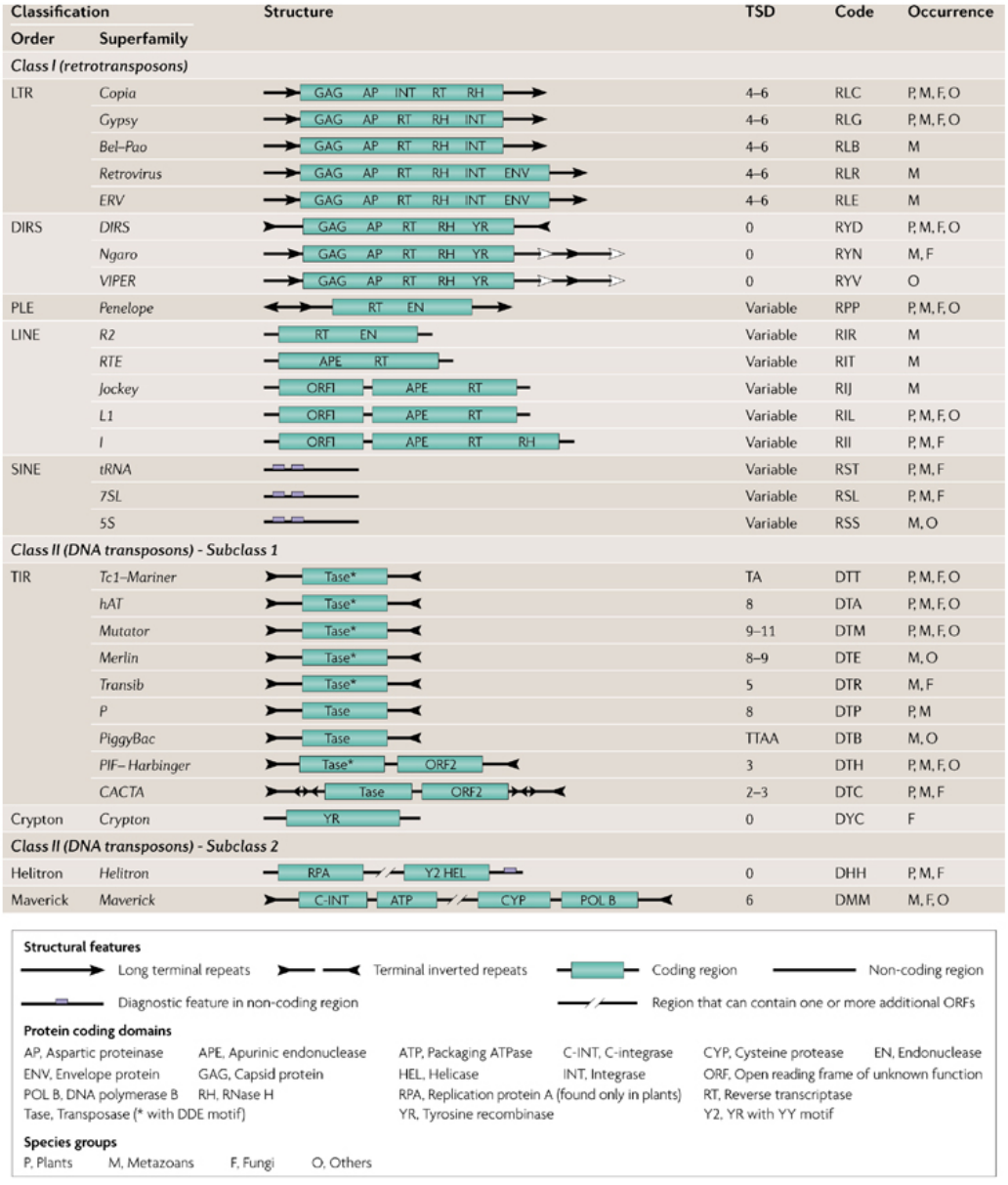


FIG. 3: Proposition de classification pour les ET. (Tiré de Wicker et coll. 2007.)

Comme l'indiquent ses auteurs, cette classification a été proposée dans un but pragmatique d'annotation et ne vise pas à refléter la phylogénie. Mais cela ne doit pas occulter le fait que les ET, en plus d'influer sur l'évolution de la structure des génomes, ont eux-même une structure très variable, beaucoup plus complexe que ce qu'un schéma tel celui de la figure 2 peut laisser paraître, ce qui complique la construction de leur phylogénie.

Un premier exemple de variations structurales est celui des éléments non-autonomes décrits précédemment. Le processus par lequel un élément devient non-autonome est a priori continu, l'élément accumulant progressivement substitutions, insertions et délétions. Parmi les éléments non-autonomes dérivant de rétro-transposons à LTR, certains sont qualifiés de LARD pour *large retrotransposons derivatives* lorsque leur taille excède quatre mille bases (Kalendar et coll. 2004), alors que les autres sont qualifiés de TRIM pour *terminal repeat retrotransposons in miniature* (Witte et coll. 2001). Mais comment distinguer à première vue un tel élément qui a continué à transposer d'un élément inactif subissant des délétions sans ne plus pouvoir transposer ? Si l'on s'intéresse à l'évolution de la structure du génome, le premier cas présente plus d'intérêt que le deuxième, les éléments non-autonomes ayant transposé représentant généralement une bien plus grande proportion du génome.

Un deuxième exemple de variations structurales correspond aux chimères résultant de la combinaison de plusieurs éléments répétés existants. Un tel cas a été montré chez le moustique *Anopheles gambiae* impliquant une famille de MITE dont l'une des répétitions terminales semble dériver d'une famille d'éléments P alors que l'autre répétition terminale semble dériver d'une famille d'éléments P différente (Quesneville et coll. 2006). Un cas encore plus extrême est celui de rétro-transposons à LTR non-autonomes appelés *Cassandra* dont chaque répétition terminale contient un gène d'ARN 5S dont la structure secondaire, très similaire des autres gènes d'ARN 5S, semble fonctionnelle (Kalendar et coll. 2008).

Un troisième exemple de variations structurales est celui d'éléments transposables dupliquant des fragments de gènes codants qui ne sont pas censés leur appartenir. Certains éléments peuvent transposer en copiant des séquences présentes au-delà de leurs extrémités, en 5′ ou en 3′. Si l'élément est dans un intron, il peut donc emporter un exon avec lui. Si le tout s'insère dans l'intron d'un autre gène, l'exon peut potentiellement être incorporé au transcrit du gène en question. Ce phénomène de trans-

duction de fragments de gènes par des ET a d'abord été observé avec les rétrotransposons L1 du génome humain (Moran et coll. 1999). Le phénomène de transduction peut aussi concerner un gène entier. Dans le génome humain, le gène AMAC a été dupliqué trois fois par ce mécanisme à l'aide d'un ET nommé SVA qui se trouve être lui-même un élément chimère (Xing et coll. 2006). Ces phénomènes d'*exon shuffling*, *exon trapping*, *gene transduction*, etc., sont aussi fréquents dans les génomes de plantes. Dans le génome du riz, un grand nombre d'ET proches de la famille *Mutator* contiennent des fragments de gènes (Jiang et coll. 2004). Les fragments étant inclus entre les répétitions terminales des ET, on parle alors de transduplication. Ces éléments appelés Pack-MULEs sont nombreux, environ 3000, et ne semblent pas générer seulement des pseudo-gènes non fonctionnels : en effet, 22% sont transcrits et parmi ceux-ci, 28 sont également traduits, un nombre vraisemblablement très sous-estimé (Hanada et coll. 2009). Par ailleurs, des résultats suggèrent que ces Pack-MULEs ont un effet sur l'expression des gènes dont ils contiennent des fragments, vraisemblablement *via* l'action de petits ARN (Hanada et coll. 2009). Le même type de variants structuraux existe en grande quantité dans le génome du maïs, mais cette fois avec des Hélitrons (Morgante et coll. 2005). Depuis que la séquence complète d'un génome de maïs est disponible (Schnable et coll. 2009), plusieurs études ont montré que les Hélitrons sont présents en un grand nombre de copies dans les génomes de maïs et que la plupart, 60%, contiennent des fragments d'un ou plusieurs gènes (Yang et Bennetzen 2009). Les mécanismes moléculaires par lesquels les ET acquièrent des fragments de gènes ne sont pas toujours bien élucidés.

La classification actuelle permet de catégoriser les ET facilement définissables. Or, les exemples de variations structurales au sein des familles d'ET sont nombreux et variés. Il est donc important de pouvoir annoter les ET au mieux, c'est-à-dire en les considérant dans toute leur diversité structurale.

1.3.3 *Démarche et objectifs de cette thèse*

Les ET sont maintenant bien connus pour présenter de grandes variations structurales au sein d'une même famille. Si l'on veut étudier leurs impacts sur l'évolution d'un génome, il faut en particulier pouvoir identifier les différentes familles d'ET présents dans ce génome, déterminer leurs variants structuraux,

en construire des séquences consensus correspondant aux entités ancestrales, puis annoter leurs fragments et enfin reconstruire les copies en connectant les fragments contigus au moment de l'insertion. L'approche *de novo* est ici encore plus indispensable étant donné que les variants structuraux d'une même famille sont la plupart du temps absents des banques de données. Au début de mes travaux de thèse, les approches *de novo* n'avaient généralement pas été testées à grande échelle. De plus, les variants structuraux n'étant pas recherchés pour eux-mêmes, il était difficile de distinguer une séquence consensus correspondant à un variant structural d'une séquence consensus correspondant à un élément tronqué et inactif. Enfin, nous ne disposions pas d'outils permettant de classifier automatiquement les séquences consensus d'ET et donc de détecter les faux-positifs.

Pendant mon travail de thèse, j'ai donc mis au point des outils informatiques d'annotation des ET prenant en compte la diversification des familles d'ET, processus par lequel des variants structuraux apparaissent et transposent (chapitre 2). Le choix des génomes de *D. melanogaster* et *A. thaliana* pour ce travail était motivé par la disponibilité de séquences de référence de qualité qui permettaient de mettre au point et de tester des outils informatiques implémentant une approche *de novo*. Dans le cadre de ces analyses, j'ai pu également montrer que ces outils permettaient de prendre en compte et de révéler l'importante variation structurale des familles d'ET de ces génomes. Dans le dernier chapitre (3), je présente les résultats de l'application de ces outils à différents génomes, notamment de plantes, champignons et insectes et je détaille l'attention apportée aux techniques de *développement logiciel* dans l'implémentation des outils. Ces travaux nous ont conduits à proposer une feuille de route pour l'annotation des ET. Dans la suite, je propose également plusieurs pistes d'amélioration afin d'incorporer d'autres approches et de rendre possible l'analyse de grands génomes ayant un fort contenu en répétitions, tel que celui du blé tendre, à l'aide des outils décrits précédemment. En termes de perspectives, je suggère quelques applications possibles de mon travail en soutien à la génétique et l'amélioration des plantes. Finalement, j'explicite comment simulation et mise au point d'algorithmes de détection se complètent pour l'étude de la dynamique des ET, inscrivant ainsi ma contributions dans un projet de recherche à long terme.

Deuxième partie

CONSIDÉRER LA DIVERSIFICATION DES ÉLÉMENTS TRANSPOSABLES DANS LES APPROCHES D'ANNOTATION DE NOVO

2

CONSIDÉRER LA DIVERSIFICATION DES ÉLÉMENTS TRANSPOSABLES DANS LES APPROCHES D'ANNOTATION DE NOVO

Ce travail a été soumis récemment. J'en présente ci-dessous un résumé en français, suivi du manuscrit en anglais.

Les éléments transposables (ET) sont des séquences d'ADN mobiles répétées, présentes dans quasiment tous les génomes procaryotes et eucaryotes étudiés jusqu'à maintenant. Ils ont de larges impacts sur la structure des génomes, leurs fonctions et leur évolution. Grâce au récent développement des méthodes de séquençage à haut-débit, de nombreux génomes ont pu être séquencés, rendant ainsi possible l'analyse comparative des dynamiques d'éléments transposables à une échelle sans précédent. Plusieurs méthodes ont été proposées concernant l'identification de novo des ET dans les génomes séquencés. La plupart commence par la détection des répétitions génomiques, mais diffère lors des étapes suivantes de définition des familles d'ET. Des annotations d'ET de qualité sont disponibles pour les séquences génomiques de *Drosophila melanogaster* et *Arabidopsis thaliana*, fournissant une base solide pour tester et valider de telles méthodes. Nous avons donc comparé les performances d'algorithmes spécifiques du regroupement de répétitions dispersées, et avons trouvé que seule l'utilisation combinée de certains algorithmes détectait les familles d'ET avec une bonne reconstruction des séquences de référence. Nous avons ensuite appliqué une nouvelle procédure pour réconcilier les différents résultats de regroupement et classifier les séquences d'ET. Cette approche a été implémentée dans un outil utilisant la boîte à outils REPET. Finalement, nous montrons que notre approche combinée met en évidence la dynamique de familles d'ET bien définies en rendant possible l'identification de variations structurales entre leurs copies. Cette approche permet ainsi d'annoter les familles d'ET et d'étudier leur diversification en une seule analyse, améliorant ainsi notre compréhension de la dynamique des ET à l'échelle du génome entier et ce pour différentes espèces.

Considering transposable element diversification in *de novo* annotation approaches

Timothée Flutre[1], Elodie Duprat[2], Catherine Feuillet[3], Hadi Quesneville[1]

1 Unité de Recherche en Génomique-Info, UR 1164, INRA centre de Versailles-Grignon, RD 10, 78026 Versailles, France.

2 IMPMC, UMR 7590, CNRS-UPMC-IPGP-Université Paris Diderot, 140 rue de Lourmel, 75015 Paris, France.

3 Génétique, Diversité et Ecophysiologie des Céréales, UMR 1095, INRA domaine du Crouël, 234 avenue du Brézet, 63100 Clermont-Ferrand, France.

ABSTRACT

Transposable elements (TEs) are mobile, repetitive DNA sequences that are almost ubiquitous in prokaryotic and eukaryotic genomes. They have a large impact on genome structure, function and evolution. With the recent development of high-throughput sequencing methods, many genome sequences have become available, making possible comparative studies of TE dynamics at an unprecedented scale. Several methods have been proposed for the *de novo* identification of TEs in sequenced genomes. Most begin with the detection of genomic repeats, but the subsequent steps for defining TE families differ. High-quality TE annotations are available for the *Drosophila melanogaster* and *Arabidopsis thaliana* genome sequences, providing a solid basis for the benchmarking of such methods. We compared the performance of specific algorithms for the clustering of interspersed repeats and found that only a particular combination of algorithms detected TE families with good recovery of the reference sequences. We then applied a new procedure for reconciling the different clustering results and classifying TE sequences. The whole approach was implemented in a pipeline using the REPET package. Finally, we show that our combined approach highlights the dynamics of well defined TE families by making it possible to identify structural variations among their copies. This approach makes it possible to annotate TE families and to study their diversification in a single analysis, improving our understanding of TE dynamics at the whole-genome scale and for diverse species.

AUTHOR SUMMARY

Since more than a decade, whole genomes are being sequenced, now on a high-throughput basis. In the vast majority of the cases, it is possible to detect the presence of mobile genetic elements named transposable elements (TEs). Such repeated sequences can represent most of the genome (e.g. about 45% in humans) and have deep impacts on the evolution of genome structure and function, making their annotation a prerequisite for further analyses. Identifying TEs experimentally, for instance by screening mutants, is a labor-intensive task not applicable on a large scale, therefore requiring computational methods. However, relying on already-known TE sequences prevent from identifying new ones. Moreover, our aim is not only to annotate TEs but although to recover the ancestral sequences that transposed in order to gain insights on TE dynamics. Here, we describe a *de novo* approach and show how to combine efficiently existing algorithms to fully recover most of the ancestral TEs in any sequenced genome. Our results highlight the emergence of structural variants during the diversification of TE families, a process reflecting the tempo and mode of TE evolution.

INTRODUCTION

Transposable elements (TEs) are DNA sequences that can move and duplicate, autonomously or with the assistance of other elements, within genomes. TEs have been described as the “ultimate parasite”, because of their ability to amplify and invade genomes for their own ends [1], and as “selfish DNA sequences” [2]. These invasion events play a particularly important role in eukaryotic genomes, probably because of the smaller population sizes of eukaryotes than of prokaryotes [3].

TEs are generally classified according to their transposition mechanism. Those transposing *via* an RNA intermediate belong to class I and are referred to as retrotransposons, whereas those transposing *via* a DNA intermediate belong to class II and are called DNA transposons [4]. Class I introns can be classified into three main orders, LTR retrotransposons (having long terminal repeats), LINEs (long interspersed nuclear elements) and SINEs (short interspersed nuclear elements), whereas class II introns are classified into DNA transposons, Helitrons and Mavericks [5]. Most TEs encode proteins that mediate their autonomous transposition. During the course of evolution, non autonomous elements have emerged from autonomous elements. Some are incomplete versions of autonomous elements, often with insertions/deletions (indels) disrupting their open reading frames (ORFs). Others are miniature versions lacking internal sequences but retaining the boundaries of the original element, making it possible for the autonomous element machinery to recognize them. MITEs (miniature inverted-repeats transposable elements) are well known examples of non autonomous elements that evolved from class II DNA transposons [6,7].

TEs are now recognized to be a major component of the structure of the genome and to affect genome size and chromosomal rearrangements [8-10]. They often account for a large proportion of the genome: 20% of the 180 Mb *Drosophila melanogaster* genome, 45% of the 3.2 Gb human genome, and more than 80% of the 17 Gb bread wheat (*Triticum aestivum)* genome [11]. These dispersed repeats can induce major chromosomal rearrangements, thereby affecting genome

organization. However, the impact of TEs is not limited to effects on genome structure. As initially suggested by Barbara McClintock [12], TEs may be seen as “controlling” elements. They may provide regulatory sequences with various effects on the adjacent genes. In particular, some silencing mechanisms involving RNA interference seem to have emerged primarily as a host response to prevent TE amplification. Thus, genes located close to TE insertions may be subject to transcriptional control due to TE repression, resulting in their epigenetic regulation [13,14]. Moreover, TEs are intrinsically able to create, modify and re-wire gene regulatory networks [15,16]. Finally, many cases of exaptation and domestication involving TEs have been reported [17,18]. For example, there are several lines of evidence to suggest that the RAG1 and RAG2 genes involved in V(D)J recombination originated from a hAT DNA transposon that was domesticated to fulfill this primordial function of the adaptive immune system [19,20].

The increases in efficiency and decreases in cost of new sequencing techniques [21] are leading to the sequencing of increasing numbers of genomes. About 1250 genome sequencing projects have already been initiated for eukaryotes, including species with large and repetitive genomes, such as maize [22]. The efficient and accurate annotation of TEs is therefore essential to our understanding of their impact on gene function and genome evolution [23].

If a routine TE annotation procedure is to be efficient, it must be both rapid and exhaustive, biologically relevant and computationally tractable. The TE annotation process can be divided into two phases: (i) the *de novo* discovery and identification of the TE families present in the genome studied and (ii) the precise, comprehensive annotation of TE copies on the chromosomes. For the second phase, an integrated pipeline has already been developed and tested [24] and this pipeline has been applied to several organisms [25-31]. For the *de novo* discovery phase, several programs and algorithms based on different assumptions have been developed, but none has yet proved entirely satisfactory. Indeed, as pointed out in a previous study [32], some programs have

very low levels of sensitivity or specificity, whereas others return too short consensus sequence (<1 kb).

In addition to describing the composition and organization of the genome, TE annotation facilitates the identification of structural variants providing useful information about genome dynamics. Several examples of structural variations in TE families have been reported [33,34] but these variations have generally been underestimated in genome-wide analyses of TEs. In this study, we addressed two questions, one concerning the challenges associated with whole-genome TE annotation, and the other relating to the identification and characterization of structural variants from the same TE family. We first compared the existing computational methods for the *de novo* identification of TEs in sequenced genomes, using the high-quality TE annotations available for the *Drosophila melanogaster* [24] and *Arabidopsis thaliana* genome sequences [29]. We then developed the TEdenovo pipeline, a tool combining several different programs, including procedures for the clustering of interspersed repeats, into a single framework. Finally, by analyzing the *D. melanogaster* and *A. thaliana* genomes with the TEdenovo pipeline, we were able to obtain new insight into TE dynamics, highlighting structural variations emerging during the diversification of TE families and identifying putative new TEs absent from reference databanks.

RESULTS

Comparative analysis of *de novo* approaches

We developed a three-step approach for comparing the efficiency of *de novo* TE detection methods (see [35] for a review), to provide a robust tool for identifying TEs in eukaryotic genomes: (i) the self-alignment of the input genomic sequences, (ii) the clustering of the resulting pairwise alignments, and (iii) the construction of a multiple alignment for each cluster from which a consensus sequence is derived (see figure S1). This process generates a databank of *de novo* consensus sequences representing putative TE families present in the genome analyzed, which can be used for the annotation of individual TE copies. We applied this three-step approach to the *D. melanogaster* release 4 and *A. thaliana* release 9 genome sequences. At each step, we evaluated several programs, comparing the efficiency with which they identified TEs with the aid of the high-quality TE sequence databanks (from the Berkeley *Drosophila* Genome Project and Repbase Update) and annotations available for these two reference genomes [24,29].

Traditionally, the quality of *de novo* consensus sequences — the extent to which they correspond to full TE reference sequences rather than truncated versions — is not assessed. Validation is instead indirect: researchers annotate a genome sequence with RepeatMasker, using Repbase Update as TE databank, and consider the resulting TE annotations as the references. They then annotate the same genome with RepeatMasker, using the *de novo* consensus as TE databank, and consider these TE annotations as predictions. These two sets of annotations are then compared, by calculating sensitivity and specificity at the nucleotide level. The criterion used to estimate the quality of the *de novo* method is therefore the extent to which *de novo* predictions and reference annotations overlap. However, as we are particularly interested in TE dynamics, we need to assess the quality of the *de novo* library itself, by evaluating the extent to which full ancestral TE reference sequences are recovered. Such sequences, which originate from the reconstruction of a

given element from its copies, are not only useful for subsequent TE annotation, but also provide a condensed view of the TEs in the genome. One way of assessing the quality of the *de novo* consensus sequences obtained with our three-step approach would be to compare these sequences with reference sequences from the Berkeley Drosophila Genome Project (BDGP) or Repbase Update databanks. However, it was clear that some of the reference sequences present in these databanks would not be present in the genomes analyzed. For example, the “P-element” reference sequence is absent from the genome sequence of the *D. melanogaster* strain used here. So, rather than using the reference databanks directly, we first constructed, for each genome, a “knowledge-based” databank comprising one consensus sequence per reference TE sequence, based on its genomic copies (see Methods). For each genome, we then compared each *de novo* databank with its corresponding “knowledge-based” databank through pairwise sequence alignments. We then calculated the sensitivity S_n*, specificity S_p* and recovery ratio R_{CC} (see Methods and table 1). This last index, the R_{CC} ratio, provides a precise measurement of the number of TE reference elements fully recovered in the *de novo* consensus sequences.

Self-alignment of the genomic sequences

The first step in the three-step *de novo* approach involves the self-alignment of the input genomic sequences, corresponding to an all-by-all comparison of the genome with itself. We evaluated two local pairwise alignment programs for this first step: a heuristic algorithm, BLASTER [36], and an exact algorithm, PALS [37]. BLASTER is a wrapper for the BLAST softwares [38]. It was used for comparisons at the genome scale. It begins by cutting long queries into batches and launching them in parallel against the subject databank. The second program, PALS, implements a filter algorithm. It first finds all exact matches of length *q* between the query and subject sequences. It then restricts the search by identifying regions (parallelograms in the alignment matrix) containing a number of hits above a given threshold. Finally, if these regions have a length above a given threshold, PALS develops a chain of hits for each region, aligns the nucleotides and returns the coordinates of the resulting matches. We launched these two programs, with stringent parameters,

on each of the target genomes, and then applied post-processing procedures to discard long segmental duplications (see Methods). This resulted in a list of pairwise alignments corresponding to repeats in the *D. melanogaster* and *A. thaliana* genome sequences. A comparison of the *de novo* consensus sequence performances of the BLASTER and PALS programs (table 1) showed that BLASTER consistently had a higher sensitivity (S_n*) and a much higher recovery ratio (R_{CC}). In most cases, its specificity (S_p*) was also better than that of PALS. Note that PALS was run with more stringent parameters than BLASTER as it cannot be used with the same values without computing time becoming intractable (see Methods). As the recovery ratio R_{CC} reflects the ability of the *de novo* approach to define TE boundaries correctly and, therefore, to recover full-length TE reference sequences, we preferred BLASTER over PALS for the self-alignment step. Moreover, in both genomes, BLASTER gave a higher genome coverage than PALS (table S1), as expected given the lower stringency of its parameters (see Methods). This first step also generated a lower limit for the repeats content of the genome: around 7% for *D. melanogaster* and 13% for *A. thaliana*.

Clustering of the all-by-all matches

In the second step, we clustered the results of the all-by-all comparisons, with the aim of gathering together, within the same cluster, all sequences belonging to the same TE family. This step is crucial to ensure the precise definition of repeat boundaries, one of the main challenges in the *de novo* detection of TEs. Due to their specific dynamics during genome evolution, TE families may differ considerably in terms of copy number, sequence divergence and insertion/deletion patterns. At this step, the aim is to cluster together all TE fragments sharing a common ancestor, with the aim of recovering the ancestral element that transposed in the past. However, TE copies diverge as they multiply and different copies may accumulate different modifications. This phenomenon results in structural variants, as depicted in figure 1. In this case, the grouping together of copies from different structural variants within the same cluster may have the undesirable consequence of the identification of consensus sequences with some features specific to one of the structural

variants with others specific to another variant. We therefore tested three programs specifically implemented for the clustering of interspersed repeats: GROUPER [36], RECON [39] and PILER [40].

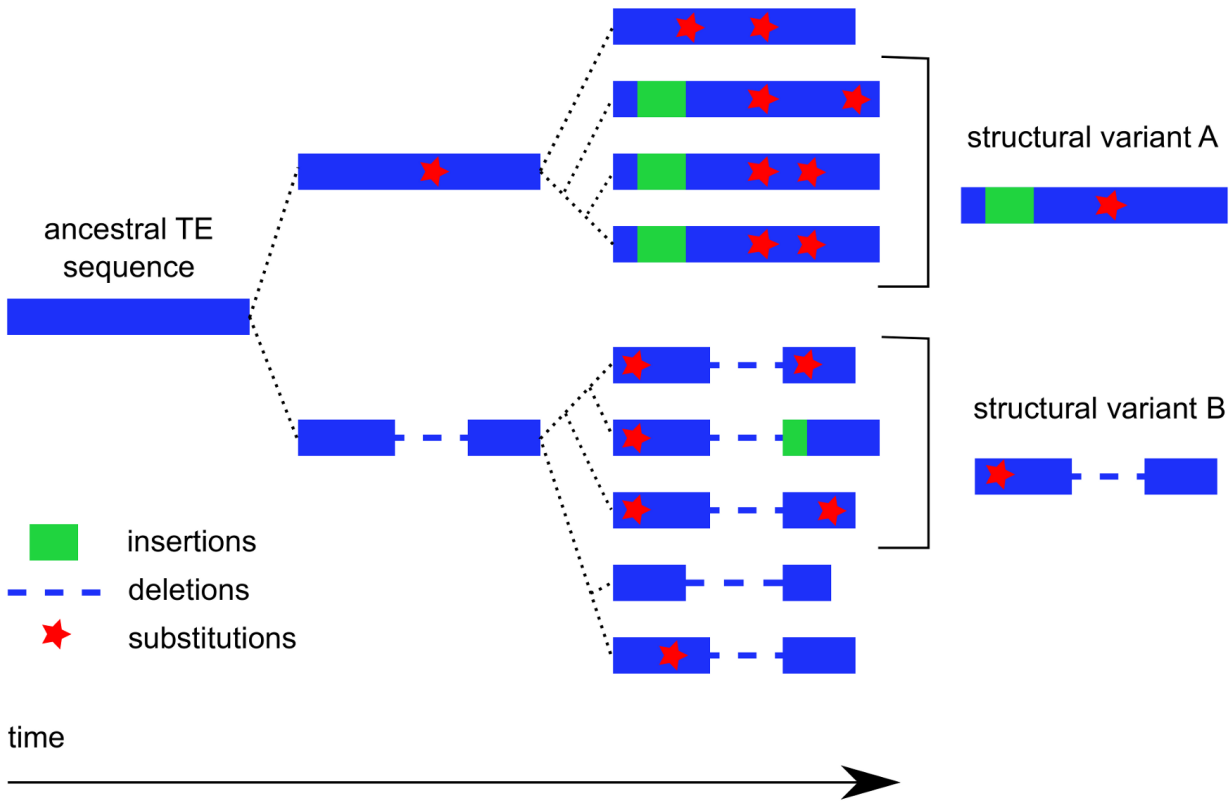


Figure1. Schematic diagram of the dynamics of a TE family with two structural variants.

As TEs are frequently inserted within each other, GROUPER first aims at retrieving the full-length copies by connecting their fragments *via* a dynamic programming algorithm applied to the pairwise alignments. It links together, in chains, the matches corresponding to fragments belonging to the same TE copy but interrupted by indels. It then uses single-link clustering to gather similar chains of matches into the same cluster, using a high-coverage constraint (95%), and merges chains corresponding to the same locus. The use of a high-coverage constraint makes it possible to

identify different structural variants from a given TE family (figure 1): two copies from two different variants overlap by less than 95% threshold, and are therefore assigned to different clusters. Compared to the previous version of the program [36], we added a procedure to remove redundancy among clusters (see Methods). RECON uses a different strategy, first trying to infer the ancestral TE copies, named “elements”, from all the fragments at each locus. It does this by single-link clustering with a low-coverage constraint (50%), followed by an additional procedure focusing on the aggregation of endpoints to ensure the correct handling of composite elements, such as segmental duplications. It then gathers the “elements” into families again by single-link clustering, but this time with a high-coverage constraint (90%), and a procedure is then applied to deal with families that are related but different, based on length ratio and similarity thresholds. We also tested the PILER suite of programs. In this work, “PILER” refers systematically to PILER-DF, which focuses on interspersed repeats. PILER first defines “piles” as lists of matches covering a maximal contiguous region. It then globally aligns these piles, rather than the individual matches themselves. This prevents bias towards the shortest match in the pile. Finally, piles that cannot be globally aligned with each other over 95% of their length are gathered into clusters.

We applied these three programs to the list of matches obtained in the genome self-alignments, and then applied procedures for discarding segmental duplications (see Methods). In terms of the *de novo* consensuses generated from the resulting clusters (table 1), GROUPER and RECON were consistently more sensitive than PILER. Moreover, GROUPER was systematically more specific than RECON and PILER, with *de novo* consensuses from GROUPER less likely to be artifacts unrelated to TE reference sequences than those obtained with the other methods. In some cases, RECON provided a better recovery ratio (R_{CC}) than GROUPER and PILER, but this ratio was always better for GROUPER than for PILER. In addition, regardless of the genome analyzed, the three programs did not necessarily generate the same number of clusters (table S2): GROUPER generated more clusters than RECON, which, in turn, generated more clusters than PILER. The number of clusters obtained by the *de novo* approach was much higher for GROUPER

and RECON than the number of TE families present in the reference databanks (BDGP or Repbase Update). We suggest that this large number of clusters reflects the high level of diversity within TE families in terms of copy divergence and nesting patterns, as assessed by clustering algorithms.

Multiple sequence alignment for each cluster

The third step, which consists in the construction of a multiple alignment for each cluster, is particularly important, because the quality of a consensus sequence depends on the quality of the multiple sequence alignment (MSA) from which it is derived. A large number of programs have been developed over the last 10 years [41], focusing on issues such as the scoring scheme, the use of templates for secondary structures and computation speed. In our case, the sequences to be aligned are very similar to each other in terms of nucleotide substitutions (identity of more than 90%), with most of the differences between them resulting from indels. The clusters generated by RECON include sequences of very different lengths (table S2). Moreover, depending on the size and TE content of the genome analyzed, it may be necessary to process several thousand clusters. We tested only a subset of all the possible programs, focusing on those most suitable for our requirements: rapid procedures capable of handling indels. We therefore focused on progressive MSA algorithms, as these algorithms generate alignments rapidly.

MAP [42] was specifically designed to handle long gaps. Such gaps frequently occur when aligning TE copies of different lengths from the same TE family. In this program, gaps are not penalized beyond a given length. CLUSTAL-W [43] is a well known progressive MSA algorithm, and was the first to propose position-specific gap penalties. MAFFT [44] addresses the issue of CPU time by using Fast Fourier Transform for the rapid detection of homologous segments. It also implements a normalized similarity matrix that is said to perform better for alignments with sequences of different lengths. This program was the fastest MSA program we tested. PRANK [45] takes into account the phylogenetic information contained in indels to distinguish insertions from deletions, to position them properly and avoid the overestimation of deletions. This sophistication

renders PRANK much slower than MAP (10 times slower for a typical cluster containing 16 sequences of 8 kb each).

For each of these multiple alignment programs, we used the matches returned by BLASTER and then clustered by GROUPER, RECON or PILER. This comparative study (table S3) shows that all the multiple alignment programs gave similar results in terms of sensitivity and specificity, whether launched after GROUPER or after PILER. This result was expected, as all the sequences in these clusters are similar in both composition and length, making them easy to align. However, PRANK was slower than the other three programs, rendering it less suitable for large genome analyses. With RECON, the best results were obtained with MAP, due to the greater heterogeneity of RECON clusters in terms of sequence length, rendering alignment more difficult. Thus, few differences were observed, but MAP clearly outperformed the other programs on RECON clusters, making it more robust than the other programs, whatever the clustering method used.

Comparison with another approach, RepeatScout

We evaluated the performance of our three-step strategy, by comparing it with another approach that also builds a databank of TE consensus sequences from a raw genome sequence. RepeatScout [46] begins by keeping high-frequency strings of length *k*, called *k*-mers. The program initially takes the most frequent *k*-mer and sets the consensus as being the *k*-mer in question. Based on the multiple alignment of all its occurrences, the program extends this consensus one nucleotide at a time, in both directions, according to a specific scoring function. For correct definition of the consensus boundaries, the scoring function is designed to allow extension of the consensus, even if shared by some alignments, but not all. Once the consensus can be extended no further, the program detects all its occurrences in the genome, and updates the initial table of *k*-mer frequencies accordingly. This procedure is applied iteratively for each *k*-mer with a frequency above a given threshold.

We applied the RepeatScout program on the *D. melanogaster* and *A. thaliana* genome sequences, with the default parameters. It constructed 1770 consensuses for *D. melanogaster* and 3417 for

A. thaliana . However, these consensuses were less sensitive and specific than those obtained with the tools described above (table S4). This is probably due to the shorter length of the consensuses identified by RepeatScout (500 bp on average) than by GROUPER (~ 2500 bp on average), RECON (~ 2000 bp on average) and PILER (~ 2800 bp on average). We obtained similar results for both genomes, indicating that this bias does not seem to be due to the input genome sequences. We hypothesize that RepeatScout fails to connect TE fragments more than a certain distance apart, thereby sometimes missing the true boundaries of a given TE copy. By contrast, GROUPER and RECON are particularly efficient at this task.

Combination of programs into a robust pipeline, TEdenovo

Based on the comparative analyses reported above, BLASTER should be used for the genome self-alignment step, followed by GROUPER or RECON for the clustering step and MAP for the multiple alignment step. However, a comparison of the consensuses obtained with all three clustering methods clearly showed that each of these methods nonetheless missed several reference TEs fully recovered by the others (figure 2). With the *D. melanogaster* genome sequence, four “knowledge-based” consensuses were retrieved intact by GROUPER only, and nine such sequences were recovered by RECON only (figure 2 A). Similarly, with the *A. thaliana* genome sequence, eight “knowledge-based” consensuses were retrieved intact only by GROUPER, whereas fifteen were retrieved intact only by RECON (figure 2 B). The three clustering methods should therefore be used in combination, for the accurate identification of TE families in genome sequences. For this reason, we decided to the three-step approach within a combined, modular pipeline named “TEdenovo”. With respect to the best single method, GROUPER for *D. melanogaster* and RECON for *A. thaliana* (table 1), the combined approach, as implemented in the TEdenovo pipeline, improved the recovery of full-length “knowledge-based” consensus sequences by 20% and 13.5%, respectively, while maintaining high sensitivity and specificity (table S5). The three approaches are combined at the clustering step, through the launching of GROUPER, RECON and PILER in parallel. The user may also choose to use PALS rather than

BLASTER at step 1 or the other MSA programs at step 3, and can even choose to apply only one clustering program at step 2, although our results suggest that this would not be wise.

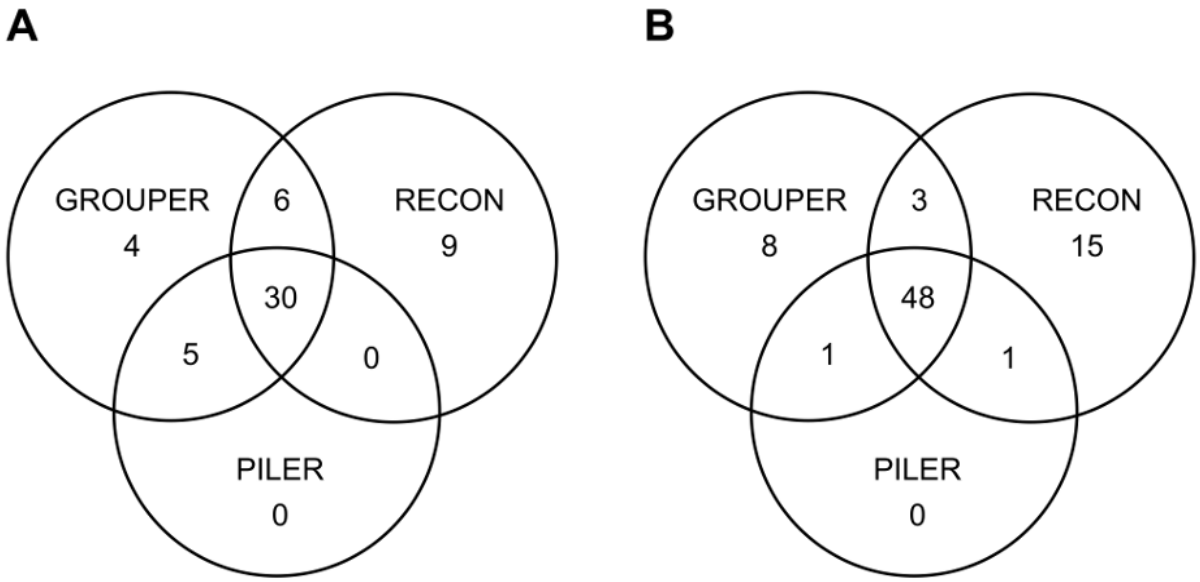


Figure 2. Venn diagram showing the gains achieved by combining several clustering programs. (A) Combining the GROUPER and RECON programs in particular makes it possible to fully recover more TE sequences than each program alone from the *D. melanogaster* genome. (B) Same conclusion from the *A. thaliana* genome.

Table 1. Sensitivity and specificity of the programs tested in the three-step *de novo* approach.

Genome	Self-alignment	Clustering	Multiple alignment	S_n*	S_p*	R_{CC}
D. mel.	BLASTER	GROUPER	MAP	80.34%	85.89%	66.20%
D. mel.	BLASTER	RECON	MAP	92.31%	73.17%	66.20%
D. mel.	BLASTER	PILER	MAP	62.39%	84.17%	51.50%
D. mel.	PALS	GROUPER	MAP	73.50%	88.75%	60.30%
D. mel.	PALS	RECON	MAP	90.60%	74.23%	51.50%
D. mel.	PALS	PILER	MAP	53.85%	76.42%	42.64%
A. tha.	BLASTER	GROUPER	MAP	60.33%	82.42%	39.00%
A. tha.	BLASTER	RECON	MAP	73.77%	61.70%	43.50%
A. tha.	BLASTER	PILER	MAP	47.21%	57.33%	32.45%
A. tha.	PALS	GROUPER	MAP	54.75%	88.38%	24.00%
A. tha.	PALS	RECON	MAP	71.80%	66.20%	27.90%
A. tha.	PALS	PILER	MAP	40.00%	59.92%	16.20%

"*D. mel.*" stands for "*D. melanogaster*" and "*A. tha.*" stands for "*A. thaliana*". The three indices S_n*, S_p* and R_{CC} correspond respectively to the measure of sensitivity, the measure of specificity and the recovery ratio when comparing a databank of TE *de novo* consensus sequences with a databank of TE reference sequences.

Classification of TE consensus sequences and identification of structural variation within TE families

Classification of the consensus sequences and elimination of redundancy

The three-step approach provided us with a set of *de novo* consensus sequences corresponding to interspersed sequences occurring at least three times in the genome studied. A two-step classification procedure was implemented to add more biological information, to filter out false-positives and to eliminate the redundancy introduced by the combined approach. There is a long-standing debate about the aims of any classification in biology [47], and the case of genomic repeats does not escape the rule: “Although the reality is that repeats [...] are a hierarchical evolutionary continuum that defies classification, it is still desirable to impose a simplistic classification that pretends that repeat families are distinct, for the purpose of practical genome annotation” [39]. In this spirit, our classification procedure begins with the detection of TE features in the consensus sequences, and a decision tree classifying each consensus as a function of these features is then produced (figure 3). In the first step, the procedure identifies terminal repeats, tandem repeats, poly-A tails and SSR-like tails (simple sequence repeats). It also aligns the consensus sequences with known TEs through blastn, blastx and tblastx, and with known genes from the host genome. The known TEs are those from Repbase Update, a curated set of TE sequences from numerous genomes [48]. Our TE classifier uses a customized version of this databank available from the Repbase Update website (http://www.girinst.org/server/RepBase/index.php), but any customized databank may be used, provided that it is appropriately formatted. In the second step, the procedure implements a decision tree based on the classification summarized in [5]. For benchmarking purposes, when analyzing *de novo* consensus sequences from the *D. melanogaster* and *A. thaliana* genomes, we removed the sequences known to belong to these species from the Repbase Update databank, to establish conditions equivalent to those for the analysis of a new genome.

The classification takes into account the degree of completeness of the *de novo* TE consensus (figure 3). For instance, if a consensus sequence has the required “structural features” — LTRs (long terminal repeats), TIRs (terminal inverted repeats) or a tail (poly-A or SSR-like) — and “coding features” — matches with known TEs in tblastx and blastx analyses — then it is considered “complete”. If it has only one of these two types of features, it is classified as “incomplete”. In the case of incompatible features, the consensus is classified as “confused”, and if the consensus has no identifiable features, it is classified as “not categorized”. We also used length parameters (table S6) benchmarked on the reference databanks — BDGP for *D. melanogaster* and Repbase Update for *A. thaliana* — to improve differentiation between presumably truncated and full-length consensus sequences. We also used length parameters when classifying a sequence having only “structural features”, to determine whether the sequence concerned was a SINE or a MITE.

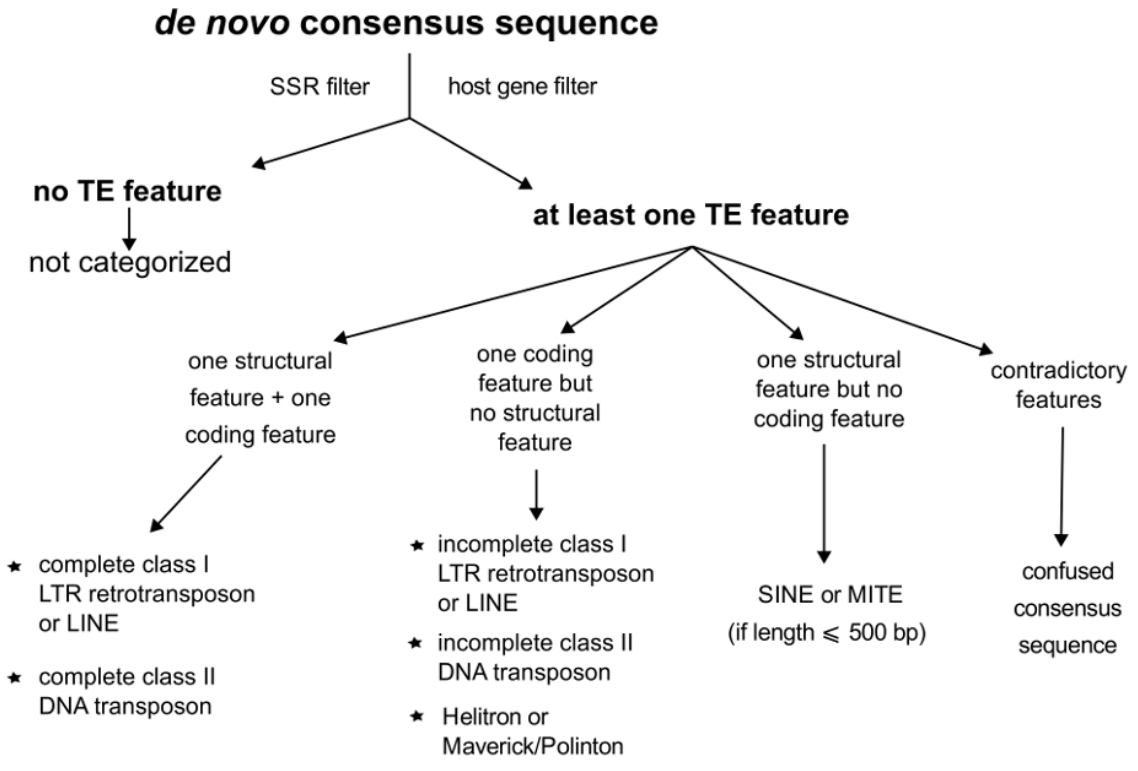


Figure 3. Simplified decision tree implemented in the TE classifier.

As shown in table 2 for *D. melanogaster* (table S7 for *A. thaliana*), our TE classifier retrieved the appropriate classification for the reference TEs, thus predicting a correct classification for *de novo* consensus sequences. Moreover, each high-level category of the TE classification (*e.g.* “class I LTR-retrotransposon”, “class I LINE”, and “class II DNA transposon”) contained similar proportions of *de novo* consensus sequences and TE reference sequences from the BDGP or Repbase Update databanks. Thus, our procedure efficiently classified *de novo* consensus sequences into the appropriate high-level categories.

The proportion of *de novo* consensus sequences classified as “incomplete” was high. In *D. melanogaster*, 75% (45% in *A. thaliana*) of the *de novo* consensus sequences classified as “incomplete” LTR retrotransposons matched only with known LTR retrotransposons from Repbase Update but contained no long terminal repeats. All the *de novo* consensus sequences classified as “incomplete” LINE retrotransposons from *D. melanogaster* (28% of those from *A. thaliana*) matched only known LINE retrotransposons from Repbase Update, but with no polyA-/SSR-like tail. Among the *de novo* consensus sequences classified as “incomplete” DNA transposons, 50% of those from *D. melanogaster* (32% of those from *A. thaliana*) matched only known DNA transposons from Repbase Update, but with no TIR. Thus, these *de novo* consensus sequences are classified as “incomplete” in *D. melanogaster* mostly because they lack terminal features, which is not the case for most such consensus sequences in *A. thaliana*. Thus, the clustering programs combined in the TEdenovo pipeline do not generate systematic bias towards “incomplete” TEs of a given kind.

Table 2. TEclassifier results for the classification of *D. melanogaster* TE sequences.

Classification	Reference TEs from the BDGP	*De novo* consensus (with redundancy)	*De novo* consensus (without redundancy)
Class I “complete” LTR retrotransposon	56	150	48
Class I “incomplete” LTR retrotransposon	2	377	209
Class I “complete” LINE	23	117	27
Class I “incomplete” LINE	17	147	57
Class I SINE	0	2	1
Class II “complete” DNA transposon	19	30	13
Class II “incomplete” DNA transposon	2	75	32
Class II MITE	0	8	5
Helitron	1	0	0
SSR	0	8	8
Host genes	0	26	11
Confused	1	20	6
No category	5	341	176
Total	126	1301	593

Moreover, only a few consensus sequences were classified as “confused”, facilitating manual curation. The consensus sequences classified as “host genes” are detailed in the supplementary materials (table S8). The high proportion of unclassified consensus sequences (“no category”) reflects the limitations of this classification scheme. We searched all these consensus sequences

for HMM profiles specific to TEs. We detected fragments of known TE profiles in only 14% of these unclassified consensus sequences (data not shown). Several of the other corresponded to Helitron reference sequences recovered intact within unclassified *de novo* consensus sequences. Helitrons are difficult to detect as they have no clear structural features other than a terminal hairpin [49]. Nonetheless, our *de novo* approach recovered all those with at least three full-length copies in the genome. This confirms the relevance of the combined *de novo* approach as implemented in the TEdenovo pipeline and shows that *de novo* consensus sequences may correspond to TEs even if they remain unclassified by this method.

As shown above, the combination of clustering programs gave better results than any single program. However, it also provided redundant consensus sequences. We therefore applied a redundancy elimination procedure, in which we considered a consensus sequence to be redundant if it was included in another sequence, over *x*% of its length, with an identity of more than *y*%. We first tested this procedure directly on the whole *de novo* databank. It resulted in the loss of many well classified consensus sequences that were shorter than the misclassified sequences. We therefore applied the redundancy procedure on the basis of the classification. We decided to remove redundant consensus sequences classified as “incomplete” when they were included within consensus sequences classified as “complete”, but not vice versa. The “80-80-80” rule [5] has been proposed as a means of identifying copies from the same TE family: two TE copies may be considered to belong to the same family if they are aligned, with 80% identity, over at least 80 bp and 80% of their respective lengths. However, as this rule was originally developed for TE copies and not for consensus sequences, we also tested more stringent parameters (table S9). We found that the best strategy for obtaining a high-quality *de novo* databank with a low level of redundancy was to remove redundant consensus sequences with more stringent parameters: “95-80-98”. This implies that a consensus sequence is removed if it is included within another consensus sequence over 98% of its length, with an identity level exceeding 95%.

Comparison of de novo *and knowledge-based TE annotations*

After the first phase of analysis with the TEdenovo pipeline, we used the *de novo* TE consensus sequence databank to detect all TE fragments and to reconstruct each TE copy in the genome of interest. This annotation phase was achieved with the TEannot pipeline [24], which is also part of the REPET package. TEannot combines several programs for detecting TE fragments, filtering out false-positives and reconstructing intact TE copies. This process involves the connection of TE fragments from the same copy, a procedure also called “TE defragmentation”. While improving the robustness of this pipeline, we notably improved the connection of TE fragments in the MATCHER algorithm (see Methods). For a comprehensive analysis of the performance of the *de novo* approach presented above, we used the TEannot pipeline on both the *D. melanogaster* and *A. thaliana* genomes, with several databanks of *de novo* consensus sequences, each obtained with a specific combination of programs from the TEdenovo pipeline. We first discarded the consensus sequences that could be unambiguously identified as SSRs or host genes. We then compared each annotation with that obtained with the reference TE libraries from the BDGP and Repbase Update databanks (see table 3).

In *D. melanogaster* (table 3), when using only one clustering method, the GROUPER databank delivered the annotation closest to that obtained with the BDGP reference databank in terms of genome coverage and copy number. Sensitivity was highest with the RECON databank and specificity was highest with the PILER databank. The annotation shows high sensitivity and specificity, together with a high level of genome coverage, for the combined approach. Similar conclusions were drawn from the annotations for *A. thaliana.* An examination of match boundaries (figure S3 and table S10) showed that the clustering methods were complementary: GROUPER gave the largest number of exact matches in *D. melanogaster* whereas RECON gave the largest number of exact matches in *A. thaliana*. Thus, the combination of clustering methods in the TEdenovo pipeline leads to the construction of a high-quality TE library delivering annotations similar to those obtained for manually curated databanks.

Table 3. TE annotation results obtained with reference databanks and *de novo* databanks.

Genome	TE databank	Consensus (having copies)	TE genome coverage	Number of copies	S_n	S_p
D. mel.	BDGP	125	10.51%	31208	NA	NA
D. mel.	GROUPER	712	10.29%	43699	81.92%	98.12%
D. mel.	RECON	437	11.05%	33072	87.77%	97.95%
D. mel.	PILER	114	8.87%	32789	74.07%	98.79%
D. mel.	G+R+P	568	11.98%	42847	91.43%	97.35%
A. tha.	Repbase	318	19.02%	41146	NA	NA
A. tha.	GROUPER	1237	18.78%	41791	79.29%	95.43%
A. tha.	RECON	1004	23.69%	49470	88.75%	91.59%
A. tha.	PILER	300	13.14%	34818	56.56%	97.05%
A. tha.	G+R+P	1232	22.77%	44059	87.03%	92.32%

"*D. mel.*" stands for "*D. melanogaster*" and "*A. tha.*" stands for "*A. thaliana*". The S_n and S_p columns correspond respectively to sensitivity and specificity results when comparing two annotations in terms of nucleotide overlaps. "G+R+P" indicates that the three programs GROUPER, RECON and PILER were used to build the databank of *de novo* consensus sequences.

We then compared the results of our analyses with those obtained with RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html), which combines RECON, RepeatScout, RepeatMasker and TRF and classifies the consensus sequences obtained. For *D. melanogaster,* RepeatModeler generated a library of 141 consensus sequences, with a sensitivity of 78% and a specificity of 76%. However, the recovery ratio of RepeatModeler (R_{CC}=21%) was much lower than that of the TEdenovo pipeline (R_{CC}=72%). This indicates that RepeatModeler recovers only a few

intact TE reference sequences. Similar results were obtained for *A. thaliana,* with all three measurements showing lower values with RepeatModeler than with TEdenovo (table S11). Running the TEannot pipeline with the consensus sequences generated by RepeatModeler resulted in a sensitivity markedly lower than that for the annotations obtained with the *de novo* library from the TEdenovo pipeline, although specificity was slightly higher (table S12). The combination of several tools is therefore not sufficient in itself to improve the results. The way in which the tools are chosen and combined is determinant. We conclude that the TEdenovo pipeline achieves a good balance between sensitivity and specificity in the *de novo* construction of a TE databank from raw genomic sequences.

We therefore developed the REPET package available online (http://urgi.versailles.inra.fr/download/repet/), into which we integrated both the TEdenovo and TEannot pipelines, with the TE classifier described above corresponding to the final step of the TEdenovo pipeline. The REPET package was specifically designed to improve speed and tractability by (i) interacting with MySQL tables at several key points to take advantage of the SQL language, and (ii) automatically handling jobs launched in parallel on a cluster *via* free batch-queuing systems, such as the Sun Grid Engine (SGE, http://www.sun.com/software/sge/), relaunching jobs in cases of cluster node failure. This package is thus specifically implemented for computationally intensive, genome-wide analyses that do not compromise the biological relevance of the results. As already pointed out by several authors, tools for *de novo* TE identification are "quite difficult to use, indicating the need for better user interfaces and auto-optimization" [32]. We facilitated the use of our tools, by concealing technical details behind interfaces, one per pipeline. The user is also provided with access to a detailed tutorial and a configuration file with default parameters.

Identification of structural variation within TE families and manual curation

After annotating the TE copies in both genomes, we investigated the structural diversity within TE families, as represented by the large number of *de novo* consensus. We focused on the TE

families for which the “knowledge-based” consensus was fully recovered by only one clustering method, GROUPER or RECON, as shown in figure 2 (tables S13 and S14). Indeed, the failure of one method (either GROUPER or RECON) to recover all the TEs would illustrate differences in the ability of these methods to take TE structural variations into account. For each of these TE families, we retrieved genomic copies detected by the *de novo* consensus and built multiple alignments (see Methods). Figure 4 (cases a, b, c and d) provides an overview of several of these multiple alignments displaying extensive structural variations. In almost all TE families, differences between genomic copies were observed, due to substitutions and indels. The clustering method generated different clusters as a function of these differences and the fragmentation of the copies. Depending on the specific features of each algorithm, the consensus will correspond to the complete TE reference sequence or a truncated version of that sequence.

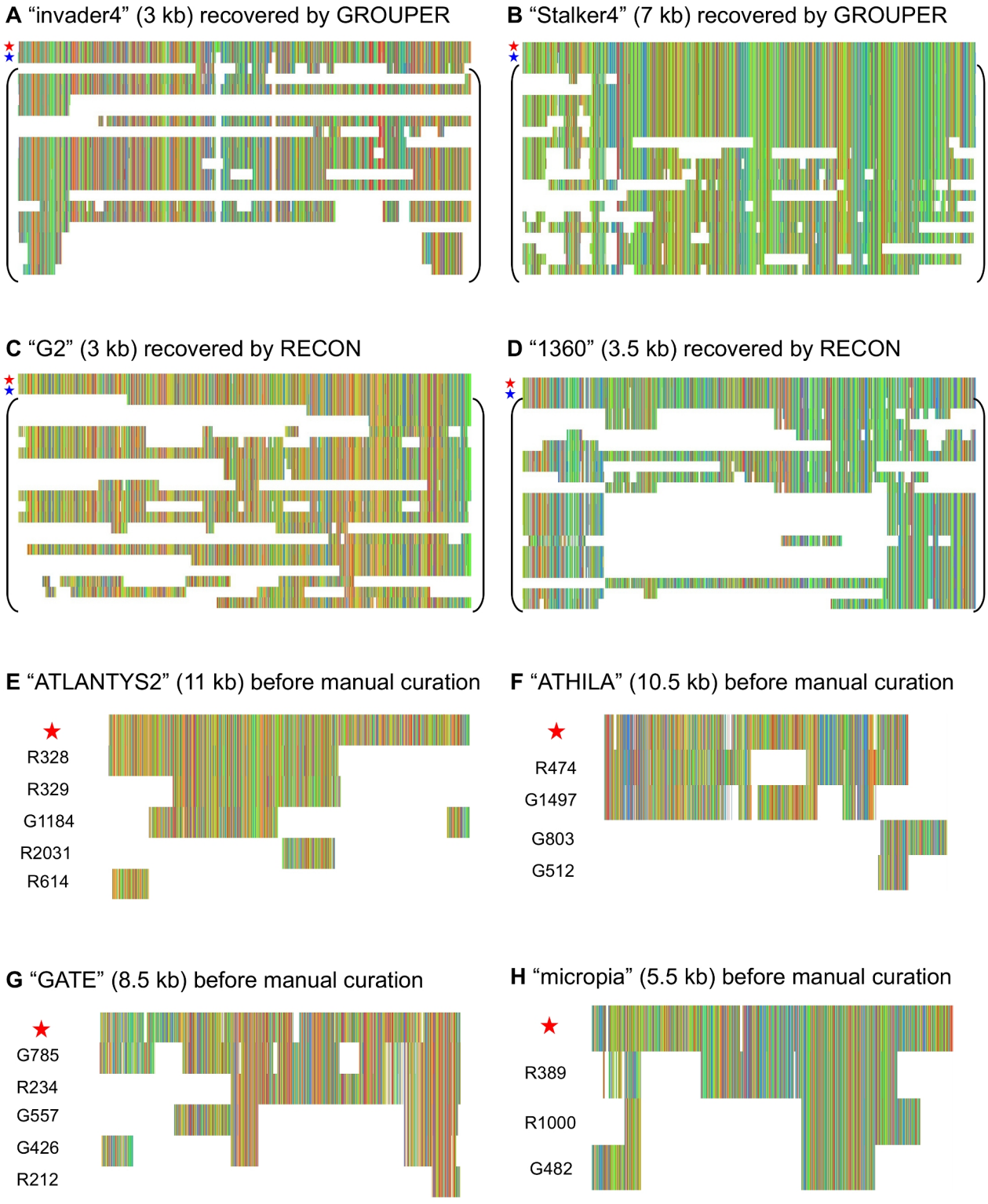


Figure 4. Extensive structural variations within several TE families. Each image provides an overview of a multiple alignment, a column being in one color if all the residues within it are

identical. In all the images, the first sequence in the multiple alignment (red star) is the TE reference sequence from a public databank (BDGP or Repbase). For alignments A to D, the second sequence (blue star) is the only *de novo* consensus in which the TE reference sequence is fully recovered by only one clustering method. All sequences below (in brackets) are TE genomic copies found by the *de novo* consensus analysis. For alignments E to H, the sequences below the TE reference sequence are *de novo* consensus that require manual curation. Beside is indicated the program that build them, “R” for RECON and “G” for GROUPER.

We then looked for features particular to the TE families for which the reference sequence was fully recovered by only one clustering method, being recovered only partially with another method. We first compared the classification of these TE reference sequences. In *D. melanogaster*, the four TE reference sequences fully recovered only by GROUPER were all LTR retrotransposons, whereas the nine TE reference sequences fully recovered only by RECON comprised four LTR retrotransposons, three LINE retrotransposons and two DNA transposons. In *A. thaliana*, the eight TE reference sequences fully recovered only by GROUPER comprised five LTR retrotransposons and three DNA transposons, whereas the fifteen TE reference sequences fully recovered only by RECON comprised six LTR retrotransposons, eight DNA transposons and one Helitron. There were therefore no clear differences in the recovery of full-length TE reference sequences obtained with different clustering methods. We also looked for differences in terms of copy number. However, the TE reference sequences fully recovered only by GROUPER or only by RECON had similar numbers of full-length and truncated copies (data not shown). Comparisons of *de novo* consensuses and “knowledge-based” consensuses, as in table 1, showed that RECON was systematically the most sensitive method, whereas GROUPER was systematically the most specific. As a result, several TE reference sequences displayed partial matches with *de novo* consensuses from RECON only. Moreover, almost 50% of the *de novo* consensus sequences from RECON lacked both boundaries of the TE reference sequences, versus less than 20% of the

consensus sequences from GROUPER. Conversely, a *de novo* library from GROUPER is likely to match with fewer TE reference sequences, but, when a match does occur, at least one boundary of the reference is likely to be correctly retrieved. Consequently, GROUPER and RECON are truly complementary, making the combined approach implemented in the TEdenovo pipeline very efficient.

Our TE classifier can combine *de novo* consensus sequences from various sources, but manual curation is still required in some cases, particularly when there are several consensus sequences from the same TE family (table S15). Figure 4 (alignments E to H) shows several examples in which manual curation of the *de novo* consensus sequences identified by GROUPER and RECON for a given TE family improves the recovery of the reference sequence. For instance, in figure 4-E, all consensus sequences from RECON are truncated before the 3' LTR of the ATLANTYS2 element whereas the G1184 consensus sequence from GROUPER connects this LTR with part of the internal region of the element. In figure 4-F, two consensuses can be removed, G512 from GROUPER, a a solo-LTR, and G803, which is chimeric. Moreover, we can add the internal region of the element present in G1497, which lacks the 3' LTR, to the truncated R474 from RECON. A similar strategy can be applied in figures 4-G and 4-H. Thus, although it is not always possible to recover the full reference sequence, we can improve the final *de novo* consensus by manual curation, making use of the multiple alignment of classified consensuses to guide informed decision-making.

The TEdenovo pipeline identified putative new TEs in the *D. melanogaster* and *A. thaliana* genomes, despite the intensity with which these genomes have been studied and annotated manually. Indeed, we found *de novo* consensus sequences that were classified as "complete" TE but had no single match in blastn with known TE reference sequences from the BDGP or Repbase Update databanks. Three such sequences were identified in *D. melanogaster* (two from GROUPER and one from RECON), and four in *A. thaliana* (three from RECON and one from PILER). For example, in *D. melanogaster*, a 7.8 kb *de novo* consensus was classified as a

"complete" LTR retrotransposon on the basis of its two long terminal repeats (each 510 nt long) and its matches with known LTR retrotransposons from Repbase Update in tblastx and blastx analyses. Furthermore, this consensus matched two HMM profiles corresponding to an integrase and an aspartic proteinase. There were four full-length copies in the genome, two of which retained their target site duplications. This *de novo* consensus sequence is probably non autonomous, as it lacks matches with HMM profiles corresponding to other LTR retrotransposon genes, and, notably, displays no match with a reverse transcriptase. However, it has enough of the typical properties of TEs and enough full-length copies to correspond to a true TE, rather than a mere segmental duplication containing TE fragments.

DISCUSSION

Combining approaches to build high-quality TE *de novo* consensus sequence databanks for sequenced genomes

Transposable elements play a key role in the structure and evolution of genomes, but their impact remains to be fully elucidated. If we are to make use of the increasing numbers of genome sequencing projects to improve our understanding of TE biology, we will need an efficient, automatic *de novo* approach for the annotation of genome TE content. Several methods praised for their rapidity and low memory requirements, such as the MDR index [50], and P-clouds [51], provide a good overview of the repeat content of a given genome. However, they are not precise enough to provide insight into the functional impact of TEs (see for instance [52] and [53]). As such, these methods are good starting points but are inadequate for a full exploration of biological questions relating to TEs.

The typical process of TE annotation, as we see it, should involve the identification of all TE copies throughout the genome and proceed in two phases. The first step is the construction of a *de novo* library of TE consensus sequences representing the TE families present in the genome. This library is then used to mine all the TE fragments present in the genome. The TE copies are then reconstructed, usually nested within each other, to unravel their intricate evolution. We describe here a new strategy for constructing databanks of TE consensus sequences by *de novo* methods. Our results for the *D. melanogaster* and *A. thaliana* genomes show that an approach combining several clustering programs provides the best outcome, in terms of sensitivity and specificity, for the identification of TE families and the annotation of their TE copies.

Previous studies have already compared tools for *de novo* TE identification, either to propose a new algorithm (as for RECON, PILER and RepeatScout) or to help experimental biologists to use the most suitable tools for their analysis [32]. In most of these studies, the results were validated

by comparing the final TE copy annotations in terms of genome coverage, rather than by evaluating the accuracy of *de novo* TE consensus sequence identification. Little attention has been paid to analyses of the *de novo* databank itself. The use of consensus sequences corresponding to truncated or artifactual TEs would clearly bias subsequent annotations. We therefore propose three indices for measuring the quality of *de novo* databanks, with respect to reference databanks. This combined approach was validated by considering the proportion of *de novo* consensus sequences corresponding to full-length reference sequences, with well-defined boundaries.

Classification of TE sequences on the basis of their biological features and the curation of TE *de novo* databanks

The tools typically used for the *de novo* identification of TE families provide no additional meta-data about the consensus sequences they build. Instead, the user has to use other tools to obtain a putative classification of the sequences obtained. Here, we implemented a combined *de novo* approach for the identification of TE families and developed a dedicated procedure for classifying sequences on the basis of TE features. We based our TE classifier on the classification proposed in [5], from class to superfamily, combining various sources of information about TEs in a modular manner. This procedure was designed to be run in isolation for the analysis of any databank of putative TE-related sequences and for integration into the TEdenovo pipeline, to facilitate its use.

The current version of our tool, TEclassifier, requires further improvement for the classifiers of less well known TEs, such as DIRS-like elements, Penelope-like elements, Crypton elements and Polintons/Mavericks. However, unlike tools like TEclass [54], which makes use of the profiles of frequent *k*-mers from the Repbase Update databank, our TE classifier guides output classification by detailing the precise TE features present on the analyzed sequences, such as terminal repeats and matches with known TEs. Our tool and REPclass [55] are very similar, but our TE classifier has several advantages. Being part of the REPET package, it benefits from the architecture of this

package and is therefore able to handle large databanks, which is often required when studying large genomes. Moreover, our TE classifier not only classifies putative TEs, it also makes it possible to filter out false-positives and to remove redundant sequences present in the input databank.

In a wider perspective, we considered manual curation, a key topic in genome annotation. Computer-based predictions for the annotations of protein-coding genes are now of high quality, but it remains difficult to predict exon-intron boundaries correctly [56]. Major efforts have been and are being made to improve the automatic annotation of protein-coding regions, but such efforts have been much less intensive for other parts of the genome, including TEs. In this study, we tried to fill this gap by designing the REPET package for the comprehensive association of all the meta-data obtained for a given TE family: (i) the *de novo* consensus sequences, (ii) the TE features used to classify them, (iii) the all-by-all comparisons from which they were built, and (iv) the TE copies they identify on the chromosomes. As a result, the manual curator can make informed decisions based on the biological features at hand, and the curated TE consensus databank can be used to provide a second release of TE annotations.

Efficiency of *de novo* approaches for recovering ancestral TEs and their structural variants from “junk”

At the core of any *de novo* approach lies the possibility of identifying new agents in a given system by searching for their fundamental properties. In our case, we were confronted with raw whole-genome sequences containing numerous repeated sequences of different kinds, some of which have specific features and a common ancestry. We were interested in identifying, from among these sequences, TEs — interspersed repeats — and, more particularly, the ancestral sequences that actually transposed. The current copies of such sequences are likely to be divergent, fragmented and nested within each other, as TE families typically display extensive structural

variation (figure 4 A-D). We showed in this study that the *de novo* approach could recover some full-length TE sequences by correctly collecting together their fragments, even if no TE feature could be used to classify them. This was the case here for several Helitron reference sequences in the *A. thaliana* genome. Similarly, we were able to identify putative new non autonomous TEs in the *D. melanogaster* and *A. thaliana* genomes that had all the essential features of TEs but were nonetheless absent from the reference databanks.

Efficient and robust tools are essential to keep pace with current whole-genome sequencing programs. Our analyses were performed on small model genomes. We therefore decided to extend the testing of our pipelines to several larger genomes, such the 400 Mb genome of rice, *Oryza sativa* (Japonica group cultivar Nipponbare). The TE *de novo* consensus sequences recovered included the Pong DNA transposon and its MITE relative mPing [57] (data not shown), which were fully recovered intact as two distinct consensus sequences. For all projects in which the aim is to sequence large and highly repetitive genomes (*e.g.* barley, hexaploid wheat), computational tools, such as the TEdenovo pipeline presented here, are likely to become increasingly useful for increasing our knowledge of the evolution of genome structure and the functional impact of TEs on neighboring genes.

METHODS

Genome sequences and TE reference databanks as benchmark

In 2000, an international consortium driven by the Berkeley Drosophila Genome Project (BDGP) sequenced, assembled and annotated the genome of an isogenic *y; cn bw sp* strain of *D. melanogaster* [58], this work now being continued at FlyBase (http://flybase.org/). In this study, we used the 118.4 Mb release 4 genome sequence corresponding mostly to euchromatin regions. Release 5 became available in 2006 and corresponds to release 4 with 50.3 Mb of additional heterochromatic regions and minor sequence corrections. We chose to work on release 4, which remains the *D. melanogaster* genomic sequence best annotated for TEs [24,25], despite the availability of the later sequence.

Flybase provides a set of natural TE sequences experimentally identified in several Drosophilidae genomes. The latest version of this data set (bdgp9.41) consists of 179 sequences corresponding to different TE families, 126 of which have been detected in *D. melanogaster* strains. Each of these 126 sequences corresponds either to a TE copy of the given family present in *D. melanogaster* or to a consensus based on TE copies for the family concerned. In the latter case, consensuses were built either as a mosaic of TE copies or following a “majority rule” from a multiple alignment. In this study, these 126 sequences are referred to as “reference TEs”. The aim of the *de novo* approach presented in this article is to reconstruct such a library of TE sequences.

In 2000, 115 Mb of the 125 Mb genome of *A. thaliana* accession Columbia had been sequenced, assembled and annotated [59]. For our analyses, we used the 119 Mb of *A. thaliana* release 9 genome sequences available from the TAIR website (http://www.arabidopsis.org/). For *A. thaliana*, a similar set of reference TEs is present in the Repbase Update databank. We used the library derived from this databank for previous studies [29] and containing 318 TE sequences.

We annotated both genomes with the TEannot pipeline [24] using the reference databanks cited above. We then constructed a consensus for each reference sequence, from the multiple alignment of copies when at least three copies longer than 100 bp were present (table S16). For *D. melanogaster*, a consensus could be obtained for only 117 of the 126 reference sequences in the BDGP databank, referred to as a "knowledge-based" consensus: 68 of these sequences corresponded exactly to their reference element, the others being truncated. For *A. thaliana*, it was possible to construct 305 "knowledge-based" consensuses for the 318 TE reference sequences in the Repbase Update databank: 154 corresponded to the entire element and the others were truncated. These consensuses were collected into "knowledge-based" databanks. As such, they represent the upper limit of what can be retrieved with the *de novo* methods compared in this study (table S17).

De novo detection of TEs

A three-step approach forms the backbone of our analysis: (i) self-alignment of the genomic sequences, (ii) clustering of the resulting matches, and (iii) multiple alignments of each cluster to build a consensus (figure S1). Different programs may be used at each step, leading to several possible combinations. For each combination, we compared the *de novo* consensus with the set of reference TEs. Other methods are also available for constructing a consensus for comparison with the "knowledge-based" databanks constructed as described above.

When comparing a *de novo* databank with a "knowledge-based" databank, it is not possible to calculate the number of false-negatives, *i.e.* the *de novo* consensus sequences incorrectly classified as not related to a TE because these consensus sequences are not present in the *de novo* library. The usual definitions of sensitivity and specificity are therefore not directly applicable. We estimated sensitivity by calculating the proportion of "knowledge-based" consensus sequences matching *de novo* consensus sequences (noted S_n*). We then estimated specificity by considering the proportion of *de novo* consensus sequences matching "knowledge-based"

consensus sequences (noted S_p*). The observation of a *de novo* consensus that matches a "knowledge-based" consensus, both aligned along their entire lengths (+- 5%) indicates that this *de novo* consensus retrieved the exact and entire "knowledge-based" consensus. A recovery ratio (noted R_{CC} for "complete-complete ratio") can be calculated as the number of "knowledge-based" consensus sequences exactly retrieved by a *de novo* consensus, divided by the number of reference sequences exactly retrieved by a "knowledge-based" consensus. In *D. melanogaster,* 68 reference sequences of 126 in the BDGP databank) were exactly retrieved by a "knowledge-based" consensus, and 154 such sequences (of the 318 in Repbase Update) were exactly retrieved in *A. thaliana*. As an illustration, in *D. melanogaster*, if a given *de novo* method recovers 60 entire "knowledge-based" consensus sequences from a given genome, R_{CC} would be 88% (60/68).

First step: self-alignment of the genomic sequences

The first step is the self-alignment of the genomic sequences in an all-by-all manner. We compared two programs, BLASTER [36] and PALS [37] for this purpose. For BLASTER, we kept matches with E-value below 10^{-300}, a length exceeding 100 bp and an identity exceeding 90% (parameters "-E 1e-300 -L 100 -I 90"). For PALS, we used the default parameters: identity exceeding 94% and a length of more than 400 bp (parameters "-length 400 -pctid 94"). If the same parameters were used for BLASTER (identity > 90% and length > 100 bp), the computations took too long, even when launched in parallel. If the parameters length > 100 bp and identity > 94% were used with both programs, BLASTER still gave better results (data not shown).

When all the matches had been retrieved, we discarded all those that were more than 20,000 bp long. This procedure was designed to filter out repeats corresponding to long segmental duplications. The issue of short segmental duplications was addressed in the next step. To speed up the computations, the input genome sequences were cut into chunks. Each chunk was then

aligned against the whole set of chunks, in parallel. Finally all the matches were collected together and redundant matches were filtered out.

Second step: clustering of the resulting pairwise alignments

The matches obtained in the first step were then clustered, and we tested three different programs for this step: GROUPER ([36], parameters "-j -C 0.95 -Z 3 -X 2 -G -1"), RECON ([39], default parameters) and PILER ([40], parameter "-trs"). The previous version of GROUPER suffered from the high redundancy among clusters, preventing its usage in practice on large genomes. Therefore, we implemented a new procedure that specifies, during the single-link clustering step, if a chain of matches is fully included within another one, as is often the case with non-autonomous TEs with respect to their autonomous counterparts. Once the clusters are built, we now remove those having less than a given number of members not included in any others (option "-X 2"). All these programs, GROUPER, RECON and PILER, return a set of clusters to which we applied several filters. First, we removed the clusters with fewer than three members. This discarded most of the short segmental duplications. Second, for large clusters, we retained only the 20 longest sequences, as keeping all the sequences would not add much information for the building of a consensus and would even introduce noise (data not shown). Third, for GROUPER, we filtered the sequences resulting from connected matches (*i.e.* chains) with a cumulative length greater than 20 kb and spanning more than 30 kb of the genome, as such sequences probably corresponded to segmental duplications.

Third step: multiple alignments and consensus construction

Finally, for each cluster, a multiple sequence alignment (MSA) was constructed, from which a consensus was derived. We compared several programs: MAP ([42], parameters "gap-size=50 mismatch=-8 gap-open=16 gap-extend=4"), CLUSTAL-W ([43], default parameters), MAFFT ([44],

parameter "--auto") and PRANK ([45], parameter "-F"). The consensus was constructed by applying a majority rule discarding columns in which all but one sequence had a gap.

Other approaches

Other methods have also been proposed. We tested REPEATSCOUT [46] and RepeatModeler (Smit and Hubley unpublished), both with default parameters.

Classification of TE consensus and the elimination of redundancy

We implemented a two-step TE classifier. The first step detects structural features of the consensus, such as terminal repeats, tandem repeats, and polyA or SSR-like tails, using programs from the REPET package (TRsearch, polyAtail) or elsewhere (TRF, [60]). It also searches for matches with known TEs, by blastx, tblastx and blastn analysis, and for matches with known genes from the host's genome, by blastn analysis. The second step is based on a decision tree (figure 3) classifying each consensus according to its length and features. The classification and the evidence underlying the classification are delivered as output. Any program looking for other TE features, such as Helitron hairpins, could easily be integrated into this framework.

We eliminated redundant consensus sequences, by discarding all those included within another consensus sequence, for at least *x*% of their length, and with at least *y*% identity. We tested several values for *x* and *y*: 95-98, 90-90 and 80-80. We tested this procedure with and without taking into account the classification of the sequences.

Genome-wide annotation of TE copies

A combined pipeline for the genome-wide annotation of TE copies is already available [24]. As input, it takes the genome sequences and a databank of TE sequences, typically that generated by the TEdenovo pipeline. It then launches BLASTER [36], RepeatMasker (http://www.repeatmasker.org/) and CENSOR [61], to map the TE sequences against the genome.

False-positives are filtered out by applying the same procedure to shuffled genomic sequences. More precisely, the genomic sequences are randomized using the “shuffle” program of the HMMER package (http://hmmer.janelia.org/). The TE reference sequences are then mapped onto these shuffled sequences with BLASTER, RepeatMasker and CENSOR, with the score for each match recorded. Finally, the matches between the TE reference sequences and the true genomic sequences are filtered according to these scores. Whereas in a previous version of the pipeline we used the highest score obtained on randomized genomic sequences to filter false-positives, we now use the 95% quantile of the scores obtained on the randomized sequences. This improvement prevents excessive filtering, using a single, very good match on randomized sequences, much better than most others. This change slightly increases TE coverage over previous estimations.

Once the matches were filtered, we began to reconstruct the TE copies, to obtain a true annotation of TE copies and not of TE fragments only. In this pipeline, two steps are used to connect several TE fragments belonging to the same TE copy: MATCHER and the “long join” procedure. As in the *de novo* library, a TE family may be represented by several consensus sequences corresponding to each of its structural variants. We improved these tools to take this into account. In terms of vocabulary, we define a “TE fragment” as a match between a TE consensus sequence and a genomic sequence, whereas we define a “TE copy” as a chain of matches, each match in the chain being a TE fragment. Note that a full-length TE copy may correspond to two TE fragments, which, when connected together, correspond to the full TE consensus sequence.

In the previous version of MATCHER, we began by combining the matches found by all three algorithms mentioned above. When two consensus sequences overlapped at a given locus, we retained the sequence with the highest score and truncated the other. We then connected the remaining matches by dynamic programming. In the current version of MATCHER, we first connect the matches by dynamic programming and then filter out overlapping chains of matches. A match that might have been filtered out in the previous version may not be filtered out in the current version, if it is chained with another match, thereby improving fragment connections. TE

annotations would be improved by taking into account chains of matches (whole TE copies) rather than TE fragments (single matches).

Once matches are connected by MATCHER, the TEannot pipeline also detects microsatellites by launching and combining the results of TRF [60], Mreps [62] and RepeatMasker. All TE copies are then combined with microsatellite coordinates to filter out short TE matches fully recovered by microsatellites.

In the same spirit, we improved the “long join” procedure. We previously sorted the chains of matches on the basis of length, as a proxy for the age of the TE copies. The rationale behind this was that a TE copy may disappear slowly due to small deletions, becoming shorter over time. We now estimate the age of a TE copy by calculating the ratio of match identity to match length, summing this ratio for all matches in the chain.

Finally, we compared the annotations obtained with *de novo* libraries and reference databanks, by calculating genome coverage and TE copy number, together with sensitivity and specificity, in terms of nucleotide overlaps (figure S3). A high sensitivity indicates that the annotation based on *de novo* consensus sequences misses few TE nucleotides (false-negatives). A high specificity indicates that the *de novo* annotation identifies few non-TE nucleotides (false-positives).

Reconstruction of TE families

We compared the patterns of diversification between TE families, by mining the genome with the *de novo* consensus sequences, using BLAT [63] for the rapid identification of well conserved genomic copies. We then constructed a multiple alignment with the TE reference sequence from the public databank, the *de novo* consensus sequence and the genomic copies identified with this sequence (see figure 4 A-D). For the identification of TE families represented by several *de novo* consensus sequences, we clustered *de novo* consensus sequences with BLASTCLUST from the NCBI-BLAST suite ([64], parameters “-S 0 -L 0.8 -b F -p F”). We then added the best TE reference sequence corresponding to each cluster, and finally built a multiple alignment. The addition of the

reference sequence after clustering prevents the *de novo* consensus sequences from being clustered together solely because they overlap with the same reference sequence. This procedure can therefore be used to assist manual curation for newly sequenced genomes without known reference sequences. Multiple alignments were checked by eye, using Jalview [65].

ACKNOWLEDGMENTS

We wish to thank Emmanuelle Permal, Joëlle Amselem and Victoria Dominguez for helpful comments on the *de novo* approach, Olivier Inizan for his work on the REPET package, Claire Hoede for her help with the TE HMM profiles, and Isabelle Luyten and Sébastien Reboux for their help with the computer infrastructure required for such an analysis. We also thank the GenOuest platform for allowing us to run our analyses on their computer cluster.

REFERENCES

1. Orgel L, Crick F (1980) Selfish DNA: the ultimate parasite. Nature 284: 604-607.

2. Brookfield JFY (2005) The ecology of the genome - mobile DNA elements and their hosts. Nature Reviews Genetics 6: 128-136.

3. Lynch M, Conery J (2003) The origins of genome complexity. Science 302: 1401-1404.

4. Finnegan D (1989) Eukaryotic transposable elements and genome evolution. Trends in Genetics 5: 103-107.

5. Wicker T, Sabot F, Hua-Van A, Bennetzen J, Capy P et al. (2007) A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8: 973-982.

6. Wessler S (1995) LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Current Opinion in Genetics and Development 5: 814-821.

7. Yang G, Nagel DH, Feschotte C, Hancock CN, Wessler SR (2009) Tuned for Transposition: Molecular Determinants Underlying the Hyperactivity of a Stowaway MITE. Science 325: 1391-1394.

8. Gray Y (2000) It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. Trends in Genetics 16: 461-468.

9. Eichler E, Sankoff D (2003) Structural dynamics of eukaryotic chromosome evolution. Science 301: 793-797.

10. Coghlan A, Eichler E, Oliver S, Paterson A, Stein L (2005) Chromosome evolution in eukaryotes: a multi-kingdom perspective. Trends in Genetics 21: 673-682.

11. Paux E, Sourdille P, Salse J, Saintenac C, Choulet F et al. (2008) A Physical Map of the 1-Gigabase Bread Wheat Chromosome 3B. Science 322: 101-104.

12. McClintock B (1956) Controlling elements and the gene. Cold Spring Harbor Symposia on Quantitative Biology 21: 197-216.

12. McClintock B (1956) Controlling elements and the gene. Cold Spring Harbor Symposia on Quantitative Biology 21: 197-216.

13. Slotkin R, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics 8: 272-285.

14. Martin A, Troadec C, Boualem A, Rajab M, Fernandez R et al. (2009) A transposon-induced epigenetic change leads to sex determination in melon. Nature 461: 1135-1138.

15. Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics 9: 397-405.

16. Herpin A, Braasch I, Kraeussling M, Schmidt C, Thoma E et al. (2010) Transcriptional Rewiring of the Sex Determining dmrt1 Gene Duplicate by Transposable Elements. PLoS Genetics 6: e1000844+.

17. Cordaux R, Udit S, Batzer M, Feschotte C (2006) Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proceedings of the National Academy of Sciences 103: 8101-8106.

18. Santangelo A, de Souza F, Franchini L, Bumaschny V, Low M et al. (2007) Ancient Exaptation of a CORE-SINE Retroposon into a Highly Conserved Mammalian Neuronal Enhancer of the Proopiomelanocortin Gene. PLoS Genetics 3: .

19. Agrawal A, Eastman QM, Schatz DG (1998) Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature 394: 744-751.

20. Zhou L, Mitra R, Atkinson PW, Burgess Hickman A, Dyda F et al. (2004) Transposition of hAT elements links transposable elements and V(D)J recombination. Nature 432: 995-1001.

21. Schuster S (2008) Next-generation sequencing transforms today's biology. Nature Methods 5: 16-18.

22. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112-1115.

23. Bourque G (2009) Transposable elements in gene regulation and in the evolution of vertebrate genomes. Current Opinion in Genetics and Development 19: 607-612.

24. Quesneville H, Bergman C, Andrieu O, Autard D, Nouaud D et al. (2005) Combined evidence annotation of transposable elements in genome sequences. PLoS Computational Biology 1: 166-175.

25. Bergman C, Quesneville H, Anxolabehere D, Ashburner M (2006) Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. Genome Biology 7: R112+.

26. Nene V, Wortman JR, Lawson D, Haas B, Kodira C et al. (2007) Genome Sequence of Aedes aegypti, a Major Arbovirus Vector. Science 316: 1718-1723.

27. Cuomo CA, G\uldener U, Xu JR, Trail F, Turgeon BG et al. (2007) The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization. Science 317: 1400-1402.

28. the 12 Drosophila Genomes Consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203-218.

29. Buisine N, Quesneville H, Colot V (2008) Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. Genomics 91: 467-475.

30. Abad P, Gouzy J, Aury JM, Castagnone-Sereno P, Danchin EGJG et al. (2008) Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. Nature Biotechnology : 909-915.

31. the International Aphid Genomics Consortium (2010) Genome sequence of the pea aphid Acyrthosiphon pisum. PLoS Biology 8: e1000313+.

32. Saha S, Bridges S, Magbanua ZV, Peterson DG (2008) Empirical comparison of ab initio repeat finding programs. Nucl. Acids Res. 36: 2284-2294.

33. Jiang N, Jordan IK, Wessler SR (2002) Dasheng and RIRE2. A Nonautonomous Long Terminal Repeat Element and Its Putative Autonomous Partner in the Rice Genome. Plant Physiol. 130: 1697-1705.

34. Quesneville H, Nouaud D, Anxolabehere D (2006) P elements and MITE relatives in the whole genome sequence of Anopheles gambiae. BMC genomics 7: 214+.

35. Bergman CMM, Quesneville H (2007) Discovering and detecting transposable elements in genome sequences. Briefings in Bioinformatics 8: 382-392.

36. Quesneville H, Nouaud D, Anxolabehere D (2003) Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes. Journal of Molecular Evolution 57.

37. Rasmussen K, Stoye J, Myers E (2006) Efficient q-gram filters for finding all epsilon-matches over a given length. Journal of Computational Biology 13: 296-308.

38. Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. Journal of Molecular Biology 215: 403-410.

39. Bao Z, Eddy S (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Research 12: 1269-1276.

40. Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. Bioinformatics 21: i152-158.

41. Notredame C (2007) Recent evolutions of multiple sequence alignment algorithms. PLoS Computational Biology 3: e123+.

42. Huang X (1994) On global sequence alignment. Computer Applications in the Biosciences 10: 227-235.

43. Thompson J, Higgins D, Gibson T (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22: 4673-4680.

44. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. Briefings in Bioinformatics 9: 286-298.

45. Loytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. Proceedings of the National Academy of Sciences 102: 10557-10562.

46. Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21: i351-358.

47. Mayr E (1981) Biological classification: toward a synthesis of opposing methodologies. Science 214: 510-516.

48. Jurka J, Kapitonov V, Pavlicek A, Klonowski P, Kohany O et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research 110: 462-467.

49. Tempel S, Nicolas J, El Amrani A, Couée I (2007) Model-based identification of Helitrons results in a new classification of their families in Arabidopsis thaliana. Gene 403: 18-28.

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

51. Gu W, Castoe T, Hedges D, Batzer M, Pollock D (2008) Identification of repeat structure in large genomes using repeat probability clouds. Analytical Biochemistry 380: 77-83.

52. Newman J, Bailey A, Fan H, Pavelitz T, Weiner A (2008) An abundant evolutionarily conserved CSB-PiggyBac fusion protein expressed in Cockayne syndrome. PLoS Genetics 4: .

53. Teixeira F, Heredia F, Sarazin A, Roudier F, Boccara M et al. (2009) A role for RNAi in the selective correction of DNA methylation defects. Science 323: 1600-1604.

54. Abrusan G, Grundmann N, DeMester L, Makalowski W (2009) TEclass--a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics 25: 1329-1330.

55. Feschotte C, Keswani U, Ranganathan N, Guibotsy M, Levine D (2010) Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome Biol Evol 2009: 205-220.

56. Brent M (2008) Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nature Reviews Genetics 9: 62-73.

57. Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR et al. (2003) An active DNA transposon family in rice. Nature 421: 163-167.

58. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD et al. (2000) The genome sequence of Drosophila melanogaster. Science 287: 2185-2195.

59. the Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796-815.

60. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research 27: 573-580.

61. Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR--a program for identification and elimination of repetitive elements from DNA sequences.. Comput Chem 20: 119-121.

62. Kolpakov R, Bana G, Kucherov G (2003) mreps: efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Research 31: 3672-3678.

63. Kent W (2002) BLAT--the BLAST-like alignment tool. Genome Research 12: 656-664.

64. Dondoshansky I (2002) Blastclust (NCBI Software Development Toolkit).

65. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191.

SUPPLEMENTARY MATERIALS

Figure S1. Flow chart of the three first steps of the TEdenovo pipeline.

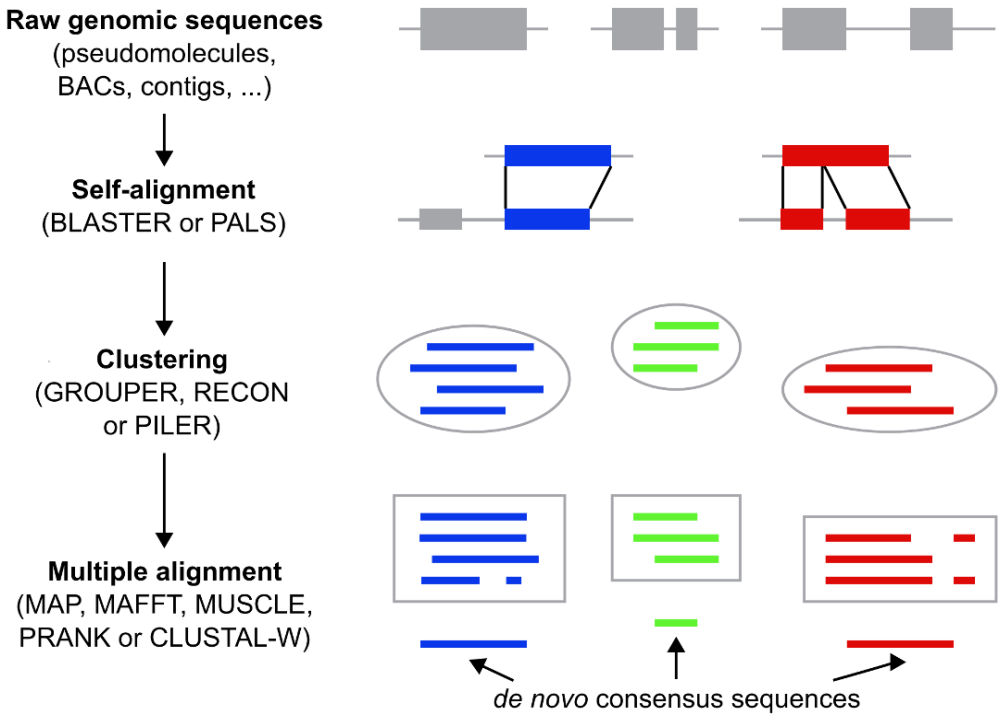


Figure S2. Flow chart of the TEannot pipeline.

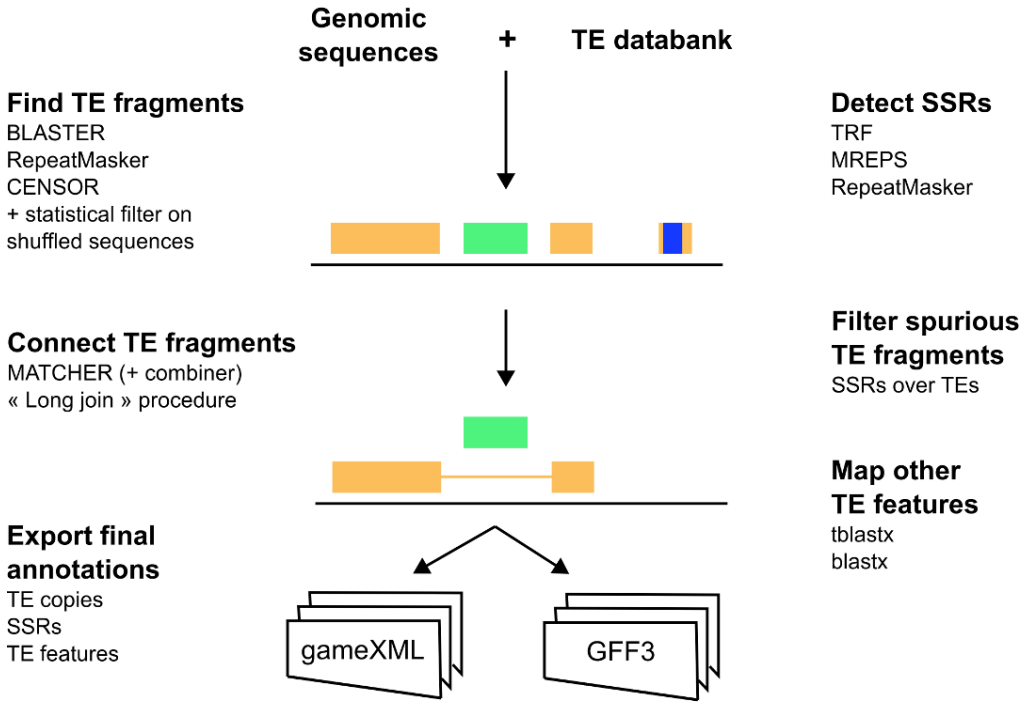
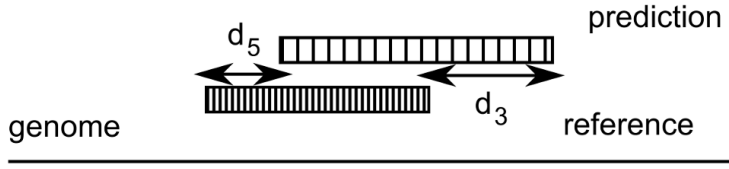


Figure S3. Comparison of two TE annotations in terms of match boundaries.



	$d_5 \leqslant 1$ bp	$1 \text{ bp} < d_5 \leqslant 10$ bp	$d_5 > 10$ bp
$d_3 \leqslant 1$ bp	exact	near exact	one-side exact
$1 \text{ bp} < d_3 \leqslant 10$ bp	near exact	equivalent	near equivalent
$d_3 > 10$ bp	one-side exact	near equivalent	similar

Table S1. Comparative analysis of self-alignment programs

Genome	Genome length (bp)	Program	Number of matches	Length on the genome (bp)	Genome coverage
D. mel.	129,919,500	BLASTER	109,882	9,636,659	7.41%
		PALS	105,059	9,590,737	7.38%
A. tha.	119,146,348	BLASTER	103,728	16,063,506	13.48%
		PALS	51,023	12,547,315	10.53%

Table S2. Comparative analysis of clustering programs

Genome	Self-alignment	Clustering	Number of clusters	CV of sequence length (mean +- sd)
D. mel.	BLASTER	GROUPER	730	0.463 +- 0.190
		RECON	451	0.711 +- 0.181
		PILER	120	0.566 +- 0.172
	PALS	GROUPER	542	0.466 +- 0.193
		RECON	485	0.742 +- 0.203
		PILER	106	0.567 +- 0.166
A. tha.	BLASTER	GROUPER	1428	0.509 +- 0.184
		RECON	1021	0.702 +- 0.186
		PILER	300	0.603 +- 0.146
	PALS	GROUPER	912	0.519 +- 0.182
		RECON	1000	0.713 +- 0.175
		PILER	242	0.631 +- 0.130

Each cluster contains various genomic sequences, which may differ in length. We therefore calculated the coefficient of variation (CV) of sequence length for each cluster. This coefficient, corresponding to the standard deviation divided by the mean, assesses the dispersion of a distribution. It is high in clusters with sequences of very different lengths, and low in clusters with sequences of similar lengths. We first calculated this coefficient for all the clusters obtained with each method. We then calculated the mean and standard deviation of these coefficients of variation between clusters. As shown in the right column of table S2, on average, RECON clusters are more heterogeneous in terms of sequence length (mean CV > 0.702) than clusters from GROUPER (mean CV < 0.519) and PILER (mean CV < 0.631).

Table S3. Comparative analysis of multiple alignment programs

Genome	Self-alignment	Clustering	Multiple alignment	S_n*	S_p*	R_{CC}
D. mel.	BLASTER	GROUPER	MAP	80.34%	85.89%	66.20%
		RECON		92.31%	73.17%	66.20%
		PILER		62.39%	84.17%	51.50%
		GROUPER	CLUSTAL-W	80.34%	85.89%	66.20%
		RECON		91.45%	72.06%	20.60%
		PILER		62.39%	84.17%	51.50%
		GROUPER	MAFFT	78.63%	85.89%	64.70%
		RECON		92.31%	73.17%	54.41%
		PILER		62.39%	84.17%	51.50%
		GROUPER	PRANK	80.34%	85.89%	66.20%
		RECON		92.31%	72.95%	61.80%
		PILER		62.39%	84.17%	51.50%
A. tha.	BLASTER	GROUPER	MAP	60.33%	82.42%	39.00%
		RECON		73.77%	61.70%	43.50%
		PILER		47.21%	57.33%	32.45%
		GROUPER	CLUSTAL-W	60.00%	82.42%	38.30%
		RECON		73.11%	60.33%	29.20%
		PILER		47.21%	57.33%	32.45%
		GROUPER	MAFFT	60.00%	82.42%	39.00%
		RECON		74.01%	61.21%	40.25%

Genome	Self-alignment	Clustering	Multiple alignment	S_n*	S_p*	R_{CC}
		PILER		47.54%	57.33%	32.45%
		GROUPER		60.00%	82.42%	39.00%
		RECON	PRANK	73.77%	61.61%	39.00%
		PILER		47.21%	57.33%	31.80%

Table S4. Results of the RepeatScout program

Genome	k-mers	Consensus	Mean length (median)	S_n*	S_p*	R_{CC}
D. mel.	5,624,530	1,770	552 (209)	94.87%	57.63%	25.00%
A. tha.	7,569,031	3,417	498 (221)	82.95%	39.54%	13.00%

S_n*: percentage of “knowledge-based” consensus sequences matching a *de novo* consensus sequence

S_p*: percentage of *de novo* consensus sequences matching a “knowledge-based” consensus sequence

R_{CC}: percentage of fully recovered “knowledge-based” consensus sequences

Table S5. Results obtained with a combination of several clustering programs in the *de novo* approach

Genome	Combination of clustering programs	Consensus sequences	S_n*	S_p*	R_{CC}
D. mel.	GROUPER	730	80.34%	85.89%	66.20%
	RECON	451	92.31%	73.17%	66.20%
	PILER	120	62.39%	84.17%	51.50%
	GROUPER + RECON	1181	93.16%	81.03%	79.40%
	GROUPER + PILER	850	81.20%	85.65%	66.20%
	RECON + PILER	571	92.31%	75.48%	73.50%
	GROUPER + RECON + PILER	1301	93.16%	81.32%	79.40%
A. tha.	GROUPER	1428	60.33%	82.42%	39.00%
	RECON	1021	73.77%	61.70%	43.50%
	PILER	300	47.21%	57.33%	32.45%
	GROUPER + RECON	2449	74.43%	73.79%	49.35%
	GROUPER + PILER	1728	63.93%	78.07%	39.60%
	RECON + PILER	1321	74.10%	60.71%	44.15%
	GROUPER + RECON + PILER	2749	74.43%	72.00%	49.35%

Table S6. Length parameters used in the TEclassifier program

Category	Threshold	Parameter value (bp)
Complete LTR retrotransposon	Maximum length	18,000
	Minimum length	3,900
Complete LINE retrotransposon	Maximum length	13,500
	Minimum length	3,400
Complete TIR DNA-transposon	Maximum length (*)	5,000
	Minimum length	900
SINE	Maximum length	500
MITE	Maximum length	500

(*) For *A. thaliana,* we used 13,000 bp for the maximum length of a complete TIR DNA-transposon.

Table S7. Results of the classification of TE sequences from *A. thaliana*

Classification	Reference TEs from Repbase	*De novo* consensus sequences (with redundancy)	*De novo* consensus sequences (without redundancy)
Class I “complete” LTR retrotransposon	121	51	25
Class I “incomplete” LTR retrotransposon	16	400	181
Class I “complete” LINE	6	9	6
Class I “incomplete” LINE	3	31	18
Class I SINE	4	0	0
Class II “complete” TIR transposon	16	14	7
Class II “incomplete” TIR transposon	46	164	99
Class II MITE	11	12	6
Helitron	5	15	11
SSR	0	8	8
Confused	0	6	4
No category	88	2039	910
Host genes	2	0	0
Total	318	2749	1275

Table S8. Consensus sequences matching known genes of *D. melanogaster*

Of the non-redundant, *de novo* consensus sequences from the *D. melanogaster* genome, 11 overlapped host transcripts over at least 95% of their length, with identity levels exceeding 90%.

#	Clustering method (cluster identifier)	All-by-all matches	Length (bp)	Transcript name	Gene name	Additional information
1	GROUPER (42)	3	454	FBtr0072628	Lysozyme B	
2	GROUPER (45)	20	613	FBtr0089196	Kif3C	See Deloger *et al.* Gene 2009
3	GROUPER (521)	4	2532	FBtr0082512	Hsp70Aa	6 Hsp70 genes are present in the genome.
4	GROUPER (607)	3	3524	FBtr0074205	Mucin 14A	Predicted CDS is composed of a repeat domain (125+ repeat units) of 95 amino acids.
5	GROUPER (709)	4	2778	FBtr0072849	sallimus	3 matches not merged by Grouper.
6	GROUPER (771)	3	3665	FBtr0076140	Mucin 68Ca	Protein sequence contains multiple repeats homologous to Salivary glue proteins. 3 matches not merged by Grouper.
7	GROUPER (825)	6	1442	FBtr0077393	Salivary gland secretion 1	Internal highly repetitive region from gene prediction.
8	PILER (110.3)	3	1133	FBtr0082785	Actin 87E	6 actin genes are present in the genome.
9	RECON (26)	3	2022	FBtr0091706	Muscle-specific protein 300	3 matches on the same chunk
10	RECON (302)	6	5238	FBtr0076820	CG32377	87-aa repeat (6 matches on the same chunk)
11	RECON (45)	20	1167	FBtr0089196	Kif3C	See #2 above.

Table S9. Parameters of the tests for redundancy elimination in TEclassifier

Genome	Redundancy elimination	Consensus	S_n*	S_p*	R_{CC}
D. mel.	With redundancy	1301	93.16%	81.32%	79.41%
	95%-98%	593	92.31%	75.72%	77.94%
	90%-90%	494	92.31%	71.86%	70.58%
	80%-80%	428	91.45%	70.33%	70.58%
A. tha.	With redundancy	2749	74.43%	71.99%	49.35%
	95%-98%	1275	74.43%	66.75%	49.35%
	90%-90%	1005	73.44%	61.59%	45.45%
	80%-80%	836	73.44%	57.66%	42.85%

For *D. melanogaster*, the elimination of redundancy with the parameters “90-90” or “80-80” resulted in the loss of six *de novo* consensus sequences that fully recovered “knowledge-based” consensus sequences. These sequences belonged to the 412, invader3, roo, springer, Stalker and Stalker4 families.

For *A. thaliana*, the elimination of redundancy with the parameters “90-90” and “80-80” resulted in the loss of, respectively, six and ten *de novo* consensus sequences that fully recovered “knowledge-based” consensus sequences. These sequences belonged to the ATCOPIA49, ATHILA4B, ATHILA4D, ATREP10, ATREP10A and VANDAL2 (for parameters “90-90”), as well as ATDNA2T9A, ATREP14, ATREP2 and ATREP7 families (for parameters “80-80”).

The remaining *de novo* consensus that matched these families were longer than the “knowledge-based” consensus. They contained insertions not present in the “knowledge-based” consensus.

Table S10. Results of coordinate comparisons for TE annotation

We compared the annotations between the various combinations of *de novo* libraries and reference databanks, in terms of match coordinates. Several cases can be distinguished on the basis of the distance between the 5' (or 3') coordinate of the test match and that of the reference match (figure S3): distance less than or equal to 1 nt, distance strictly greater than1 nt but less than or equal to 10 nt, and distance strictly greater than 10 nt.

Genome	Prediction	GROUPER	RECON	PILER	G+R+P
D. mel.	Exact	10975	6438	8084	8124
	Near exact	3553	2817	2794	3388
	Equivalent	890	825	789	972
	Near equivalent	5683	5839	5321	6774
	One-side exact	10169	7493	7429	8903
	Similar	18606	19026	15440	21502
	New TE	11077	8864	6874	13533
A. tha.	Exact	3423	4448	2346	3986
	Near exact	2590	2672	2171	2706
	Equivalent	917	916	927	1051
	Near equivalent	8207	8138	7020	7992
	One-side exact	8671	9626	6242	8716
	Similar	36651	38690	30977	36139
	New TE	8614	15482	6588	13023

Same data in percentages, knowing the number of TE fragments for *D. melanogaster*: G=43699; R=33072 ; P=32789; GRP=42857; and *A. thaliana*: G=41791; R=49470; P=34818; GRP=44059.

Genome	Case	Prediction	GROUPER	RECON	PILER	G+R+P
D. mel.	1-to-1	Total 1-to-1	57.34%	58.54%	64.07%	54.98%
		Exact	27.84%	26.07%	27.57%	25.60%
		Near exact	9.68%	9.23%	8.84%	9.32%
		One-side exact	21.72%	20.18%	19.56%	20.35%
		Equivalent	2.27%	2.37%	2.36%	2.53%
		Near equivalent	12.16%	13.49%	13.31%	13.47%
		Similar	26.34%	28.67%	28.36%	28.73%
	1-to-0	New TE	24.80%	22.45%	18.58%	27.65%
	1-to-n	Chimera	17.86%	19.01%	17.34%	17.37%
A. tha.	1-to-1	Total 1-to-1	53.96%	48.10%	54.87%	47.69%
		Exact	11.78%	14.27%	9.19%	14.67%
		Near exact	6.38%	6.40%	5.98%	7.16%
		One-side exact	19.16%	19.35%	16.53%	19.29%
		Equivalent	1.67%	1.75%	1.90%	2.18%
		Near equivalent	14.49%	13.54%	14.37%	14.53%
		Similar	46.52%	44.68%	52.03%	42.17%
	1-to-0	New TE	18.38%	27.46%	16.86%	25.95%
	1-to-n	Chimera	27.66%	24.44%	28.28%	26.35%

Table S11. Comparison of the performances of the RepeatModeler and TEdenovo databanks

Genome	Pipeline	Consensus	S_n*	S_p*	R_{cc}
D. mel.	RepeatModeler	141	77.78%	75.89%	19.11%
	TEdenovo	593	92.31%	75.72%	77.94%
A. tha.	RepeatModeler	177	45.90%	57.63%	5.19%
	TEdenovo	1351	74.43%	66.75%	49.35%

S_n*: percentage of “knowledge-based” consensus sequences matching a *de novo* consensus

S_p*: percentage of *de novo* consensus sequences matching a “knowledge-based” consensus

R_{CC}: percentage of fully recovered “knowledge-based” consensus sequences

Table S12. Comparison of the performances of TE annotation with the databanks of *de novo* consensus sequences from RepeatModeler and TEdenovo

Genome	TE library	Consensus	TE genome coverage	Number of copies	S_n	S_p
D. mel.	BDGP	125	10.51%	31208	NA	NA
	TEdenovo	568	11.98%	42847	91.43%	97.35%
	RepeatModeler	141	9.80%	28789	80.18%	98.47%
A. tha.	Repbase	318	19.02%	41146	NA	NA
	TEdenovo	1232	22.77%	44059	87.03%	92.32%
	RepeatModeler	175	15.14%	35432	65.14%	96.60%

Table S13. Reference sequences entirely retrieved by one clustering method but not by the others, in the *D. melanogaster* genome

Family	Classification	Length of the reference	Number of full-length fragments	Number of full-length copies	Program retrieving the entire reference	Results of the other programs
invader3	Class I LTR	5484	3	4	GROUPER	R:CI / P:NA
invader4	Class I LTR	3105	1	2	GROUPER	R:II / P:IC
Stalker	Class I LTR	7256	3	3	GROUPER	R:II / P:IC
Stalker4	Class I LTR	7359	2	3	GROUPER	R:II / P:IC
1360	Class II TIR	3409	2	2	RECON	G:IC / P:IC
Doc2-element	Class I LINE	4789	1	3	RECON	G:IC / P:II
G2	Class I LINE	3102	4	4	RECON	G:IC / P:II
gypsy4	Class I LTR	6852	0	3	RECON	G:IC / P:NA
Max-element	Class I LTR	8556	2	2	RECON	G:IC / P:NA
R1A1-element	Class I LINE	5356	2	2	RECON	G:IC / P:NA
rover	Class I LTR	7318	2	2	RECON	G:IC / P:II
Tabor	Class I LTR	7345	2	2	RECON	G:IC / P:II
Tc1-2	Class II TIR	1644	1	1	RECON	G:IC / P:NA

In the column on the far right, “G” stands for GROUPER, “R” for RECON and “P” for PILER. “CI” indicates that the reference sequence matches completely (over more than 95% of its length) a *de novo* consensus, which matches incompletely (over less than 95% of its length). “IC” means that the reference sequence matches incompletely whereas the *de novo* consensus matches completely. “II” indicates that both

sequences match incompletely. “NA” indicates that the reference sequence matches none of the *de novo* consensus sequences.

Table S14. Reference sequences entirely retrieved by one clustering method but not by the others, in the *A. thaliana* genome

Family	Classification	Length of the reference sequence (bp)	Number of full-length fragments	Number of full-length copies	Clustering method retrieving the entire reference sequence	Results of the other clustering methods
ATDNA1T9A	Class II TIR	3552	3	3	GROUPER	R:IC / P:II
ATENSPM5	Class II TIR	8717	0	4	GROUPER	R:IC / P:NA
ATGP2N	Class I LTR	5233	0	2	GROUPER	R:IC / P:II
ATGP6	Class I LTR	5695	0	1	GROUPER	R:IC / P:II
ATHILA	Class I LTR	10492	2	3	GROUPER	R:II / P:II
ATHILA2	Class I LTR	10923	4	5	GROUPER	R:II / P:IC
ATHILA6A	Class I LTR	11611	1	1	GROUPER	R:II / P:IC
VANDAL15	Class II TIR	5332	1	1	GROUPER	R:IC / P:NA
ARNOLD2	Class II TIR	15490	2	2	RECON	G:IC / P:NA
ATCOPIA1	Class I LTR	5139	2	2	RECON	G:NA / P:II
ATCOPIA10	Class I LTR	5186	3	3	RECON	G:NA / P:NA
ATCOPIA31	Class I LTR	4626	2	2	RECON	G:NA / P:NA

Family	Classification	Length of the reference sequence (bp)	Number of full-length fragments	Number of full-length copies	Clustering method retrieving the entire reference sequence	Results of the other clustering methods
ATCOPIA49	Class I LTR	5220	5	6	RECON	G:II / P:NA
ATDNAI27T9B	Class II TIR	2090	2	2	RECON	G:IC / P:II
ATENSPM6	Class II TIR	8825	2	9	RECON	G:IC / P:NA
ATGP2	Class I LTR	7623	4	6	RECON	G:IC / P:II
ATGP5	Class I LTR	6359	1	2	RECON	G:IC / P:NA
ATREP14	Class II Helitron	737	11	12	RECON	G:IC / P:II
ATREP16	Class II TIR	1391	2	2	RECON	G:NA / P:NA
BRODYAGA1	Class II TIR	1184	8	8	RECON	G:II / P:II
LIMPET1	Class II TIR	1874	2	2	RECON	G:IC / P:II
SIMPLEHAT1	Class II TIR	1059	4	4	RECON	G:IC / P:NA
VANDAL2	Class II TIR	15253	3	6	RECON	G:IC / P:IC

In the column on the far right, “G” stands for GROUPER, “R” for RECON and “P” for PILER. “CI” indicates that the reference sequence matches completely (over more than 95% of its length) a *de novo* consensus, which matches incompletely (over less than 95% of its length). “IC” means that the reference sequence matches incompletely whereas the *de novo* consensus matches completely. “II” indicates that both sequences match incompletely. “NA” indicates that the reference sequence matches none of the *de novo* consensus sequences.

Table S15. List of TE families represented by several *de novo* consensus sequences

Genome	Reference sequence(s)	Consensus from GROUPER	Consensus from RECON	Consensus from PILER
D. mel.	Stalker, Stalker2 and Stalker4	15	3	0
	FB	11	8	0
	invader1	5	1	1
	GATE	3	2	0
	297	3	1	1
	baggins	1	4	0
	invader4	2	2	1
	Idefix	1	2	1
	springer, gypsy3	0	3	0
	mdg1	2	0	1
	micropia	1	2	0
A. tha.	ATHILA6A, ATHILA6B	8	3	0
	SIMPLEHAT2	7	3	1
	ATREP1, ATREP2	5	2	2
	ATGP2N	4	1	0
	ARNOLDY1, ARNOLDY2	2	5	0
	ARNOLD1 to 4	2	5	0
	VANDAL2	3	3	0

	ATLANTYS2	1	4	0
	ATHILA4D	3	2	1
	VANDAL6	2	2	0
	ATREP2A	1	0	4
	HELITRONY2	1	2	0
	ATMUNX1	3	1	0
	ATREP4	2	2	0

These TE families were formed using BLASTCLUST, with a coverage of 80%. Manual curation may result in these families being represented by fewer consensus sequences than indicated.

Table S16. TE reference sequences for which no "knowledge-based" consensus could be built

Genome	Reference sequence	Length (bp)	Comments
D. mel.	Helitron	564	no copy
	BS4	754	less than 3 copies
	Q-element	759	not possible to build a consensus
	Stalker3T	372	no copy more than 100 bp long
	Penelope	804	"knowledge-based" consensus sequence could not identify a genomic copy
	P-element	2907	known to be absent from the sequenced strain
	TART-A	13424	consensus does not match its reference sequence, known to be present at telomeres
	TART-B	10654	consensus does not match its reference sequence, known to be present at telomeres
	TART-C	11124	consensus does not match its reference sequence, known to be present at telomeres
A. tha.	ATCOPIA6	4718	consensus does not match its reference sequence with the parameters used
	ATCOPIA18	2384	only two copies (1 truncated and 1 full-length)
	ATCOPIA30	4237	consensus does not match its reference sequence with the parameters used
	ATCOPIA31A	4664	only two copies (1 truncated and 1 full-length)
	ATCOPIA47	5063	three copies (2 truncated, including 1 less than 100 bp long, 1 full-length)
	ATCOPIA80	4280	only two copies (1 truncated and 1 full-length)
	ATCOPIA84	4882	only two copies (1 truncated and 1 full-length)
	ATCOPIA85	3878	consensus does not match its reference

Genome	Reference sequence	Length (bp)	Comments
			sequence with the parameters used
	ATCOPIA91	5270	five copies (3 less than 100 bp long and 1 full-length)
	ATREP19		consensus does not match its reference sequence with the parameters used
	DRL1	250	only one copy
	TA1_AT	514	only one copy
	TA12	949	consensus does not match its reference sequence with the parameters used

Table S17. Comparison of "knowledge-based" libraries with reference databanks

Genome	Sequences in the reference databank	Sequences in the "knowledge-based" library	S_n*	S_p*	R_{CC}
D. mel.	126	117	94.44%	100.00%	100.00%
A. tha.	318	309	98.43%	100.00%	100.00%

S_n*: percentage of "knowledge-based" consensus sequences matching a *de novo* consensus sequence

S_p*: percentage of *de novo* consensus sequences matching a "knowledge-based" consensus sequence

R_{CC}: percentage of fully recovered "knowledge-based" consensus sequences

Troisième partie

DISCUSSION ET PERSPECTIVES

3

DISCUSSION ET PERSPECTIVES

3.1 L'ANNOTATION DES ÉLÉMENTS TRANSPOSABLES À L'ÈRE DU SÉQUENÇAGE HAUT-DÉBIT DES GÉNOMES EUCARYOTES

3.1.1 *Application des outils informatiques développés à divers génomes*

Les années 1990-2000 ont été caractérisées par le séquençage des premiers génomes d'eucaryotes, principalement des organismes modèles tels *Saccaromyces cerevisiae* (Goffeau et coll. 1996), *Drosophila melanogaster* (Adams et coll. 2000), *Arabidopsis thaliana* (the Arabidopsis Genome Initiative 2000), mais aussi celui de l'homme (the International Human Genome Sequencing Consortium 2001). En avril 2010, la base de données GOLD (`http://www.genomesonline.org/`) recensait 129 projets de séquençage d'eucaryotes complétés et publiés. Ce chiffre correspond à une estimation basse des séquences disponibles en pratique, étant donné les recommandations en vigueur à propos de la mise à disposition des séquences génomiques avant leur publication (the Welcome Trust 2003). En effet, la base de données GOLD recensait également 1339 projets de séquençage en cours. Quelque soit le but de l'analyse d'une séquence génomique, l'annotation des ET en représente une étape essentielle. L'annotation des gènes codant pour les protéines est grandement améliorée si les ET ont été repérés au préalable, de telle sorte qu'un ET ne puisse être considéré par erreur comme étant un gène de l'hôte. Cependant, cette appellation *gène de l'hôte* est trompeuse et prête à confusion. Les ET sont en effet dignes d'être annotés pour eux-mêmes, les preuves expérimentales de leur implication dans de nombreux processus biologiques étant nombreuses (chapitre 1).

Les génomes récemment séquencés n'appartenant plus seulement à des organismes modèles pour lesquels les ET sont connus, il est nécessaire d'utiliser des approches *de novo* afin d'identifier les familles d'ET présentes en reconstruisant les séquences ancestrales (consensus) pour en détecter leurs copies. J'ai été impliqué dans de nombreux projets de ce type au cours

de mes trois années de doctorat. Le premier projet d'annotation auquel j'ai participé a concerné un génome de l'espèce *Meloidogyne incognita* (50 Mb) et a fait l'objet d'une publication (Abad et coll. 2008, en annexe de ce manuscrit A.1). Ce nématode parasite les racines de la plupart des plantes cultivées et est donc responsable de baisses de rendement occasionnant des pertes colossales estimées à 157 milliards de dollars annuellement. L'annotation de ce génome à l'aide des outils TEdenovo et TEannot m'a permis d'estimer à 36% le contenu en ET et de classifier les familles présentes dans ce génome. J'ai également décrit la distribution du nombre de copies par catégorie d'ET ainsi que leur âge respectif, celui-ci étant approximé par le pourcentage d'identité entre une copie génomique et la séquence consensus qui l'a identifiée (figure 4). Ces résultats ont ainsi permis d'avoir une première idée de l'importance et de la dynamique des ET dans ce génome.

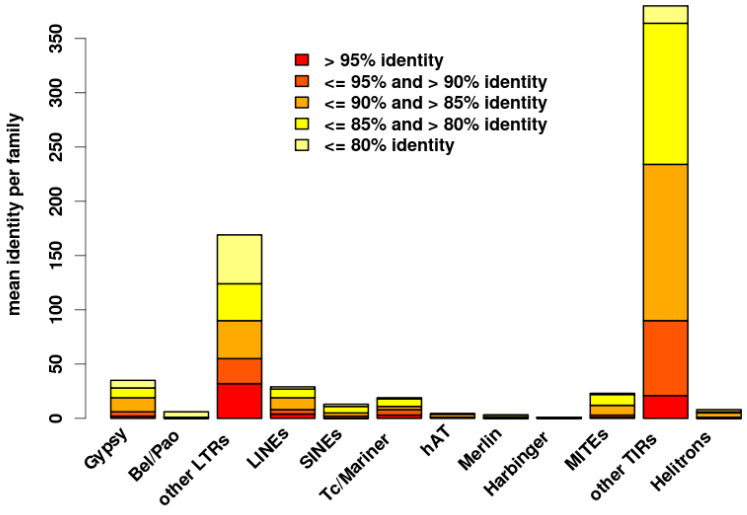


FIG. 4: Distribution de l'âge des ET dans le génome de *M. incognita*. L'âge est approximé par l'identité de séquence entre chaque copie et la séquence consensus qui l'a détectée. Les ET sont regroupés par super-famille. (Tiré de Abad et coll. 2008.)

Ce type d'analyses ne nécessitant que la séquence brute d'un génome et pouvant être réalisé entièrement automatiquement, les outils que je développais ont très vite été utilisés au sein du laboratoire. Je me suis donc attaché à les rendre rapidement opérationnels sur le plan technique, tout en améliorant la pertinence

biologique des résultats qu'ils génèrent. J'ai ainsi participé au projet ANR *HOLOCENTRISM* (07 BLAN 0057) avec Emmanuelle Permal en post-doctorat à l'URGI dans le cadre d'une collaboration avec des chercheurs de l'INRA de Montpellier et Rennes. Le but de ce projet était d'analyser la structure holocentrique des chromosomes de certains insectes et de la relier à la plasticité de ces génomes, à leur évolution et au contrôle épigénétique de l'expression génétique.

Lors de la division cellulaire de chromosomes monocentriques, les micro-tubules se fixent en une région particulière, le centromère. A l'inverse, les chromosomes holocentriques sont attachés aux microtubules sur toute leur longueur et ont donc de multiples points d'attache (figure 5). Dans les génomes monocentriques, le taux de crossing-overs diminue généralement en se rapprochant du centromère, jusqu'à devenir nul (Coop et Przeworski 2006), alors que les génomes holocentriques comme celui de *Caenorhabditis elegans* ne présentent pas les mêmes caractéristiques puisqu'ils n'ont pas de centromères localisés (Barnes et coll. 1995). Or la recombinaison méiotique a un très grand impact sur l'évolution des génomes. En effet, les modèles de génétique des populations montrent que diminuer le taux de crossing-overs revient à abaisser le taux de fixation des allèles avantageux, c'est-à-dire l'impact de la sélection naturelle : c'est l'effet de Hill-Robertson (Hill et Robertson 1966, Roze et Barton 2006). Comme la majorité des insertions d'ET sont supposées être délétères, on pourrait s'attendre à une accumulation d'ET dans les zones de faible recombinaison méiotique, et donc à une corrélation négative entre ET et taux de crossing-overs. Mais la relation entre distribution des ET et taux de recombinaison méiotique n'est pas encore clairement comprise (Rizzon et coll. 2002, Wright et coll. 2003, Drouaud et coll. 2006).

Nous nous sommes donc intéressés à la distribution des ET dans des génomes holocentriques en analysant, à l'aide des outils TEdenovo et TEannot, plusieurs séquences génomiques synténiques entre trois espèces de Lépidoptères (15 BAC) : le ver à soie *Bombyx mori*, *Helicoverpa armigera* et *Spodoptera frugiperda*. Ce travail a fait l'objet d'une publication (d'Alençon et coll. 2010, en annexe de ce manuscrit A.2). L'étude a montré que le taux de ruptures de synténie dans ces espèces extrapolé à partir des données est quatre fois plus élevé que chez les Drosophilidae chez lesquelles le taux est lui-même deux fois plus élevé que chez les mammifères. Nous avons également montré une corré-

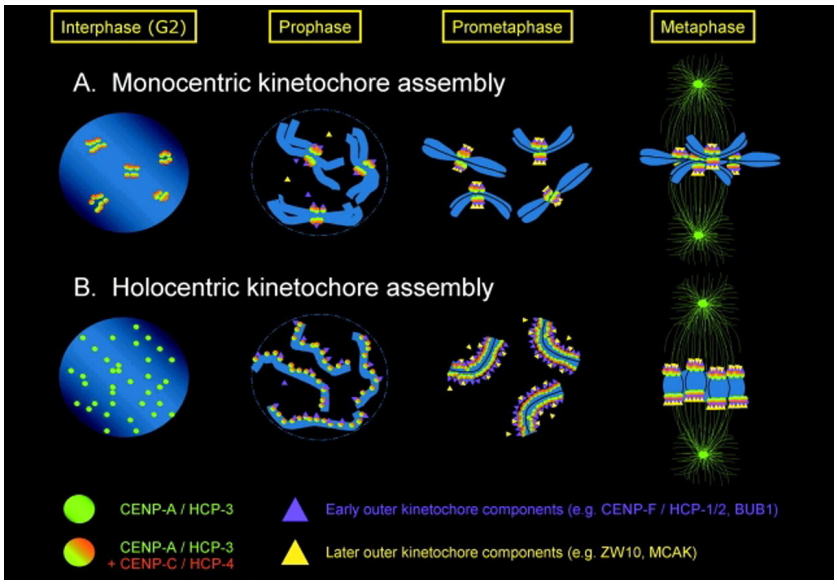


FIG. 5: Différence dans l'assemblage des kinétochores sur des chromosomes monocentriques et holocentriques. (Tiré de Dernburg 2001.)

lation entre rupture de synténie et densité en ET (figure 6). En parallèle de ces travaux, Emmanuelle Permal a réalisé, toujours à l'aide des outils TEdenovo et TEannot, l'annotation des ET dans les génomes holocentriques du ver à soie, *B. mori*, et du puceron du pois, *Acyrthosiphon pisum* (the International Aphid Genomics Consortium 2010). L'annotation des ET dans ces génomes a indirectement représenté une phase importante de mon travail de mise au point technique car ces génomes étaient les plus grands analysés jusqu'alors par les outils du package REPET : 446 Mb for *A. pisum* et 514 Mb for *B. mori*, soit plus de quatre fois le génome d'*A. thaliana*.

L'un des grands changements de ces dernières années provient du fait que l'on dispose de la séquence quasi-complète de nombreux génomes apparentés. Ce changement d'ordre de grandeur en terme de données disponibles permet de rechercher des signaux évolutifs en comparant plusieurs génomes entre eux. Ces approches regroupées sous le vocable *génomique comparée* tirent profit des relations phylogénétiques à différentes échelles de temps entre les espèces dont un génome est séquencé (Flowers et Purugganan 2008). Grâce mon travail, il est devenu envisageable d'étudier la dynamique de l'ensemble des ET dans plusieurs génomes en utilisant le même outil informatique, sans

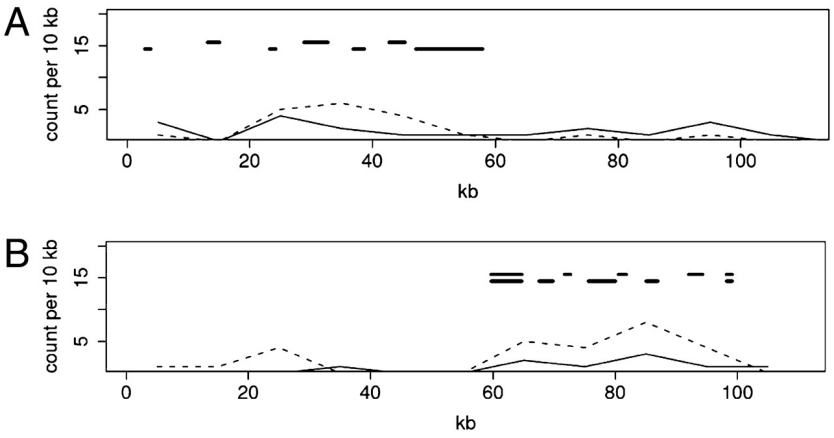


FIG. 6: Corrélation entre les points de rupture de synténie et la densité en ET chez des Lépidoptères. **A**. region CYP4M de *S. frugiperda*. **B**. région EcR de *H. armigera*. Les lignes pointillées correspondent à la densité en ET, les lignes pleines fines à la densité en gènes, et les lignes pleines épaisses aux points de rupture de synténie. (Tiré de d'Alençon et coll. 2010.)

être obligé de se restreindre à quelques familles ou quelques types d'ET (les rétrotransposons à LTR par exemple). Dans le cas où des méthodes différentes sont utilisées pour analyser deux génomes et que le contenu en éléments transposables de ces génomes est différent, il n'est pas toujours évident ni possible de distinguer si ces différences sont dues à un phénomène biologique ou à un biais présent seulement dans l'une des deux méthodes. Ainsi, le fait de pouvoir utiliser un même outil, même si lui aussi présente des biais, permet au moins d'espérer que les biais en question soient présents systématiquement dans chaque génome analysé. Les différences observées entre deux génomes devraient donc plus vraisemblablement refléter des mécanismes biologiques distincts ou des différences dans la dynamique des ET.

Plusieurs projets de ce type ont été initiés pendant mon doctorat. J'ai notamment démarré un projet concernant plusieurs génomes de plantes (tableaux 1 et 2). Etudier le contenu en ET de deux génomes du même genre, ici *Arabidopsis thaliana* et *Arabidopsis lyrata*, permet d'étudier la dynamique des ET chez deux génomes proches mais dont les organismes ont des caractéristiques biologiques contrastées. En effet, ces deux espèces ont divergé il y a environ 5 ou 10 millions d'années. Elles sont de plus suffisamment proches pour que l'on puisse inférer l'état ancestral d'un polymorphisme entre différentes accessions d'*A. tha-*

liana simplement en regardant l'état correspondant chez *A. lyrata*. De façon intéressante, cette dernière a un génome deux fois plus grand qu'*A. thaliana* (205 Mb contre 120 Mb) et un mode de reproduction différent, majoritairement par allo-fécondation. Enfin, *A. lyrata* est une espèce endémique caractérisée par des populations bien délimitées, au contraire d'*A. thaliana* reconnue pour sa capacité de dispersion. Ces caractéristiques ont donc très vite suscité notre intérêt pour comprendre les relations entre traits de vie et dynamique des ET. Ces études, aujourd'hui possibles grâce aux outils de REPET, sont en cours au laboratoire.

TAB. 1: Résultats comparatifs de la première étape de l'outil TEdenovo appliqué à plusieurs génomes de plantes.

Organisme	Taille du génome (en Mb)	Nombre d'alignements	Couverture du génome
A. thaliana	119	103728	13.48%
A. lyrata	207	1378876	32.65%
B. distachyon	270	2096350	30.16%
O. sativa	372	6027252	33.54%
V. vinifera	487	12002121	39.77%

TAB. 2: Résultats comparatifs des étapes 2 à 5 de l'outil TEdenovo appliqué à plusieurs génomes de plantes.

Organisme	Groupes formés par GROUPER	Groupes formés par RECON	Groupes formés par PILER	Banque finale de consensus
A. thaliana	1428	1021	300	1275
A. lyrata	7206	3737	1117	5214
B. distachyon	12772	3848	866	6307
O. sativa	19721	5218	1566	9844
V. vinifera	18241[1]	8397	1631	13444

[1] Seuls les groupes ayant au moins 4 membres ont été gardés.

J'ai également initié des analyses préliminaires sur le génome du riz, *Oryza sativa*, et celui de *Brachypodium distachyon*. *B. distachyon* est aujourd'hui le génome séquencé le plus proche de

celui du blé hexaploïde *Triticum aestivum* bien qu'il soit beaucoup plus petit (270 Mb contre 17 Gb) et ai divergé il y a 30 millions d'années. Du fait de sa taille gigantesque, le génome de *T. aestivum* n'est pas encore séquencé mais l'établissement d'une carte physique pour son chromosome le plus grand, le 3B (Paux et coll. 2008), indique que l'assemblage d'un génome aussi répété est concrètement possible. La séquence de ce chromosome est d'ailleurs l'objectif d'un projet ANR (3BSEQ, http://urgi.versailles.inra.fr/index.php/urgi/Projects/3BSeq) porté par le laboratoire GDEC à l'INRA de Clermont-Ferrand avec lequel j'ai effectué ma thèse en cotutelle. A ce titre, j'ai fréquemment collaboré avec Philippe Leroy et Frédéric Choulet de l'INRA de Clermont afin d'incorporer l'outil TEannot dans l'outil d'annotation TriAnnot (https://gpi.versailles.inra.fr/triannot) qui permet aux chercheurs du monde entier d'annoter les séquences de blé qu'ils produisent dans leurs laboratoires respectifs. Des travaux sont encore en cours pour déterminer les paramètres optimaux pour l'annotation des ET du blé mais l'outil TEannot a déjà été intégré avec succès à l'interface web grâce au travail de Philippe Leroy (manuscrit en préparation). Dans le cadre de ma thèse, j'ai également participé à l'annotation des ET dans le génome de la vigne *Vitis vinifera* (the French–Italian Public Consortium for Grapevine Genome Characterization 2007), un travail en cours mené par Nathalie Choisne de l'URGI.

Bien qu'à cette étape d'acquisition de données il soit trop tôt pour tirer des conclusions en terme de génomique comparée, l'ensemble de ces travaux préliminaires d'annotations de génomes de plantes démontrent la faisabilité d'une telle approche grâce aux outils TEdenovo et TEannot. Dans les années à venir, les autres génomes de plantes actuellement disponibles tels le maïs (Schnable et coll. 2009), le sorgho (Paterson et coll. 2009), le concombre (Huang et coll. 2009), la papaye (Ming et coll. 2008), etc, pourront être ainsi analysés. Dans tous les cas, les ET de ces génomes ont déjà été annotés, notamment pour le maïs (Baucom et coll. 2009), mais la diversité des outils utilisés ne permet pas toujours de comparer objectivement les annotations. Enfin, à la différence des autres approches *de novo*, l'outil TEdenovo tente d'identifier les variants structuraux d'ET. L'étude de ces variants au sein des génomes de plantes sus-cités est prometteur si l'on en croit les résultats préliminaires montrant des variations substantielles du nombre de groupes pour une couverture du génome équivalente (tableaux 1 et 2). En effet, comme mon-

tré précédemment (chapitre 2), l'outil TEdenovo construit plus de groupes que le nombre de familles d'ET identifiées préalablement chez *D. melanogaster* et *A. thaliana*. Ceci est dû au fait que l'outil TEdenovo cherche à détecter les variants structuraux des familles. Ainsi, lorsque l'analyse d'un génome produit un plus grand nombre de groupes que celle d'un autre génome proche, alors qu'une proportion similaire de ces deux génomes est constituée d'ET, il est vraisemblable que la diversification des familles dans le génome ayant beaucoup de groupes ait été plus importante que dans l'autre génome.

Dans le même esprit et du fait de mes recherches sur la reconstruction des séquences ancestrales d'ET, j'ai été en relation directe avec Joëlle Amselem travaillant à l'URGI sur de nombreux génomes de champignons. Responsable de l'annotation de ces génomes et de la mise à disposition des informations s'y rapportant, elle a pu, grâce à REPET, entreprendre une analyse de génomique comparée visant à estimer l'impact des mécanismes visant à empêcher l'expansion des ET. Elle s'est en particulier intéressée à certains mécanismes spécifiques des champignons comme le RIP (*repeat-induced point mutations*, Selker 1990). L'utilisation des outils TEdenovo et TEannot a donc été déterminant dans l'obtention des annotations fiables et comparables d'ET (tableaux 3 et 4) nécessaires à ces analyses.

D'autres projets ont également bénéficié de mon travail tel l'annotation des ET dans le génome d'Ectocarpus siliculosus (Cock et coll. 2010) et l'analyse comparative des dynamiques d'ET dans 12 génomes de Drosophilidae en collaboration avec Anna-Sophie Fiston-Lavier du laboratoire de Dmitri Petrov à Stanford University (États-Unis).

Ces projets ont pu se concrétiser grâce aux nombreuses possibilités ouvertes par l'arrivée d'une approche d'annotation *de novo* performante et fiable. En effet, à l'ère du séquençage haut-débit et de la génomique comparée, la pertinence des outils informatiques ne se mesure plus uniquement à l'aune des résultats qu'ils renvoient. Il devient indispensable de disposer d'outils robustes, capables d'analyser de grandes quantités de données variant en taille et complexité. Cet aspect du travail d'un bio-informaticien est souvent considéré comme relevant d'une activité d'ingénierie de développement et non d'une activité de recherche. En conséquence, les aspects *architecture logicielle* et robustesse du code ont peut-être tendance à être trop souvent né-

TAB. 3: Résultats comparatifs de la première étape de l'outil TEdenovo appliqué à plusieurs génomes de champignons.

Organisme	Taille du génome (en Mb)	Nombre d'alignements	Couverture du génome
Botrytis cinerea T4	40	813	0.68%
Botrytis cinerea 05.10	43	2116	1.72%
Sclerotinia sclerotiorum	38	13197	7.88%
Leptosphaeria maculans	45	695940	36.24%
Blumeria graminis	91	453530	63.71%
Melampsora larici-populinia	101	191650	44.98%
Tuber melanosporum	125	3735967	60.10%

gligés, ce qui revient à alimenter le champ disciplinaire d'outils informatiques difficilement maintenables et utilisables hors du contexte dans lequel ils ont été testés.

La pertinence biologique ainsi que la fiabilité et la robustesse des outils TEdenovo et TEannot a permis l'analyse du contenu en ET de nombreux génomes. Ces deux derniers aspects, fiabilité et robustesse, sont cependant trop souvent laissés de côté lors d'une démarche de recherche en bioinformatique. Or, il existe des solutions efficaces, suffisamment flexibles pour convenir à tout chercheur.

3.1.2 *Importance du développement logiciel dans la recherche en bioinformatique*

Depuis quelques années, la recherche en biologie se caractérise par un développement soutenu d'approches expérimentales générant un grand volume de données, notamment les techniques de séquençage dites de nouvelle génération (NGS, Metz-

TAB. 4: Résultats comparatifs des étapes 2 à 5 de l'outil TEdenovo appliqué à plusieurs génomes de champignons.

Organisme	Groupes formés par GROUPER	Groupes formés par RECON	Groupes formés par PILER	Banque finale de consensus
Botrytis cinerea T4	72	19	10	31
Botrytis cinerea 05.10	305	30	8	42
Sclerotinia sclerotiorum	606	102	62	337
Leptosphaeria maculans	3890	86	25	1850
Blumeria graminis	8503	1071	265	1920
Melampsora larici-populinia	2916	1764	461	2029
Tuber melanosporum	12776	791	306	2597

ker 2009). Cela concerne le séquençage de génomes complets bien sûr comme décrit dans le paragraphe précédent, mais pas uniquement. Le re-séquençage de génomes est en plein essor. Il consiste à séquencer à faible couverture le génome de plusieurs individus de la même espèce en s'appuyant sur une séquence de référence préalablement établie. Ces approches sont devenues de plus en plus courantes comme le montrent les projets phares *1000 génomes humains* et *1001 génomes d'A. thaliana*. De plus, le séquençage NGS permet des analyses fonctionnelles à haut débit au travers du séquençage en masse des transcrits ARN d'une population de cellules (*RNA-seq*, Wang et coll. 2009) et des fragments d'ADN liés à des protéines récupérées par immunoprécipitation (*ChIP-seq*, Park 2009). Enfin, ces techniques permettent également de réaliser du génotypage à très haut débit. Les programmes de sélection animaux et végétaux s'orientent de plus en plus vers les marqueurs basés sur les séquences et la détection des milliers voir millions de polymorphismes de type SNP en une seule expérimentation. Toutes ces approches, qui

deviennent courantes dans de nombreux laboratoires, génèrent d'importants volumes de données (de l'ordre de plusieurs dizaines de millions à plusieurs milliards de courtes séquences). Enfin, des approches expérimentales haut-débit concernant cette fois la détection de protéines et l'imagerie cellulaire sont aujourd'hui en plein développement.

Toutes ces techniques nécessitent le développement d'outils informatiques permettant d'analyser leurs résultats. En effet, les quantités de données produites représentent un défit pour leur analyse par l'expérimentateur. Celui-ci, souvent habitué à travailler sur un nombre restreint de locus, est alors amené à analyser l'intégralité d'un génome *via* des données qu'il ne sait pas toujours manipuler sur son ordinateur personnel. Ces deux aspects-clés, la génération de données expérimentales à haut-débit et leur analyse automatisée, représente non seulement un changement d'échelle mais induit surtout un changement de méthodes d'analyse. L'expérimentateur a besoin aujourd'hui de résultats standardisés obtenus grâce à des méthodes robustes et rapides, que seul des analyses automatiques peuvent réaliser. L'analyse de données haut-débit suit aujourd'hui une approche industrielle.

Néanmoins, l'industrialisation des analyses ne correspond pas toujours à la réalité actuelle. En effet, le développement des techniques expérimentales haut-débit et la mise au point d'outils informatiques robustes ne se font pas toujours aux mêmes endroits. De plus, ces activités suivent également un processus de recherche, par tâtonnement, *via* le développement de prototypes dont le but premier n'est en général pas d'être robuste. Le développement de ces logiciels fait appel tout autant à l'ingénierie qu'à la science, mais comme l'indique Sean Eddy (HHMI, Janelia Farm Research Campus) :

> *our culture values science, not engineering ; the result : a software literature full of good ideas that don't get fully baked, tools that work in one place but aren't portable.*

A ma connaissance, ces questions ne sont généralement pas abordées dans une thèse de bio-informatique. Cela me paraît regrettable, non seulement parce que tout chercheur y est confronté en pratique, mais aussi parce que des solutions existent et qu'elles méritent d'être diffusées, étant donné leur efficacité. J'ai pu effectuer ma thèse dans un fort contexte d'ingénierie, l'URGI, un laboratoire de recherche hébergeant une plateforme bio-informatique. La plateforme a pour mission de mettre à disposition de la communauté scientifique données, outils et

expertise. L'un des défis de ces dernières années a été de faire communiquer les aspects recherche et ingénierie au laboratoire. A ce titre, j'ai été fortement impliqué dans ce chantier, de concert avec Olivier Inizan qui m'a initié aux méthodes de développement dites *Agiles* alors que je lui transmettais mon expertise sur les ET. De là est née l'équipe de développement *pipelines* à l'URGI utilisant comme base de travail le code informatique sur lequel j'avais travaillé. Je ne vais pas détailler ici l'ensemble des méthodes *Agiles*, mais montrer dans la suite quel bénéfice un chercheur en bio-informatique peut en tirer dans son travail quotidien.

Dans les années 1990, plusieurs responsables de projet de développement logiciel à travers le monde ont remis en question le long cycle de développement traditionnel constitué d'une phase de conception, suivi d'une phase de réalisation et clôt par une phase de livraison (figure 7). Un tel cycle présente certaines lourdeurs pouvant occasionner des retards, voir même une inadéquation du produit final au regard des besoins actuels qui ont évolué depuis la phase initiale de conception. De nouvelles pratiques ont donc été proposées, beaucoup plus réactives grâce à des cycles de développement courts, et mettant l'accent sur les relations humaines au sein de l'équipe de développement comme entre les développeurs et les clients. En 2001 a été écrit le manifeste Agile (`http://agilemanifesto.org/`) visant à définir les valeurs et principes de ces nouvelles méthodes. Il est à noter que celles-ci peuvent s'appliquer à toute gestion de projet, et pas seulement à la conception de logiciels. Cette flexibilité permet de s'inspirer de ces méthodes, notamment dans le cadre d'un travail de recherche en bio-informatique.

Le cas du chercheur est particulier en ce que son travail est collaboratif par nature, et doit (devrait ?) donc favoriser le transfert d'expertise. De plus, il est à la fois développeur d'un logiciel ainsi que premier utilisateur pour ses propres recherches. Enfin, il conçoit toujours des prototypes et non des produits finis au sens où on l'entend généralement. Au vu de ces caractéristiques, il est compréhensible que la méthode traditionnelle de développement logiciel ne soit que rarement utilisée en bio-informatique impliquant qu'aucune méthode n'est en faite vraiment utilisée. Or le besoin existe car même la mise au point d'un prototype nécessite de la modularité et tout prototype doit faire preuve de robustesse au regard de la quantité de données à manipuler. La méthode dite *extreme programming* (XP) fait des recommanda-

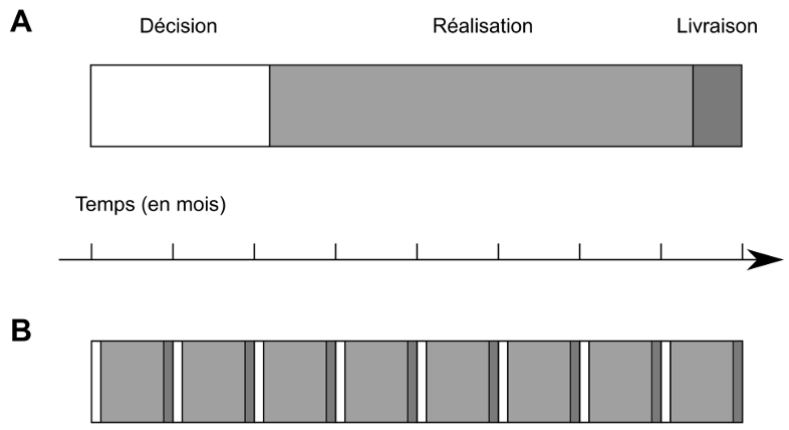


FIG. 7: Deux façons de concevoir le cycle de développement d'un logiciel. **A**. Conception traditionnelle. **B**. Conception prônée par l'*extreme programming*.

tions précises sur le plan technique (Beck 1999). J'ai donc commencé en cours de thèse à appliquer l'un des piliers de cette méthode : le *test-driven development*. Cela consiste d'abord à coder un test, dit *unitaire*, que la fonctionnalité désirée devra passer, puis à coder la fonctionnalité en question, et enfin à la modifier jusqu'à ce qu'elle passe le test écrit au préalable.

A première vue, une telle pratique n'est pas efficace puisqu'elle oblige le chercheur à écrire beaucoup plus de code que prévu, ce qui semble être une perte de temps. Or elle est au contraire primordiale lorsqu'on se préoccupe d'avoir une action efficace et durable, et ce à plus d'un titre. Dans tout projet de recherche en bio-informatique, la quantité de code généré augmente rapidement, il est donc fréquent de voir apparaître des erreurs non-triviales. Disposer de tests déjà écrits permet de détecter ces erreurs facilement et très tôt, ce qui occasionne des gains de temps très appréciables sur le long terme. De plus, le code informatique, quel qu'il soit, doit pouvoir rester *vivant*, c'est-à-dire utilisé mais surtout repris et amélioré par d'autres. Or, notre expérience montre que rien n'est plus clair qu'un test pour comprendre le but d'une fonctionnalité donnée lorsqu'on n'est pas soi-même le développeur de celle-ci. Enfin, une activité de recherche n'est prédictible que jusqu'à un certain point. Il est fréquent d'ajuster, voir de réorienter, ses questions scientifiques, et ce sur un pas de temps court. Dans ce cadre, il est indispensable de disposer d'un code informatique modulaire en lequel on a confiance afin de le faire évoluer dans d'autres directions

que celles prises initialement. Cette réflexion sur la conduite de projets vers un but qui peut évoluer au cours du temps est à rapprocher de travaux fondamentaux récents sur l'optimisation (Kashtan et coll. 2007). Ces travaux, inspirés par des considérations biologiques, montrent que la vitesse à laquelle un objectif est atteint peut être considérablement augmentée dans le cas où l'objectif varie de façon modulaire au cours du temps.

Pour une efficacité maximale, les tests unitaires écrits en amont du développement de chaque fonctionnalité sont exécutés quotidiennement, permettant ainsi une intégration continue du code généré. Cependant, ce type de tests correspond à des fonctionnalités de faible étendue. Il est donc pertinent et nécessaire de disposer, en plus, de tests plus larges vérifiant la cohérence biologique des résultats d'une analyse bioinformatique lancée dans son intégralité et non sa seule exécution technique. Dans ce cas, on parle de tests fonctionnels. Ces tests consistent à comparer l'analyse d'un génome faite avec la nouvelle version de l'outil, avec les résultats de l'analyse faite avec la version précédente, ces résultats ayant été validés manuellement. Au cours de ma thèse, j'ai donc mis au point un tel système d'intégration continue sur les outils TEdenovo et TEannot. Ces outils sont exécutés quotidiennement sur le chromosome 4 de *D. melanogaster*, et sur le génome entier de cet organisme avant une distribution publique des outils. Grâce à ce système, l'industrialisation mentionnée précédemment devient possible et prend même tout son sens.

L'application de quelques recommandations des méthodes Agiles a rendu robuste les outils, permettant ainsi leur utilisation sur de nombreux génomes. Grâce à cela, nous avons pu acquérir une expertise suffisante sur la question de l'annotation des ET dans les génomes eucaryotes *via* des approches *de novo* afin d'écrire une revue sur ces questions.

3.1.3 *Feuille de route pour l'annotation des éléments transposables dans les génomes eucaryotes*

Ce travail a été soumis récemment. J'en présente ci-dessous un résumé en français, suivi du manuscrit en anglais.

Traditionnellement présentés sous l'appellation *junk DNA*, les ET sont maintenant reconnus pour leur rôle dans l'évolution des génomes, à travers leur impact sur les réarrangements chromosomiques et l'expression des gènes. De plus, avec les méthodes

actuelles à haut-débit, le séquençage de génomes d'organismes non-modèles est techniquement faisable, et c'est le processus d'annotation qui représente actuellement un goulot d'étranglement pour l'analyse de ces génomes, particulièrement en ce qui concerne les ET. Le but de cette revue est donc de proposer une feuille de route aux chercheurs impliqués dans de tels projets de génomique abordant ces questions. A chaque étape du processus d'annotation des ET, de la définition des familles à l'annotation des copies, nous listons plusieurs outils fréquemment utilisés, tout en mettant en évidence leurs complémentarités. Nous donnons également un ensemble de bonnes pratiques requises pour la curation manuelle lors de l'interprétation des séquences consensus d'ET et de leurs copies annotées.

Roadmap for annotating transposable elements in eukaryote genomes

AUTHORS

Emmanuelle Permal, Timothée Flutre, Hadi Quesneville

Unité de Recherches en Génomique Info- URGI (UR1164) – INRA – Centre de Versailles 78026 cedex France

ABSTRACT

Once caricatured as mere “junk DNA”, transposable elements (TEs) are now known to have been important in genome evolution, particularly through effects on rearrangements and gene expression. Current high-throughput techniques make sequencing genomes of even non-model organisms feasible, such that the annotation process represents now a bottleneck to genome analysis, especially when dealing with transposable elements. The aim of this review is to provide a roadmap for researchers involved in genome projects addressing this issue. At each step of the TE annotation process, from the identification of TE families to the annotation of TE copies, we list several widely-used tools and consider their complementarities. We also explain the good practice required for manual curation when interpreting TE consensus and copy annotations.

keywords: *Transposable elements, genome annotation, sequence analysis*

INTRODUCTION

Transposable elements (TEs) are mobile genetic elements that shape the eukaryotic genomes in which they are present. They are virtually ubiquitous and make up, for instance, 20% of a typical *D. melanogaster* genome [1], 50% of a *H. sapiens* genome [2], and 85% of a *Z. mays* genome [3]. They are classified into two classes depending on their transposition mode: via RNA for class I retro-transposons; and via DNA for class II transposons [4]. Each class is also subdivided into several orders, super-families and families, this hierarchy being detailed elsewhere [5]. Due to their unique ability to transpose, TEs are main determinants of genome size [6, 7] and cause of genome rearrangements [8, 9]. Once described as the “ultimate parasites” [10], TEs are commonly found regulating the expression of neighboring genes [11, 12], or even domesticated to provide a specific function beneficial to the host [13-16].

Thanks to the development of new sequencing techniques, the number of sequenced eukaryotic genomes available, from different species or from different populations, is constantly increasing. However, the first step of the analysis, *i. e.* accurate annotation, remains a major challenge, and particularly as concerns TE copies. The correct genome annotation of genes but also TEs is an indispensable part of thorough genome-wide studies. Consequently, efficient computational methods have been proposed for TE annotation [17]. But, as the pace at which genomes are sequenced is unlikely to decrease in the coming years, the process of TE annotation has to be democratized.

The first aim of this review is to establish a clear road-map detailing the order in which computational tools (or combinations of such tools) should be used. The second aim is to provide some hints on manual curation, a still-necessary procedure once computational predictions have been obtained.

DE NOVO DETECTION OF TRANSPOSABLE ELEMENTS

Currently, efficient computational methods are required to identify TEs. Each method is based on specific assumptions that have to be understood to optimize selection and combination of the methods appropriate for any particular analysis goal.

Computing highly-repeated words

Software, such as the recent TALLYMER [18] and P-CLOUDS [19], was designed to find repeats quickly in genome sequences by counting highly frequent words of a given length *k*, called *k*-mers. These programs are very useful for providing a rapid view of the repeated fraction in a given set of genomic sequences, especially with unassembled sequences. However, they do not provide much detail about the transposable elements present in these sequences. Their output only corresponds to highly repeated regions without indicating precise TE fragment boundaries or TE family assignment. These methods are quick and simple to use but allow only limited biological interpretations.

Other methods also start by counting frequent *k*-mers but then go on to try to define consensus. ReAS [20] applies this approach directly on shotgun reads. For each frequent *k*-mer, a multiple alignment of all short reads containing it is built and then extended iteratively. REPEATSCOUT [21] has a similar approach but works on assembled sequences. These tools return a library of consensus sequences. However, although their results are more biologically relevant than those of previous methods, the consensuses are usually too short and correspond to truncated versions of ancestral TEs. Substantial manual inspection and editing is therefore needed to obtain a meaningful list of consensus sequences.

All-by-all alignment and clustering of interspersed repeats

Repeats can also be identified by self-alignment of the genomic sequences. In practice, once an assembled genome is available, the analysis starts by an all-by-all alignment of the input sequences.

Several tools can be used for this. Some are heuristic and use BLAST or BLAST-like algorithms. Tools, such as BLASTER [22], perform this search by launching BLAST [23] repeatedly over the genome sequences. Others, such as PALS, are exact algorithms called “filters” [24]. As the amount of input data is usually high, the computations are intensive. Therefore stringent parameters are applied: good results are obtained with BLAST-like tools when matches shorter than 100 bp or with identity below 90% or with an E-value above 1e-300 are dismissed. To speed up the computations, such alignment tools can be easily launched in parallel on a computer cluster.

With these parameters, only closely related TE copies will be used to reconstruct the ancestral sequence of the TE family. Note that the aim of this step is not to recover all TE copies, but to use those that are well conserved to build a robust consensus. Our experience indicates that this is a crucial assumption for successful reconstruction of a valid consensus. Even with these stringent criteria, this approach is still more sensitive than other methods for identifying repeats. However, it is also the most computer intensive and misses single-copy TE families because at least two copies are required for detection by self-alignment. Finally, as most TEs are shorter than 25 kb, segmental duplications can be filtered out by removing longer matches.

Once the matches corresponding to repeats have been obtained, they need to be clustered into groups of similar sequences. Each cluster corresponds, hopefully, to copies of a single TE family. However, TEs are interspersed repeats, often nested within each other and diverging with the evolution of the genome, so the task is not straightforward. As a result, algorithms have been designed to cluster the identified sequences properly, limiting the artifacts induced by nested and deleted TE copies and non-TE repeats such as segmental duplications. Various tools exist and make different assumptions about (i) the TE sequence repertoire of families, (ii) the evolutionary dynamics of TE sequences, (iii) nested patterns and (iv) repeat numbers.

GROUPER [25] starts by connecting fragments belonging to the same copy by dynamic programming, and then applies a single link clustering algorithm with a 95% coverage constraint between copies of the same cluster. The rational here is to detect copies that have the same

length as they most probably correspond to mobile entities. Indeed, copies diverge rapidly by accumulating deletions leading to copies with different size. Only copies that are almost intact can transpose and then conserve their original, presumably functional, size. This method is believed to allow the identification of mobile structural variants of a given TE family, *i.e.* related groups of mobile elements differing by their sizes.

RECON [26] also starts with a single link clustering step, but then uses aggregation of the matches' end-points to constitute clusters corresponding to nested repeats. Indeed, nested repeats exhibit a specific pattern in alignments of sequences obtained in an all-by-all genome comparison: the sequence ends of any one inner repeat are all in the same position.

PILER-DF [27] identifies lists of matches covering a maximal contiguous region and defines them as piles, and then builds clusters of globally alignable piles. The rational is here identical to that used by GROUPER where copies of identical length are sought.

Once clusters are defined, a filter is usually applied to retain only those having at least three members, thus discarding the vast majority of segmental duplications. Finally, for each remaining cluster, a multiple alignment is built from which a consensus sequence is derived. Numerous algorithms are available for this but must comply with the following criteria: (i) speed because the number of clusters is usually very high, and (ii) ability to handle sequences of different lengths appropriately, which is the case for the clusters generated by RECON. MAP [28] and MAFFT [29] comply with these criteria and give good results. Taking the 20 longest sequences is generally sufficient to build the consensus. The set of consensus sequences obtained then represents a condensed view of all TE families present in the genome being studied.

For the easily defined TE families, *i. e.* those with full-length copies that are very similar to each other, all clustering methods will find roughly the same consensus. However, for the other families, which may be numerous, different methods generate different clusters because they rely on different assumptions. Therefore manual curation is required to find a correct set of representative sequences.

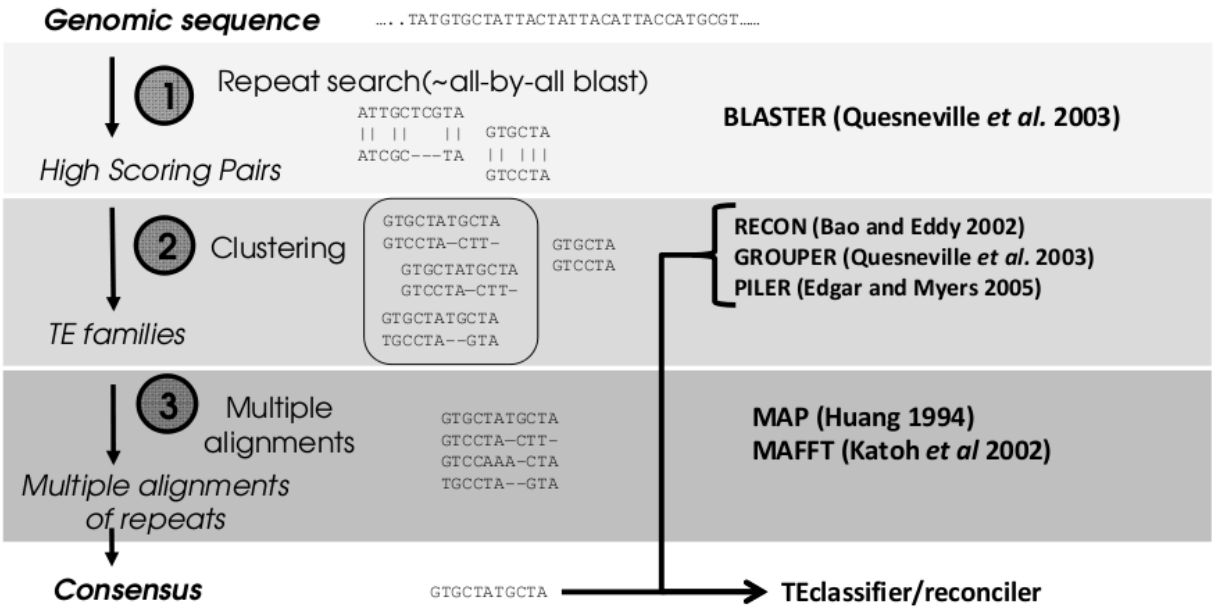


Figure 1: Workflow of the 3-step *de novo* TE detection pipeline [30]

An overall all-by-all genome comparison strategy has been implemented in a pipeline called TEdenovo (Fig.1). This pipeline is part of the REPET package [30] and was designed to be used on a computer cluster for fast calculations. It allows the use of different software at each step to exploit the best strategy according to the genome size and the TE identification goal.

Structure-based methods

TE copies can also be detected by their structural features. This approach depends on a prior knowledge of TE structure such as Long Terminal Repeats (LTR) or Terminal Inverted Repeats (TIR).

The majority of TEs in plant genomes are class I LTR retro-transposons, so many bioinformatic tools have been developed for their detection. They search first for the LTR structure of these retrotransposons and then for other internal features of these elements. LTRharvest [31] is one such tool recently developed to detect LTR TEs in large genomic sequences.

Class II TEs, with the exception of Helitrons, are structurally characterized by TIRs. They can be sought by searching for their TIRs, the approach used by a tool called Must [32]. It was designed

to find MITE elements by searching for TEs containing TIRs adjacent to Direct Repeats (i.e. the Target Site Duplication, also called TSD), distant by less than 500bp. It first finds all TIRs in the genomic sequence and then, using sequence alignments, predicts MITE candidates.

Structural-based *de novo* identifications are generally efficient methods giving good results. Unfortunately, only well-known TEs can be found, and only those that have a strong structural signature. Some TEs do not have such characteristics and thus cannot be found by this type of approach.

Proofs of TE mobility

The presence of a long indel identified by sequence alignments between two closely related species may be an indication of the presence of a TE. The rest of the genome can then be searched for the presence of this sequence to assess its repetitive nature. This approach has already been used [33] and appears to work well with recent TE insertions. Indeed, only insertions that occur after speciation can be detected. Using several alignments with species diverging at different times may lead to more TEs being identified [33] as each alignment would allow detection of TEs inserted at different times. However, one limitation is the difficulty of correctly aligning long genomic sequences from increasingly divergent species.

This idea could be also used within a genomic sequence using segmental duplications. A long indel apparent in sequence alignments of genomic duplications may also be an indication of the presence of a TE [34]. Various controls are needed, however, to confirm the TE status of the sequence. For example, TE features such as terminal repeats (e.g. LTR, TIR) or similarity to other TE sequences could be used. This approach only detects TE insertions that occur after the duplication event and may thus be limited to rare events.

Many TE families generate a double-strand break when they insert into the DNA sequence. The break is caused by the TE enzymatic machinery that generally cuts the DNA with a shift between the two DNA strands. After the insertion, the DNA repair generates a short repeat of few nucleotides (up to 11) at each end; these repeats are called TSD for Target Site Duplication. TSD

are hallmarks of a transposition event, but they can be difficult to find in old insertions because they are short, and they can be altered by mutation or deletion events. In addition the size of the TSD depends on the family and not all TEs generate TSD upon insertion.

CLASSIFICATION AND CURATION OF TRANSPOSABLE ELEMENT SEQUENCES

When they amplify, TE copies may nest within each other in complex patterns [1], thereby fragmenting the elements. With time, they accumulate (i) deletions that can slice copies, (ii) point substitutions and (iii) short indels [35]. All these events generate complex remnants of TEs. Various *de novo* tools use these remnants to try to infer the ancestral sequence that actually transposed.

When starting with a self-alignment of genomic sequences, an optimal strategy would be to use several tools and even combine them. However, whatever the tools, each *de novo* approach can encounter difficulties when trying to distinguish true transposable elements from segmental duplications, multi-member gene families, tandem repeats and satellites. It is therefore strongly recommended to confirm that the predicted sequences can be classified as being TEs. Nevertheless ambiguous cases will always remain, for which automatic analysis still need to be complemented by manual curation.

Classification

Sequences believed to correspond to TEs can be classified according to their similarity to known TEs as those recorded in databases like RepbaseUpdate [36]. For instance, a tool called TEclass [37] implements a support vector machine using oligomer frequencies to classify TE candidates.

However, for most unknown TE sequences obtained *via de novo* approaches from non-model organisms, classifying them will require to specific identification of several TE features (see [5] for complete description). By searching for structural features, such as terminal repeats, one can identify long terminal repeats specific to class I LTR retro-transposons, terminal inverted repeats

specific to the class II DNA transposons, and poly-A or SSR-like tails specific to class I non-LTR retrotransposons. Then, the comparison of TE candidates with a reference data-bank *via* blastn, blastx and tblastx usually provides hints about the classification, as long as TE candidate has equivalents in this kind of reference data-bank. Therefore, it is also judicious to search for matches for TE-specific protein profiles in TE sequences. For example, the presence of a transposase is highly indicative of a class II DNA transposon. Such protein profiles can be obtained from the Pfam database that gathers together protein families represented by multiple sequence alignments and hidden Markov models (HMM) [38]. These profiles are used by programs such as HMMER to find matches within the input sequences.

MGEScan-non-LTR [39] identifies and classifies non-LTR TEs in genomic sequences using probabilistic models. It is based on the structure of the 12 TE clades known as non-LTR TEs. It uses two separate HMM profiles, one for Reverse Transcriptase (RT) and one for endonuclease (APE) that are both well conserved among non-LTR TEs. Other tools classify TE sequences according to their features, usually *via* a decision tree. The TEclassifier in the REPET package [30] searches for all the features listed above. REPCLASS [40] also searches for all the features listed above, but it does not use HMM profiles. In addition, REPCLASS allows TE candidates to be filtered on the basis of the number of copies they have in the genome. TEclassifier interestingly allows the removal of redundancy among potential TE sequences based on their classification. Hence, a well-classified TE candidate will not be filtered out because it is included within a large segmental duplication. This tool is particularly useful to reconcile several TE reference libraries obtained independently.

Identification of families

Once unknown TE sequences have been classified, manual curation is still needed. Some consensus sequences may still be unclassified and there may still be some redundant consensus sequences. Manual curation is crucial because the annotation of TE copies described in the next section depends on the quality of the TE library. One way to curate a library of TE consensuses is to gather these sequences into clusters that could represent TE families. A tool like BLASTCLUST

in the NCBI-BLAST suite [41] can quickly build such clusters *via* simple link clustering based on sequence alignment coverage and identity. Eighty percent identity and coverage, as proposed by [5], gives good results. Typical clusters will contain well-classified consensuses (*e. g.* class I – LTR – Gypsy element) as well as unclassified consensuses (without structural features, and no sequence similarity either with known TEs or any TE domain).

Then, computing a multiple sequence alignment (MSA) for each cluster gives a useful view of the relationships between the consensuses such that it is possible to assess whether they belong to the same TE family. One of the programs detailed above, MAP [28] or MAFFT [29], could be used. It can also be informative to build a MSA with the consensus, and with the genomic sequences from which these consensuses were derived and/or the genomic copies that each of these consensuses can detect. In such cases, we advise first building a single MSA for each consensus with the genomic sequences it detects, and then building a global MSA by aligning these multiple alignments together, using for instance the “profile” option of the MUSCLE program [42]. Finally, after a visual check of the MSA with the evidence used to assign a classification to the consensus, it is then possible to tag all consensuses in the same cluster with the most frequent TE class, order, super-family and family, as far as possible (Fig.2). The MSA can be also edited by splitting it or deleting sequences to obtain a MSA corresponding to a single TE family. Indeed, in some cases, chimeric consensuses are easily identified by looking at such MSA and can then be either removed from the library if artifactual or used to build a new TE family if several copies support it.

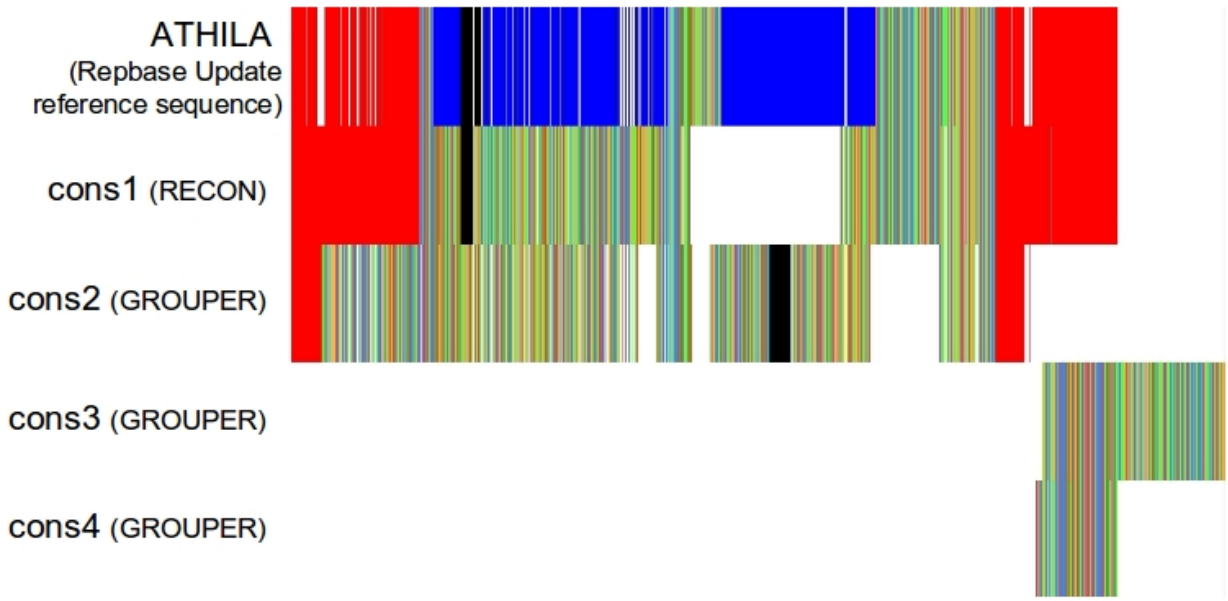


Figure 2: Alignment (Jalview [54] screenshot) of *de novo* TE consensus sequences with Athila, the best-matching known TE from Repbase Update. They are represented with some of the features shown: LTRs (red), ORFs (blue) and matches with HMM profiles (black). The differences between the consensuses obtained by different methods, here RECON (cons1) and GROUPER (cons2, cons3, cons4), are illustrated. Manual curation would remove cons3 as it corresponds to a single LTR with short sequences not present in the Athila family and cons4 as it corresponds to a LTR probably formed from the Athila solo-LTRs of the genome. A good consensus for the family would be a combination of cons1 and cons2.

Phylogenies of TE family copies and/or consensuses provide another view of the members in a TE family. This can help the curation if the cluster has many members, or when two or more sub-families are present. In such cases, sub-families can be hard to detect by examination of the MSA alone, but become evident in a phylogeny. Such phylogenies can be constructed from the MSA with currently available software, including the PhyML program [43]. However, as most phylogeny programs do not consider gaps, branch length may be biased when consensuses are of very different lengths.

ANNOTATION OF TRANSPOSABLE ELEMENT COPIES

This second phase involves annotating all TE copies in the genome, resolving the most complex degenerate or nested structures. This requires a library of reference sequences representing the TE families. In the best case, the library is both exhaustive and non-redundant, *i. e.* each ancestral TE, autonomous or not, is represented by a single consensus sequence. We usually use the manually curated library built in the previous phase, *i.e.* the *de novo* TE detection, as detailed in the previous section. Then, with these reference sequences, all TE fragments are mapped and connected if belonging to the same TE copies.

Detecting TE fragments

The first step corresponds to mining the genomic sequences with the TE library *via* local pairwise alignments. Several tools were designed specifically for this purpose, such as REPEATMASKER [44], CENSOR [45, 46] and BLASTER [25]. Some of these tools incorporate scoring matrices to be used with particular GC percentages, as is the case for isochores in the human genome. All these tools propose a small set of parameter combinations depending on the level of sensitivity required by the user.

Although similar, these tools are complementary. We have shown previously that combining these three programs is the best strategy [22]. The MATCHER program [25] can then be used to assess these multiple results and keep only the best.

Whatever the parameters used for the pairwise alignments, some of the matches will be false positives, e. g. a TE reference sequence matching at a locus although no TE is present. For protein-coding genes full-length cDNAs can be used for confirmation; unfortunately, there is no equivalent in TE annotation. To assess the false positive risk, an empirical statistical filter is implemented in the TEannot pipeline (REPET package). The genomic sequences are shuffled and screened with the TE library. Finally, only the matches on the true genomic sequences having a score higher than the best match obtained on the shuffled genomic sequences will be kept. This procedure guarantees that none of the observed matches can be obtained with random

sequences.

Filtering satellites

Short simple repeats (SSRs) are short motifs repeated in tandem. Many transposable element sequences contain SSRs but they are also present in the genome independently of TEs. It is therefore necessary to filter out TE matches if they are restricted to SSR that the TE consensus may contain. This is easily achieved by annotating SSRs and then removing TE matches included in SSR annotations. Several efficient programs, for example TRF [47], MREPS [48] and REPEATMASKER [44], are available for this. These three programs are launched in parallel and their results subsequently combined in the TEannot pipeline (REPET package).

Satellites are longer motifs, around 100 bp long, also repeated in tandem. As such they are not TEs although they are sometimes difficult to distinguish because they contain part of them. PILER-TA [27] detects pyramids in a self-alignment of the genomic sequences. These pyramids can be used to make a consensus of the satellite unit motif. These consensuses can then be aligned on the whole genome to find all their occurrences and to discriminate them from TEs.

Connecting TE fragments to recover TE copies

Even when TE fragments have been mapped on the genome, the work is only half-finished. Indeed, TE copies can be disrupted into several fragments. A complete TE annotation requires retrieving all copies, and thus joining fragments belonging to the same copy when it has transposed.

The first, historical method was manual curation using dotplots. However, apart from being tedious and curator-dependent, it is impractical for large genomes. It requires that the curator has detail knowledge about transposable elements and it ignores the age of nested fragments, hence possibly leading to incongruities. Therefore several computational approaches have been proposed. Many of them are reviewed in the article by Pereira [49].

Joining TE fragments to reconstruct a TE copy is known as the “chain problem” as it corresponds

to finding the best chain of local pairwise alignments. The optimal solution is found *via* dynamic programming as implemented in MATCHER [25]. Subsequently, additional considerations related to the biology of TEs can be taken into account. Two TE fragments distant from each other but mostly separated by other TE fragments (e. g. at least 95% as in heterochromatin) can be joined as long as the TE fragments between them are younger. The age can be approximated using the percentages of identity of the matches between the TE reference sequences and the fragments. This procedure is implemented in the TEannot pipeline (Fig.3) under the procedure name “long join”.

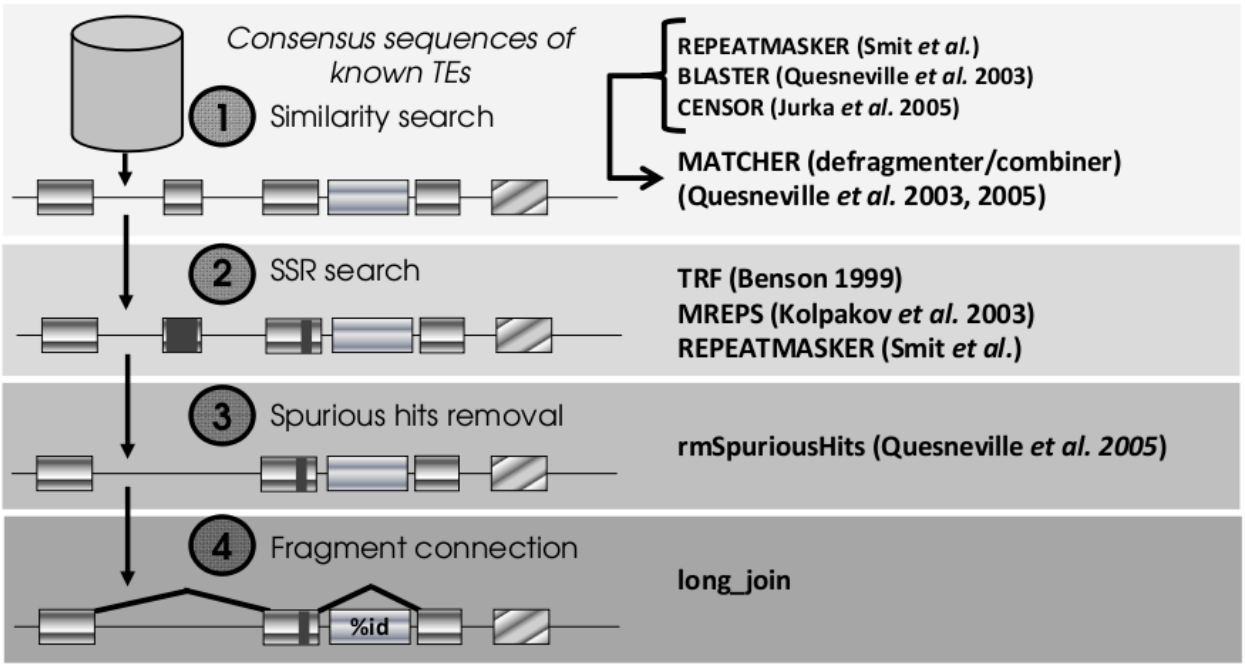


Figure 3: The four steps of the TEannot pipeline [22]

DISCUSSION

The contribution of TEs to genome structure and evolution, and their impact on genome assembly has generated increasing interest in the development of new methods for their computational analysis. The most common strategy is to detect pairs of similar sequences at different locations in a self-against-self genome comparison, and then cluster these pairs to obtain repeat families. These methods are not specific to TEs and they therefore find repeats generated by many different processes including tandem repeats, segmental duplications and satellites, in addition to TE sequences. Moreover, TE copies can be highly degenerate, deleted or nested, so repeat detection methods can make errors in detecting individual TE copies and consequently in defining TE families. In view of these issues, we believe that existing automatic approaches still need to be supplemented by expert manual curation. At this step, careful examination is required as some identified families that may appear as artifactual can in fact be previously unidentified or unusual TE families. Indeed, well documented cases show that some TEs families can appear confusing as they may (i) include cellular genes or parts of genes (*e.g.* pack-MULEs [50] or Helitrons [51]), (ii) be restricted to rDNA genes (*e.g.* the R2 Non-LTR retroelement superfamily [52], or (iii) form telomeres (in *Drosophila* [53]). Close examination of non-canonical cases may reveal new and interesting TE families or particular transposition events (*e.g.* macrotranspositions [8]).

Knowledge-based TE detection methods (i.e. based on structure or similarity to distant TEs) have distinct advantages over *de novo* repeat discovery methods in that they capitalize on prior knowledge captured in the large number of previously reported TE sequences. Thus, they are more likely to detect *bona fide* TEs, including even those present as only a single copy in the genome. However, these methods are not well suited

to detecting new TEs (especially of distinctly new types). Moreover, these methods have intrinsic ascertainment biases. For example, miniature inverted repeat transposable elements (MITEs) and short interspersed nuclear elements (SINEs) will be underrepresented if we rely exclusively on similarity-based methods because these TEs are composed entirely of non-coding sequences.

We have assessed the relative benefits of using different programs for TE detection, clustering and multiple alignments [22, 30]. Our investigations suggest that only combined approaches, using both *de novo* and knowledge-based TE detection methods are likely to produce reasonably comprehensive and sensitive results. In view of this, the REPET package [30] has been developed, composed of two pipelines, TEdenovo and TEannot. These pipelines launch several different prediction programs in parallel and then combine their results to improve the accuracy and exhaustiveness of the detection.

Sequencing costs have dropped dramatically and sequences have thus become easier to obtain. However, sequence analysis remains a major bottleneck. Efficient analysis pipelines are required. They need to be quick and robust to scale up the pace of data production, and also exploit the knowledge of the few specialists able to perform genome analysis on a large scale such that these technologies are made available to a large community of scientists.

KEY POINTS

Thanks to the development of new sequencing techniques, the number of sequenced eukaryotic genomes is ever increasing. However, their accurate annotation remains a major challenge, and in particular as concerns TE. Combined approaches, using both *de novo* and knowledge-based TE detection methods are likely to produce reasonably comprehensive and sensitive results. Nevertheless, existing automatic approaches still need to be supplemented by expert manual curation.

ACKNOWLEDGMENTS

This work was supported in part by grants from the Agence Nationale de la Recherche (Holocentrism project, to HQ [grant number ANR-07-BLAN-0057]), the Centre National de la Recherche Scientifique - Groupement de Recherche “Elements Transposables” and. TF was supported by a PhD studentship form the the Institut National de la Recherche Agronomique. EP was supported by a Post-Doctoral fellowship form the Agence Nationale de la Recherche.

REFERENCES

1. Bergman CM, Quesneville H, Anxolabehere D, et al. *Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome.* Genome Biol, 2006. **7**(11): p. R112.
2. Lander ES, Linton LM, Birren B, et al. *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
3. Schnable PS, Ware D, Fulton RS, et al. *The B73 maize genome: complexity, diversity, and dynamics.* Science, 2009. **326**(5956): p. 1112-5.
4. Finnegan DJ. *Eukaryotic transposable elements and genome evolution.* Trends Genet, 1989. **5**(4): p. 103-7.
5. Wicker T, Sabot F, Hua-Van A, et al. *A unified classification system for eukaryotic transposable elements.* Nat Rev Genet, 2007. **8**(12): p. 973-82.
6. Petrov DA. *Evolution of genome size: new approaches to an old problem.* Trends Genet, 2001. **17**(1): p. 23-8.
7. Piegu B, Guyot R, Picault N, et al. *Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice.* Genome Res, 2006. **16**(10): p. 1262-9.
8. Gray YH. *It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements.* Trends Genet, 2000. **16**(10): p. 461-8.
9. Fiston-Lavier AS, Anxolabehere D, Quesneville H. *A model of segmental duplication formation in Drosophila melanogaster.* Genome Res, 2007. **17**(10): p. 1458-70.
10. Orgel LE, Crick FH. *Selfish DNA: the ultimate parasite.* Nature, 1980. **284**(5757): p. 604-7.

11. Feschotte C. *Transposable elements and the evolution of regulatory networks.* Nat Rev Genet, 2008. **9**(5): p. 397-405.

12. Bourque G. *Transposable elements in gene regulation and in the evolution of vertebrate genomes.* Curr Opin Genet Dev, 2009. **19**(6): p. 607-12.

13. Zhou L, Mitra R, Atkinson PW, et al. *Transposition of hAT elements links transposable elements and V(D)J recombination.* Nature, 2004. **432**(7020): p. 995-1001.

14. Bundock P, Hooykaas P. *An Arabidopsis hAT-like transposase is essential for plant development.* Nature, 2005. **436**(7048): p. 282-4.

15. Santangelo AM, de Souza FS, Franchini LF, et al. *Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene.* PLoS Genet, 2007. **3**(10): p. 1813-26.

16. Kapitonov VV, Jurka J. *RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons.* PLoS Biol, 2005. **3**(6): p. e181.

17. Bergman CM, Quesneville H. *Discovering and detecting transposable elements in genome sequences.* Brief Bioinform, 2007. **8**(6): p. 382-92.

18. Kurtz S, Narechania A, Stein JC, et al. *A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes.* BMC Genomics, 2008. **9**: p. 517.

19. Gu W, Castoe TA, Hedges DJ, et al. *Identification of repeat structure in large genomes using repeat probability clouds.* Anal Biochem, 2008. **380**(1): p. 77-83.

20. Li R, Ye J, Li S, et al. *ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun.* PLoS Comput Biol, 2005. **1**(4): p. e43.

21. Price AL, Jones NC, Pevzner PA. *De novo identification of repeat families in large genomes.* Bioinformatics, 2005. **21 Suppl 1**: p. i351-8.

22. Quesneville H, Bergman CM, Andrieu O, et al. *Combined evidence annotation of transposable elements in genome sequences.* PLoS Comput Biol, 2005. **1**(2): p. 166-75.

23. Altschul SF, Madden TL, Schaffer AA, et al. *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

24. Rasmussen KSJ, Myers EW. *Efficient q-Gram Filters for Finding All e-Matches over a Given Length*. in *RECOMB*. 2005.

25. Quesneville H, Nouaud D, Anxolabehere D. *Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes.* J Mol Evol, 2003. **57 Suppl 1**: p. S50-9.

26. Bao Z, Eddy SR. *Automated de novo identification of repeat sequence families in sequenced genomes.* Genome Res, 2002. **12**(8): p. 1269-76.

27. Edgar RC, Myers EW, *PILER: identification and classification of genomic repeats.* Bioinformatics, 2005. **21 Suppl 1**: p. i152-8.

28. Huang X. *On global sequence alignment.* Comput Appl Biosci, 1994. **10**(3): p. 227-35.

29. Katoh K, Misawa K, Kuma K, et al. *MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.* Nucleic Acids Res, 2002. **30**(14): p. 3059-66.

30. Flutre .T, Quesneville H. *The REPET package*. http://urgi.versailles.inra.fr/development/repet

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

32. Chen Y, Zhou F, Li G, et al. *MUST: a system for identification of miniature inverted-repeat transposable elements and applications to Anabaena variabilis and Haloquadratum walsbyi.* Gene, 2009. **436**(1-2): p. 1-7.

33. Caspi A, Pachter L. *Identification of transposable elements using multiple alignments of related genomes.* Genome Res, 2006. **16**(2): p. 260-70.

34. Le QH, Wright S, Yu Z, et al. *Transposon diversity in Arabidopsis thaliana.* Proc Natl Acad Sci U S A, 2000. **97**(13): p. 7376-81.

35. Blumenstiel JP, Hartl DL, Lozovsky ER. *Patterns of insertion and deletion in contrasting chromatin domains.* Mol Biol Evol, 2002. **19**(12): p. 2211-25.

36. Jurka J, Kapitonov VV, Pavlicek A, et al. *Repbase Update, a database of eukaryotic repetitive elements.* Cytogenet Genome Res, 2005. **110**(1-4): p. 462-7.

37. Abrusan G, Grundmann N, DeMester L, et al. *TEclass--a tool for automated classification of unknown eukaryotic transposable elements.* Bioinformatics, 2009. **25**(10): p. 1329-30.

38. Finn RD, Mistry J, Tate J, et al. *The Pfam protein families database.* Nucleic Acids Res. **38**(Database issue): p. D211-22.

39. Rho M, Tang H. *MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes.* Nucleic Acids Res, 2009. **37**(21): p. e143.

40. Feschotte C, Keswani U, Ranganathan N, et al. *Exploring Repetitive DNA Landscapes Using REPCLASS, a Tool That Automates the Classification of Transposable Elements in Eukaryotic Genomes.* Genome Biol Evol, 2009. **2009**: p. 205-20.

41. NCBI. *NCBI blast suite* - ftp://ftp.ncbi.nih.gov/blast/

42. Edgar RC. *MUSCLE: a multiple sequence alignment method with reduced time and space complexity.* BMC Bioinformatics, 2004. **5**: p. 113.

43. Guindon S, Gascuel O. *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.* Syst Biol, 2003. **52**(5): p. 696-704.

44. Smit AFA, Hubley R, Green P. *RepeatMasker Open-3.0.* 1996-2004, Institute for Systems Biology.

45. Jurka J, Klonowski P, Dagman V, et al. *CENSOR--a program for identification and elimination of repetitive elements from DNA sequences.* Comput Chem, 1996. **20**(1): p. 119-21.

46. Kohany O, Gentles AJ, Hankus L, Jurka J et al. *Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor.* BMC Bioinformatics, 2006. **7**: p. 474.

47. Benson G. *Tandem repeats finder: a program to analyze DNA sequences.* Nucleic Acids Res, 1999. **27**(2): p. 573-80.

48. Kolpakov R, Bana G, Kucherov G. *mreps: Efficient and flexible detection of tandem repeats in DNA.* Nucleic Acids Res, 2003. **31**(13): p. 3672-8.

49. Pereira V. *Automated paleontology of repetitive DNA with REANNOTATE.* BMC Genomics, 2008. **9**: p. 614.

50. Jiang N, Bao Z, Zhang X, et al. *Pack-MULE transposable elements mediate gene evolution in plants.* Nature, 2004. **431**(7008): p. 569-73.

51. Morgante M, Brunner S, Pea G, et al. *Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize.* Nat Genet, 2005. **37**(9): p. 997-1002.

52. Eickbush TH, Burke WD, Eickbush DG, et al. *Evolution of R1 and R2 in the rDNA units of the genus Drosophila.* Genetica, 1997. **100**(1-3): p. 49-61.

53. Clark AG, Eisen MB, Smith DR, et al. *Evolution of genes and genomes on the Drosophila phylogeny.* Nature, 2007. **450**(7167): p. 203-18.

54. Clamp M, Cuff J, Searle SM, et al. *The Jalview Java alignment editor.* Bioinformatics, 2004. **20**(3): p. 426-7.

3.2 CONSOLIDATION ET AMÉLIORATION DE L'APPROCHE D'ANNOTATION *de novo*

L'approche *de novo* de détection d'ET présentée précédemment (chapitre 2) repose sur l'idée que ceux-ci, de par leur capacité à transposer, sont présents dans un génome en de multiples copies. Vu sous cet angle, identifier les ET revient à identifier des répétitions dans une chaîne de caractères, puis regrouper les répétitions similaires les unes aux autres, en dériver une séquence consensus et enfin isoler celles correspondant aux ET de celles correspondant à d'autres types de répétitions.

Cette stratégie, bien que simple à comprendre, n'est pas triviale à mettre en œuvre de manière efficace. En effet, les termes *répétitions* et *similaires* demandent à être caractérisés par des paramètres tels que nombres et tailles des répétitions, niveau de similarité, etc. L'action de *regrouper* nécessite de définir un critère permettant d'obtenir des groupes homogènes, regrouper par exemple toutes les copies d'une même famille d'ET. La mise en œuvre d'une procédure automatique est soumise à des contraintes techniques de temps de calcul, d'espace mémoire et d'espace de disque. Finalement, la curation manuelle encore nécessaire pour répondre aux questions biologiques de façon pertinente doit être facilitée pour être rapide.

Lors de mon travail de recherche, je me suis confronté à toutes ces questions et ai pu y apporter quelques réponses (chapitre 2). Il reste cependant encore des pistes à explorer pour améliorer l'efficacité de l'identification *de novo* des ET, notamment concernant la définition des variants structuraux aboutissant à représenter une famille d'ET par plusieurs séquences consensus.

3.2.1 *Détection des répétitions et regroupement des variants structuraux d'éléments transposables*

L'approche *de novo* présentée précédemment (chapitre 2) tente de prendre en compte la diversification des familles d'ET lors de l'annotation des copies. Ce processus de diversification sculpte les génomes, sa compréhension est donc essentielle à celle de la dynamique des génomes. Quand une nouvelle copie apparaît par transposition, les séquences des deux copies vont diverger au fil du temps. Cette divergence se matérialise par une accumulation de substitutions et d'indels. Si une copie mutée a toujours la capacité de transposer, ses copies posséderont ses mutations. Ce sont précisément ces copies que nous cherchons à regrouper

et identifier comme variants d'une famille d'ET plus large. Une famille est alors caractérisée par un ensemble de variants structuraux qui retracent la diversification des copies de cette famille. A l'heure actuelle, la communauté s'est entendue sur une règle empirique : deux copies d'au moins 80 paires de bases appartiennent à la même famille si elles s'alignent sur au moins 80% de leur longueur avec une identité supérieure à 80% (Wicker et coll. 2007). Cette règle est pertinente une fois que les copies ont été identifiées, mais dans une approche *de novo* telle que la nôtre, après une comparaison du génome contre lui-même, une telle règle n'est pas utilisable en l'état. Comme expliqué dans l'article décrivant le programme RECON (Bao et Eddy 2002), regrouper les répétitions avec cette règle uniquement ne permet pas d'inférer correctement les familles d'ET car la plupart des copies sont plus ou moins tronquées et les duplications segmentaires peuvent contenir des fragments d'ET.

La reconstruction des séquences ancestrales est tributaire de la première étape de comparaison du génome avec lui-même. Cette comparaison permet d'obtenir les coordonnées génomiques des répétitions suffisamment similaires les unes aux autres et ayant des tailles proches de celles d'ET. Comme montré précédemment (chapitre 2), plusieurs programmes peuvent être utilisés avec des efficacités variables. Par exemple, entre PALS qui implémente un algorithme exact et BLASTER qui est basé sur l'heuristique BLAST, ce dernier offre le meilleur compromis entre sensibilité et rapidité. D'autres programmes ont été mis au point, notamment YASS (Noé et Kucherov 2005) qui présente des perfectionnements par rapport à BLAST. Ces deux algorithmes commencent par détecter de courtes régions ancres appelées *graines* avant d'essayer de les étendre. Alors que BLAST utilise deux graines exactes non-chevauchantes appartenant à la même diagonale dans une taille de fenêtre fixée, YASS utilise des groupes de graines chevauchantes avec un nombre maximal de substitutions et pouvant être sur des diagonales proches mais différentes, autorisant ainsi les indels. Les résultats des auteurs du programme montrent une amélioration de la sensibilité tout en gardant des temps de calcul équivalent (Noé et Kucherov 2004). J'ai également testé un autre programme appelé BLAT (Kent 2002) dont l'algorithme calcule un index de tous les mots de longueur k non-chevauchants dans le génome (k=11 pour l'ADN). Cet index sert ensuite à détecter des *matches* entre une séquence d'intérêt et le génome. Ce programme est surtout utilisé pour détecter rapidement la localisation d'un gène dans le génome en

utilisant des transcrits sous forme d'ADN complémentaire. En ce sens, il pourrait être performant pour la détection de répétitions longues et peu divergentes.

J'ai donc testé ces deux nouveaux programmes dans la première étape de l'outil TEdenovo avec les génomes de *D. melanogaster* (tableau 5) et *A. thaliana* (tableau 6), rappelant par la même occasion les résultats obtenus avec BLASTER et PALS. Toutes les options ne peuvent pas toujours être données directement aux programmes. Ainsi, YASS a été lancé en ne gardant que les *matches* ayant une E-value inférieure à 10^{-300} et BLAT a été lancé en ne gardant que les *matches* ayant une identité supérieure à 90%. Une fois leurs alignements obtenus, seuls ont été conservés ceux ayant une E-value inférieure à 10^{-300}, une identité supérieure à 90% et une longueur entre cent et vingt mille nucléotides. Dans le meilleur des cas, nous voudrions un programme capable de détecter rapidement des répétitions recouvrant une bonne partie du contenu en ET du génome, tout en évitant de trop les fragmenter en plusieurs alignements.

TAB. 5: Comparaison du temps de calcul et de la couverture du génome de *D. melanogaster* entre différents programmes utilisés pour la détection des répétitions lors de la première étape de l'outil TEdenovo.

Programme	Durée moyenne d'un job	Nombre d'alignements	Couverture du génome
BLASTER	70 sec	109882	7.41%
PALS	49 sec	105059	7.38%
YASS	32 min	31973	6.52%
BLAT	22 min	691347	9.34%

Cette première comparaison montre que le programme BLASTER offre le meilleur compromis entre vitesse et couverture du génome. En effet, BLAT fournit une meilleure couverture du génome mais au prix de calculs beaucoup plus longs et d'une très grande fragmentation des répétitions. Yass quant à lui construit moins d'alignements, ceux-ci étant plus longs, mais au prix de calculs là aussi beaucoup plus longs que BLASTER, d'une plus faible couverture du génome. Cependant, dans le but de reconstruire les séquences ancestrales d'ET, se contenter de comparer le temps de calcul et la couverture du génome n'est pas suffisant. Ainsi, afin de véritablement tester l'efficacité de ces programmes,

TAB. 6: Comparaison du temps de calcul et de la couverture du génome d'*A. thaliana* entre différents programmes utilisés pour la détection des répétitions lors de la première étape de l'outil TEdenovo.

Programme	Durée moyenne d'un job	Nombre d'alignements	Couverture du génome
BLASTER	2 min	103728	13.48%
PALS	53 sec	51023	10.53%
YASS	42 min	34553	10.48%
BLAT	39 min	1078698	19.24%

j'ai regroupé leurs répétitions avec le programme GROUPER lors de l'étape 2 de l'outil TEdenovo et j'ai comparé les séquences consensus obtenues avec les banques de référence comme expliqué précédemment (chapitre 2). J'ai exécuté cette procédure à partir des répétitions détectées par chaque programme d'alignement pour les génomes de *D. melanogaster* (tableau 7) et *A. thaliana* (tableau 8).

TAB. 7: Validation des séquences consensus obtenues à partir des répétitions détectées par différents programmes d'alignement dans le génome de *D. melanogaster*.

Programme	Sensibilité (Sn*)	Spécificité (Sp*)	Ratio des complets (R_{CC})
BLASTER	80.34%	85.89%	66.20%
PALS	73.50%	88.75%	60.30%
YASS	70.09%	90.58%	55.90%
BLAT	91.45%	88.11%	52.95%

Dans les deux génomes, BLAT atteint le meilleur résultat en sensibilité. Il est également le meilleur en spécificité chez *A. thaliana*. Cependant, une plus faible proportion des séquences consensus produites à partir de ses alignements sont complètes, comparé à BLASTER. Quant à YASS, sa spécificité est très bonne mais sa sensibilité est la plus faible dans les deux génomes, et trop peu de séquences consensus sont considéres comme complètes. Ainsi, là encore, les résultats montrent que BLASTER reste la meilleur option à l'heure actuelle. A l'avenir, si l'on souhaite améliorer les performances de cette première étape, il

TAB. 8: Validation des séquences consensus obtenues à partir des répétitions détectées par différents programmes d'alignement dans le génome d'*A. thaliana*.

Programme	Sensibilité (Sn*)	Spécificité (Sp*)	Ratio des complets (R_{CC})
BLASTER	60.33%	82.42%	39.00%
PALS	54.75%	88.38%	24.00%
YASS	45.57%	89.12%	17.50%
BLAT	66.89%	90.56%	37.00%

semble qu'il faille mettre au point un algorithme spécifiquement optimisé pour cette question. Pour ce faire, il serait intéressant de tester le programme GLINT (Faraut et Courcelle, en préparation), spécifiquement mis au point pour aligner des génomes complets entre eux.

Concernant la deuxième étape de l'outil TEdenovo, il existe également des pistes qui pourraient être explorées à l'avenir. Cette deuxième étape correspond au regroupement des répétitions ainsi identifiées dans des groupes de façon à représenter une partie de l'histoire évolutive d'une famille d'ET. Nous avons pu montrer précédemment (chapitre 2) qu'il est nécessaire de combiner deux programmes, GROUPER et RECON, afin de reconstruire le plus de séquences de référence représentant les ET ancestraux. Or cela a notamment pour conséquence de former des groupes redondants dont il faut ensuite construire un alignement multiple à partir duquel sera dérivée une séquence consensus qui, finalement, sera enlevée par l'outil TEclassifier. Un gain de temps très appréciable serait de pouvoir éviter cette redondance dès la deuxième étape de l'outil TEdenovo. Ce gain de temps devient même crucial pour l'analyse des grands génomes (tableaux 1 et 2).

Ces deux algorithmes sont complémentaires. RECON, a l'avantage sur GROUPER de rassembler en un même groupe des copies de taille hétérogène, ce qui permet de ne pas construire un nombre trop abondant de groupes. En retour, pour éviter de former des groupes composites avec des duplications segmentaires contenant des ET, RECON implémente plusieurs procédures identifiant et éliminant les éléments composites. Cependant, les résultats montrent que les banques de consensus issus de groupes de RECON ont une faible spécificité (chapitre 2), ce qui revient à dire qu'une grande proportion de consensus ne

semble pas correspondre à des ET mais à d'autres types de répétitions. A l'inverse, les banques de consensus issus de GROUPER sont bien plus spécifiques.

Par ailleurs, RECON gère l'existence de familles proches les unes des autres, et permet ainsi de distinguer les ET non-autonomes des ET autonomes dont ils sont dérivés. Cependant, certains cas de variants structuraux ne peuvent vraisemblablement pas être identifiés. Ainsi, un ET chimérique, bien qu'ayant transposé plusieurs fois, risque d'être interprété comme étant une duplication segmentale. De même, dans le cas d'une famille dont les quelques copies sont toutes fragmentées, si les fragments ne s'alignent entre eux que sur de petites fractions, RECON reconstruira des groupes pour les fragments et non un seul avec les copies réelles. Sur ces points, GROUPER vise justement à d'abord reconstruire les copies en connectant les fragments qui peuvent l'être, puis à regrouper les copies en appliquant une contrainte de couverture très grande (95%). Le risque est alors de générer un trop grand nombre de groupes. Et en effet, GROUPER construit plus de groupes que RECON qui déjà en construit beaucoup plus que PILER. De plus, appliquer le TEclassifier aux consensus dérivés des groupes de GROUPER montre que ceux-ci sont généralement très redondants.

Cependant, GROUPER, en commençant par reconstruire les copies fragmentées, permet de distinguer des variants structuraux. Par exemple, pour la famille *Doc* chez *D. melanogaster*, GROUPER construit deux consensus très similaires entre eux, l'un différant de l'autre principalement par une seule délétion centrale (figure 8). La séquence de référence connue faisant 4700 paires de bases et la délétion faisant 330 paire de bases, les copies tronquées ne recouvrent les copies complètes qu'à 93%, et sont donc regroupées dans un autre groupe. Dans ce cas, il est évident que le consensus complet permettra de trouver également les copies possédant la délétion, et que donc la présence de ce variant pourra être identifiée a posteriori en observant un alignement multiple des copies. On peut alors se demander s'il est toujours pertinent de distinguer ces variants avec des consensus différents malgré leur réalité biologique. En effet, lors de la phase d'annotation des copies, les deux consensus risquent d'interférer, empêchant des connections entre fragments de se faire correctement. Cependant, l'élément *Doc* est un retrotransposon de type LINE, type connu pour leurs variants structuraux différant par leur extrémité 5'. Il n'est généralement pas fréquent de voir transposer un LINE ayant une délétion interne. Or ici, le

consensus correspondant au variant possédant la délétion a été formé à partir de 5 séquences génomiques qui possèdent toutes cette délétion centrale. Il est donc peu probable que ces 5 copies soient apparues par duplications mais bien par transposition. Ainsi, il n'est peut-être pas nécessaire de garder les deux consensus pour l'annotation des copies avec l'outil TEannot, mais il est important de mettre en évidence ce variant structural, à l'origine peut-être d'un élément non-autonome. Dans le cas de grands génomes, travailler sur plusieurs milliers de consensus évite de travailler sur plusieurs centaine de milliers de copies, voire plus. La banque de consensus peut donc être vue aussi comme un résumée de la dynamique des ET dans le génome, dont l'intérêt et la pertinence vont bien au-delà d'une simple collection de séquences permettant de masquer ou annoter des répétitions.

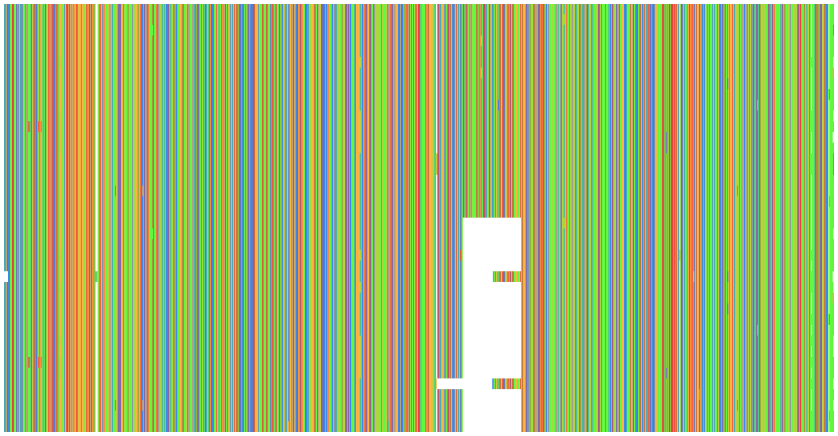


FIG. 8: Alignement multiple des répétitions géomiques à partir desquelles deux variants structuraux sont mis en évidence pour l'ET *Doc* dans le génome de *D. melanoasgter*

Partant des diagrammes de Venn présentés précédemment (chapitre 2) et des exemples de variants structuraux tel celui présenté ci-dessus (figure 8), il est clair que les programmes GROUPER et RECON sont complémentaires. Une piste d'amélioration serait donc d'augmenter la stringence de GROUPER pour identifier spécifiquement les variants structuraux importants quantitativement, issus de familles très fragmentées. RECON serait quant à lui exécuté comme d'habitude afin de continuer à identifier toutes les autres familles. Ainsi GROUPER serait utilisé seulement pour reconstruire les ET que RECON ne pourrait pas détecter et génèrerait donc peu de groupes redondants. En pratique, on pourrait augmenter la stringence de GROUPER en mo-

difiant le critère d'inclusion dans un groupe. Pour l'instant, deux alignements de répétitions sont regroupées si ils se chevauchent sur au moins 95% de leur longueur. Ce critère dépend de la taille, il est donc moins stringent pour les longues répétitions. Il pourrait être amélioré si l'on ne considère plus un pourcentage de couverture, mais un nombre de bases différentes sur l'ensemble de la séquence ou sur les extrémités. En augmentant la stringence du critère de regroupement, plus de groupes n'auront que deux copies et seront donc éliminés.

Cependant, cette modification ne peut résoudre la faible spécificité des groupes construits par RECON. Une autre piste consiste à unifier les deux algorithmes de GROUPER et RECON en un seul algorithme. Celui-ci commencerait par connecter les fragments répétés par programmation dynamique comme dans GROUPER, puis les chaines de fragments obtenues seraient regroupées à l'aide des procédures de RECON. La difficulté consiste à garder les variants structuraux tout en évitant la redondance parmi les groupes. Il faut donc ajouter des règles de décisions concernant les indels internes. Par exemple, dans quel cas une copie tronquée doit-elle être fusionnée avec une autre ou bien doit-on créer un nouveau groupe pour elle ? Ceci passe par une meilleure compréhension de la diversification des copies et son implémentation dans un algorithme. Plus précisément, un travail d'analyse des indels (longueurs et positions) est nécessaire au design d'un tel algorithme.

Pour aider à ces développements méthodologiques, il serait souhaitable de développer un outil de simulation de séquences d'ET dans des génomes, permettant de tester en pratique les algorithmes lors de leur mise au point. Un tel outil ne serait pas uniquement utile pour les développements méthodologiques puisqu'il permettrait de tester également différents scénarios évolutifs et donc la capacité des algorithmes à identifier correctement les familles et leurs variants. Un tel outil de simulation pouvant être envisagé sous différents angles, je développe ce thème à la fin de ce chapitre (voir 3.3.2).

Plusieurs pistes sont possibles pour tenter d'améliorer les algorithmes utilisés dans les deux premières étapes de l'outil TEdenovo. Concernant la détection des répétitions, BLASTER reste à l'heure actuelle la meilleure alternative. Concernant le regroupement de ces répétitions, une piste prometteuse consisterait à unifier les algorithmes de GROUPER et RECON. Mais pour être mené à bien, un tel travail méthodologique nécessite

de développer en parallèle un simulateur permettant de tester pertinemment le nouvel algorithme.

3.2.2 *Intégration des approches basées sur la structure*

L'approche *de novo* canonique est souvent supposée ne pouvoir détecter que les familles d'ET ayant un grand nombre de copies, les auteurs entendant généralement par là la présence d'au moins 10 copies. Mais un tel argument n'est pas pertinent. En effet, pour que l'approche *de novo* canonique telle qu'implémentée dans l'outil TEdenovo construise une séquence consensus qui corresponde, sur toute sa longueur (+-5%), à la séquence de référence d'une famille d'Et, il faut que cette famille ait au moins trois copies dans le génome, c'est-à-dire à trois locus distincts, et que ces trois copies s'alignent, au moins de proche en proche, les unes avec les autres.

De plus, les trois copies doivent avoir au moins 90% d'identité les unes avec les autres mais n'ont pas forcément besoin d'être chacune complète ni en un seul fragment. En effet, le programme GROUPER permet de connecter les fragments pour obtenir la copie qui a transposé, et le programme RECON permet de regrouper des copies de taille différentes (+- 5%). De plus, la construction d'un consensus à partir d'un alignement multiple de copies ne prend pas en compte les insertions présentes seulement dans une seule copie, et peut combler une délétion si au moins deux copies ne l'ont pas. Malgré ces propriétés, une famille d'ET dont il ne reste que des fragments qui se chevauchent peu ou pas, ou qui n'a plus qu'une seule copie, ne pourra pas être reconstruite par notre approche *de novo*.

Dans le but de pallier à ce genre de cas, de nombreuses méthodes cherchant des motifs structuraux caractéristiques des ET peuvent être utilisées. Ces méthodes peuvent donc identifier une séquence d'Et n'ayant pas d'autres copies dans le génome, tant que celle-ci possède une structure reconnaissable.

La plupart de ces méthodes ont pour objectif de détecter dans une séquence génomique les copies complètes de rétrotransposons à LTR : LTR_STRUC (McCarthy et McDonald 2003), LTR_par (Kalyanaraman et Aluru 2006), LTR_Rho (Rho et coll. 2007), LTR_FINDER (Xu et Wang 2007) et LTRharvest (Ellinghaus et coll. 2008). L'article décrivant cette dernière l'a comparée à toutes les précédentes. Elles ont en commun de chercher des répétitions dégénérées pouvant potentiellement corres-

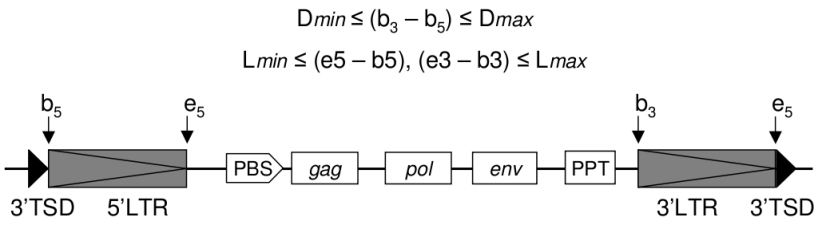


FIG. 9: Structure typique d'un rétrotransposon à LTR (Tiré de Ellinghaus et coll. 2008.)

pondre à des LTR (*long terminal repeats*), comme ayant une taille dans un intervalle donné et séparé par une distance dans un intervalle donné (figure 9). Certaines peuvent aussi chercher la présence d'autres caractéristiques de rétrotransposons à LTR, comme les TSDs (*target site duplications*), les ORFs *gag* et *pol*, le PBS (*protein binding site*), etc. Ces méthodes diffèrent surtout dans leur implémentation, LTRharvest étant la plus rapide et économe en mémoire grâce à son utilisation des algorithmes de Vmatch (ref). Cependant, les auteurs de cette méthode insistent sur le fait qu'elle ne peut détecter en premier lieu que les séquences d'éléments les plus récents, ceux qui possèdent les LTR les plus similaires. Ce n'est qu'ensuite que ces séquences peuvent être utilisées par d'autres outils pour détecter d'autres copies divergentes et fragmentées.

Il existe également des méthodes de détection basées sur la structure adaptées aux autres types d'ET (Bergman et Quesneville 2007, Lerat 2009). Ainsi, les méthodes visant à détecter les transposons à ADN et les MITEs recherchent des occurrence de TIR de la même façon que les méthodes précédentes recherchent des LTR. Même certains hélitrons, pourtant pauvres en caractéristiques structurales conservées, peuvent être détectés *via* des approches structurales, et une méthode sophistiquée a déjà été mise en place (STAN, Nicolas et coll. 2005). Cette méthode utilise pour leur détection un certain type de grammaire non contextuelle, les *string variable grammars*, afin de rechercher la présence de motifs particuliers, comme une épingle à cheveux (*hairpin*), dans une chaine de caractère (figure 10). Associée à un arbre de suffixe pour indexer le génome, une telle méthode est très rapide. En améliorant le motif recherché de manière itérative, cette méthode a aussi permis de mettre à jour les différentes combinaisons de motifs terminaux chez les Hélitrons dans le génome d'*A. thaliana* (Tempel et coll. 2007).

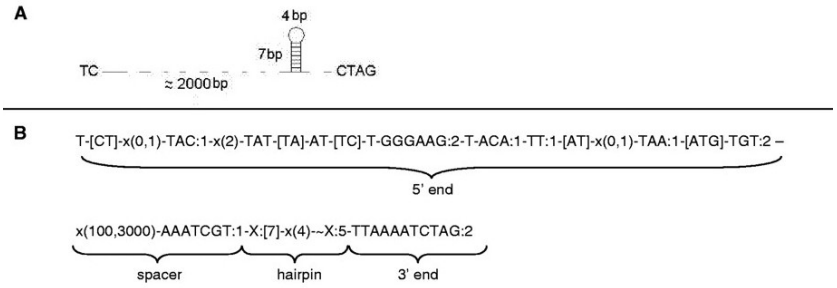


FIG. 10: Structure typique d'un Hélitron. **A**. Modèle d'AtREP3 de Kapitonov et Jurka (2001). **B**. Modèle affiné suite à l'utilisation itérative de STAN sur une banque de copies d'AtREP3. (Tiré de Nicolas et coll. 2005.)

Il serait intéressant d'intégrer ce type de méthodes dans le cadre de notre approche *de novo* et de nos outils. Pour un génome donné, nous commencerions par exécuter les trois premières étapes de l'outil TEdenovo, c'est-à-dire l'alignement du génome avec lui-même pour détecter les répétitions, suivi de leur regroupement par les programmes GROUPER, RECON et PILER, et enfin la construction d'un alignement multiple pour chaque groupe à partir duquel une séquence consensus est construite. Ensuite, nous pourrions exécuter le programme LTRharvest sur le génome étudié et récupérer la séquence des ET prédits comme étant des rétrotransposons à LTR. Il suffirait donc de construire des séquences consensus à partir des copies génomiques identifiées, d'ajouter celles-ci à la banque de consensus issue de l'outil TEdenovo, puis de classifier cette banque élargie à l'aide de l'outil TEclassifier présenté précédemment (chapitre 2), et enfin d'enlever la redondance en son sein, comme ce qui correspond actuellement aux étapes 4 et 5 de l'outil TEdenovo. Toute autre approche basée sur la structure des ET pourrait également être intégrée par ce moyen-là, démontrant encore l'utilité du TEclassifier pour la réconciliation des prédictions.

Notre analyse préliminaire montre que LTRharvest, lancé avec toutes les options et les paramètres par défaut, prédit 243 rétrotransposons à LTR dans la version 4 du génome de *D. melanogaster*, et 137 dans la version 9 du génome d'*A. thaliana*. Les séquences chez *D. melanogaster* ont une taille moyenne de 6622 pb (médiane à 7176 pb) et celles chez *A. thaliana* ont une taille moyenne de 5977 pb (médiane à 5186 pb). La qualité de ces séquences peut être estimée en essayant de les aligner avec les séquences consensus *knowledge-based* de rétrotransposons à LTR de leur génome respectif : 60 séquences parmi les 117 de la

banque *kb-BDGP* pour *D. melanogaster*, et 127 séquences parmi les 305 de la banque *kb-Repbase* pour *A. thaliana*. De plus, les séquences détectées par LTRharvest peuvent aussi être comparées aux séquences consensus construites par TEdenovo et classifiées comme étant des rétrotransposons à LTR : 258 séquences parmi 593 consensus *de novo* pour *D. melanogaster*, et 206 séquences parmi 1275 consensus *de novo* pour *A. thaliana*. Dans les tableaux 9 et 10, je rappelle également la comparaison de toutes les séquences consensus de TEdenovo avec les séquences consensus *knowledge-based* de leur génome respectif.

TAB. 9: Validation des copies détectées par LTRharvest et des séquences consensus obtenues avec TEdenovo dans le génome de *D. melanogaster*.

Comparaison	Sensibilité	Spécificité	Ratio des complets
TEdenovo versus tout kb-BDGP	92%	76%	53/68
LTRharvest versus LTR de kb-BDGP	70%	78%	15/60
LTRharvest versus LTR de TEdenovo	54%	93%	21/258

TAB. 10: Validation des copies détectées par LTRharvest et des séquences consensus obtenues avec TEdenovo dans le génome d'*A. thaliana*.

Comparaison	Sensibilité	Spécificité	Ratio des complets
TEdenovo versus tout kb-BDGP	74%	67%	76/154
LTRharvest versus LTR de kb-BDGP	56%	61%	14/127
LTRharvest versus LTR de TEdenovo	82%	85%	18/206

Dans les deux génomes, la sensibilité et le ratio des complets entre les séquences de LTRharvest et les séquences consensus *knowledge-based* sont plus faibles qu'entre les séquences de TE-

denovo et les séquences consensus *knowledge-based*. Cependant, une telle comparaison est difficile à interpréter étant donné que LTRharvest se focalise sur les éléments complets et néglige les autres. Par ailleurs, dans leurs comparaisons avec les séquences consensus *knowledge-based*, il est intéressant de noter que la spécificité des prédictions de LTRharvest oscille entre 61% et 78%. Cela signifie que, dans ces deux génomes bien connus, entre 20% et 40% des prédictions de LTRharvest ne s'alignent avec aucune séquence consensus *knowledge-based*. Seraient-ce des faux positifs comme proposé récemment (Lerat 2009) ? Il possible d'imaginer que ces nouvelles prédictions soient en fait de vrais éléments à LTR qui ne fassent pas partie de ce que cet auteur a considéré comme étant l'annotation de référence. Cette interpétation est renforcée par le fait que la spécificité des prédictions de LTRharvest augmente lorsqu'on les compare avec les consensus *de novo* de TEdenovo.

De plus, la comparaison des séquences prédites par LTRharvest avec les consensus *de novo* classifiés comme rétrotransposons à LTR de l'outil TEdenovo montre que certaines prédictions de LTRharvest ne s'alignent avec aucun consensus *de novo*, et réciproquement. Cela suggère qu'il serait intéressant de combiner les prédictions de LTRharvest avec les séquences consensus de l'outil TEdenovo, par exemple avant d'exécuter le TEclassifier. Bien sûr, de nombreux tests sont encore à effectuer afin de s'assurer de la pertinence des prédictions ainsi qu'afin d'optimiser la façon dont LTRharvest est exécuté par TEdenovo.

Les résultats préliminaires d'identification des ET par des approches basées sur la structure sont encourageants. Ils ouvrent la voie à une intégration de ces algorithmes dans l'outil TEdenovo. Mais à côté de ces pistes pour améliorer sensibilité et spécificité des résultats, des axes d'amélioration sont encore nécessaires en ce qui concerne l'application de l'outil aux génomes très répétés.

3.2.3 *Stratégies pour l'analyse des grands génomes*

En parallèle des pistes décrites ci-dessus visant à améliorer la pertinence biologique des résultats de l'outil TEdenovo, il est également nécessaire d'améliorer les performances techniques, notamment le temps de calcul requis pour analyser un génome, ce dans la perspective d'avoir un jour prochain à annoter des génomes de très grande taille et très riches en ET tels que ceux

du blé tendre (17 Gb) et de l'orge (5 Gb). Il n'est pas toujours évident d'estimer le temps d'exécution d'outils qui distribuent une partie de leurs tâches en parallèle. En effet, le temps total d'une étape peut dépendre de l'activité des autres utilisateurs avec lesquels la grappe de calcul est partagée. Pour avoir une idée cependant de l'augmentation du temps d'analyse en fonction de la taille du génome, j'ai comparé le temps d'exécution de la première étape de l'outil TEdenovo lancé en parallèle, et ceci sur plusieurs génomes de taille croissante (tableau 11). Ces analyses ont été réalisées sur une grappe de calcul HP possédant 128 nœuds de quatre cœurs cadencés à 3 GHz. Le temps de calcul pouvant dépendre de l'encombrement de la grappe, les chiffres ci-dessous correspondent à la somme des temps de chaque job (un job étant un calcul lancé en parallèle). Si la durée cumulée des jobs est d'une heure et qu'il y a 60 jobs, l'étape parallélisée ne durera que quelques minutes. Dans le cas de la première étape de l'outil TEdenovo, un job compare un sous-ensemble du génome avec le génome entier *via* le programme BLASTER.

TAB. 11: Temps de calcul de la première étape de l'outil TEdenovo en fonction de la taille du génome.

Organisme	Taille du génome (en Mb)	Temps cumulés des jobs	Proportion du génome en répétitions
L. maculans	45	70 min	35%
A. thaliana	120	3 h	13.5%
D. melanogaster	130	2 h 45 min	7.4%
A. lyrata	207	16 h	32.6%
B. distachyon	270	19 h	30.2%
O. sativa	372	64 h	33.5%

Ces résultats confirment l'intuition selon laquelle le temps de calcul nécessaire pour comparer un génome avec lui-même augmente avec la taille du génome. Il apparaît qu'en l'état actuel des choses, un génome tel que celui du maïs (2.5 milliards de paires de bases) ne peut être analysé tel quel par les outils TEdenovo et TEannot. La limite supérieure actuelle concerne des génomes d'au plus 500 millions de paires de bases, par exemple le riz (390 millions de paires de bases) et la vigne (490 millions de paires de bases). De plus, les résultats obtenus sur les dif-

férents génomes analysés par l'outil TEdenovo (3.1.1) ont aussi montré que le nombre de groupes construits lors de l'étape 2 augmente avec le contenu en ET du génome. Ceci indique que le temps de calcul peut substantiellement varier pour des génomes de taille équivalente, selon leur contenu en ET et donc selon leur histoire évolutive respective. En effet, le temps de calcul de l'outil TEdenovo dans son ensemble dépend beaucoup du nombre de groupes construits à l'étape 2, ceci reflétant le nombre de familles d'ET ainsi que leur diversification.

Il existe cependant plusieurs possibilités permettant de tenter l'analyse de grands génomes. J'ai déjà abordé certaines pistes précédemment en évoquant notamment la mise au point d'un algorithme de regroupement unifiant les programmes actuels GROUPER et RECON (3.2.1). Nous avons vu que cela éviterait les groupes redondants entre les deux méthodes, diminuant le nombre d'alignements multiples à construire et donc le nombre de consensus à classifier. Cependant, comme décrit précédemment, ce travail délicat n'est pas encore entamé car nous manquons encore de connaissances sur la diversification des familles d'ET.

Concernant l'outil TEannot, une possibilité serait de mettre des valeurs par défaut pour l'étape de filtre. En effet, à l'heure actuelle, trois méthodes sont utilisées, BLASTER, REPEATMASKER et CENSOR, pour aligner les séquences d'ET avec le génome, et ceci se répète après avoir mélangé aléatoirement les nucléotides du génome. Seuls sont gardés les alignements sur le vrai génome ayant un score plus élevé que le quantile à 95% des scores des meilleurs alignements sur le génome remanié. Les résultats actuels (tableau 12) montrent que ces valeurs seuils sont relativement similaires d'un génome à l'autre, à part pour BLASTER. L'étape d'alignement sur le génome remanié pourrait donc être évitée dans certains cas, résultant en un gain de temps significatif. Cependant, là aussi, beaucoup de facteurs sont en jeu, tels la composition nucléotidique du génome et celle de la banque d'ET, et il faudrait s'assurer qu'utiliser des valeurs par défaut soit adapté dans une majorité de cas.

Enfin, en plus de ces pistes, nous pouvons envisager une alternative peut-être plus originale. Dans les grands génomes, il est fréquent de mettre à jour des régions génomiques sur lesquels plusieurs ET sont insérés les uns dans les autres. Une étude pionnière chez le maïs en montre un bon exemple (figure 11), une étude poussée de ces réseaux d'insertions a été menée chez

TAB. 12: Quantile à 95% des scores maximaux des *matches* entre la banque *de novo* d'ET et le génome randomisé.

Organisme	BLASTER	REPEATMASKER	CENSOR
A. thaliana	102	89	233
D. melanogaster	55	82	214
A. lyrata	157	103	243
B. distachyon	0	94	226

D. melanogaster (Bergman et coll. 2006), et récemment l'analyse de 13 contigs de plusieurs mégabases sur le chromosome 3B du blé a bien illustré ce type de structuration (Choulet et coll. 2010).

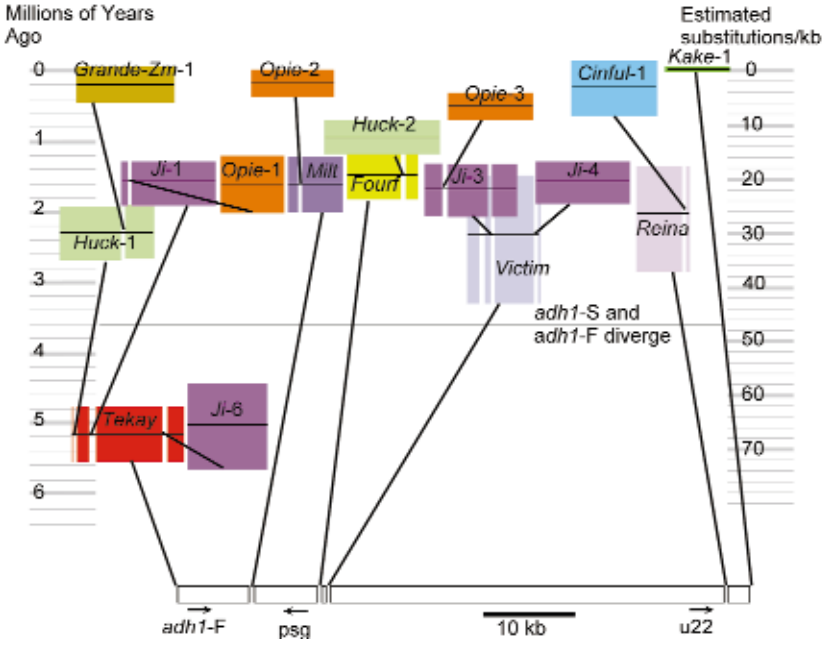


FIG. 11: Dynamique temporelle des rétrotransposons dans la région adh1-F du maïs. Les boîtes colorées représentent des rétrotransposons, les coupures au sein des boîtes représentent les insertions, les lignes horizontales au milieu des boîtes correspondent aux dates d'insertions, la hauteur des boîtes indiquant l'écart-type de l'estimation (Tiré de SanMiguel et coll. 1998.)

En commençant par détecter les copies les plus récentes, celles qui sont le plus conservées, il est alors possible de les exciser du génome pour rétablir la continuité des copies qui étaient interrompues par l'insertion de celles-ci. Il devient alors plus facile d'identifier ces copies plus anciennes par un deuxième passage

de l'outil TEdenovo sur le génome excisé. La première étape ne se concentrant que sur les copies très récentes peut être optimisée avec des paramètres stringents. Il devrait alors être possible d'annoter ainsi un génome de grande taille de manière itérative. Les paramètres utilisés seraient de moins en moins contraignants à chaque itération, la taille du génome diminuant au rythme des copies d'ET détectés.

Appliquons cette idée au génome de *D. melanogaster*. Lors de la première itération, BLASTER est utilisé à la première étape de l'outil TEdenovo, ne conservant que les répétitions ayant entre elles au moins 94% d'identité, au lieu des 90% généralement utilisés. Puis PILER est utilisé et, au lieu de 120 groupes, seuls 109 sont construits, à partir desquels des séquences consensus sont établies. Une fois celles-ci classifiées et la redondance enlevée, il reste 89 séquences consensus, parmi lesquelles seules 41 sont utilisées par l'outil TEannot, les autres étant filtrées car n'étant pas classifiées ou étant classifiées comme étant des ET incomplets. Ensuite, l'outil TEannot peut être lancé, mais seulement avec le programme REPEATMASKER. Le filtre statistique empirique est appliqué mais l'annotation des micro-satellites est évitée. Une fois l'annotation obtenue, seules les copies ayant une identité supérieure à 90% sont conservées et ce sont elles qui sont excisées du génome, résultant en une diminution de la taille du génome de 6%. Ce chiffre paraît faible mais il est à comparer avec la couverture du génome obtenue juste après la première étape de TEdenovo, couverture de 7.4% dans le cas de *D. melanogaster*. La majeure partie des copies récentes ont donc été retirées du génome.

Bien sûr, de tels résultats sont très préliminaires et il reste à voir qu'elle sera la valeur d'une telle appropche sur l'analyse de très grands génomes. A chaque étape, les valeurs utilisées pour les paramètres doivent être testées et l'analyse globale des deux génomes de référence doit être comparée, entre l'approche normale du chapitre 2 et cette approche itérative. Mais l'idée de d'abord détecter les copies les plus récentes puis de les exciser avant de relancer l'analyse mérite d'être testée plus en avant. En effet, il ne serait peut-être plus indispensable, à la première itération, de détecter les répétitions de manière exhaustive. Par exemple, afin d'accélérer encore plus cette procédure, des algorithmes autres que ceux de BLASTER, basés par exemple sur des arbres ou des tableaux de suffixes, pourraient être utilisés dans un premier temps afin d'identifier les répétitions les plus conservées. Une fois la première itération achevée, la deuxième

itération se concentrant sur les répétitions plus divergentes et le génome étant plus petit, il serait possible d'utiliser BLASTER à nouveau afin d'être plus sensible.

L'analyse du contenu en ET des grands génomes nécessite la mise en place de stratégies particulières. L'une d'elle consisterait à appliquer les outils TEdenovo et TEannot de façon itérative, retirant progressivement les copies du génome, d'abord les plus récentes et donc les mieux conservées, pour tenter d'exhumer ensuite les copies les plus divergentes.

3.2.4 *Mise en place d'outils ergonomiques de curation manuelle*

Toute recherche scientifique nécessite, à un moment ou à un autre, de pouvoir visualiser les objets étudiés sous une forme ou une autre afin de mieux les appréhender, ou tout simplement dans le but de vérifier les résultats souvent obtenus par des techniques indirectes. Dans le contexte de l'annotation des génomes, le chercheur est confronté à une séquence d'ADN pouvant atteindre plusieurs milliards de paires de bases. Le défi consistant à détecter et identifier les portions fonctionnelles débute donc toujours par une phase d'analyse informatique, celle-ci se focalisant principalement sur les segments du génome codant a priori pour une protéine. Cependant, malgré les recherches actives depuis plusieurs décennies dans ce domaine, la mise au point de modèles sophistiqués (Foissac et coll. 2008) et l'abondance de données permettant de s'aider d'un deuxième génome, voire plus, pour annoter le premier (Lin et coll. 2008), une phase de curation manuelle reste toujours indispensable et le restera probablement, étant donné que les techniques de séquençage avancent vers du très haut débit et que la qualité des assemblages peine à suivre.

Généralement seuls les gènes codants sont examinés à la main, mais cela revient tout de même à plusieurs milliers ou dizaines de milliers d'entités. En pratique, dans le cadre d'efforts communs d'annotation, le consortium international responsable du projet répartit la tâche de telle sorte qu'un laboratoire spécialiste d'une certaines famille de gènes soit chargé de corriger l'annotation réalisée automatiquement. A cette fin, il est indispensable de disposer d'outils permettant de réaliser la curation manuelle, ceux-ci devant être ergonomiques et disponibles *via* internet, et devant permettre la synchronisation des données, leur stockage

pérenne et leur confidentialité tant que le résultat des analyses n'est pas publié.

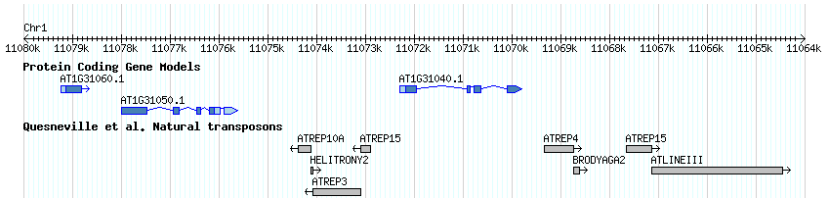


FIG. 12: Annotations du génome d'*A. thaliana* visualisables avec Gbrowse sur le site de TAIR.

Devant l'importance de la tâche, de tels outils ont été développés et continuent d'être améliorés. Parmi ceux-ci, Gbrowse est utilisé couramment comme visualisateur d'annotations. Il ne permet pas de les corriger mais seulement de les observer. Etant particulièrement ergonomique, il est utilisé notamment par le site web de référence pour le génome de *D. melanogaster* et celui d'*A. thaliana*. Apollo quant à lui est un éditeur d'annotations mis au point initialement pour le génome de *D. melanogaster* et utilisé maintenant dans les projets d'annotations de nombreux autres génomes. Cet outil peut être relié à une base de données, ce qui permet à des chercheurs dispersés dans plusieurs pays de corriger le même génome de manière simple et efficace. Ces deux outils font maintenant parti de la même suite logicielle *GMOD* utilisée pour stocker et visualiser des annotations génomiques (`http://gmod.org/`).

La biologie n'est pas la seule discipline à générer une quantité considérable de données mais ces-dernières présentent la particularité d'être très hétérogènes. Le génome, de par sa position clé dans le développement d'un organisme et l'évolution d'une population, représente naturellement le point focal vers lequel les observations expérimentales de biologie moléculaire et cellulaire se rattachent : étude de mutants, diversité génétique, expression spatio-temporelle des gènes, sites de liaison protéines-ADN, etc. De plus, chaque région génomique peut être le sujet de publications, parfois anciennes, qui contiennent des informations importantes et nécessitent donc d'être reliées à cette région donnée. La génomique est donc reconnue comme étant un champ scientifique nécessitant des outils d'intégration de données performants et pertinents.

Dans le cas qui nous concerne, l'annotation des ET, il est possible d'utiliser les outils existants. Le résultat final de l'annotation des ET est d'ailleurs toujours visible sur les visualisateur d'annotations tel Gbrowse, mais ceux-ci ne mettent à disposition que les locus auxquels les ET peuvent être trouvés, et omettent de préciser à quelle partie d'un Et les copies correspondent (signaux terminaux, cadres ouverts de lecture, motifs de régulation). De plus, les outils d'édition d'annotation comme Apollo ne permettent pas seuls d'annoter les ET de manière optimale. En effet, s'il est encore possible de vérifier manuellement les annotations automatiques d'ET dans un génome de faible taille comme *D. melanogaster*, cela devient vite impossible dans de grands génomes répétés car il y a souvent bien plus de copies d'ET que de gènes. De plus, nous avons vu précédemment (chapitre 2) que la qualité d'une annotation *de novo* dépend surtout de la banque de séquences consensus utilisées. C'est donc cette banque qu'il faut examiner en priorité.

Ainsi, dans un premier temps, l'outil TEdenovo peut être utilisé pour construire une banque de séquences consensus représentant les différentes familles d'ET présentes dans le génome étudié ainsi que leur variants structuraux. Le but de la curation manuelle telle qu'expliquée précédemment (3.1.3) est de retirer de la banque de consensus les séquences encore redondantes entre elles ainsi que les séquences chimériques. En pratique, cela correspond aux séquences qui n'ont pu être retirées automatiquement par l'outil TEclassifier ou bien pour lesquelles on dispose de caractéristiques contradictoires. Afin de nettoyer cette banque de consensus, l'annotateur a besoin de naviguer simplement entre plusieurs types d'informations, la séquence consensus faisant le lien entre eux. Avant de prendre une décision, il faut donc connaître, pour chaque consensus, le nombre de séquences génomiques qu'il représente (étape 1 de l'outil TEdenovo), la méthode de regroupement grâce à laquelle il a été construit (étape 2 de l'outil TEdenovo) et ses caractéristiques d'ET (détectées par le TEclassifier).

Il est également nécessaire de connaître les relations des séquences consensus les unes avec les autres. En effet, plusieurs consensus correspondent souvent à une seule et même famille d'ET, pour laquelle ils représentent différents variants structuraux. Il est inutile pour la phase d'annotation des copies avec l'outil TEannot de garder tous les consensus d'une même famille, notamment ceux correspondant à des versions tronquées de la séquence complète. Ainsi, nous commençons d'abord par

regrouper les consensus appartenant à une même famille, par exemple avec l'outil BLASTCLUST (Dondoshansky 2002), puis nous construisons, pour chaque groupe de consensus, un alignement multiple. En plus des informations listées dans le paragraphe précédent, le curateur a maintenant aussi besoin de visualiser des alignements multiples, voire des phylogénies, avant de prendre une décision de garder ou non tel ou tel consensus dans la banque.

Afin de permettre la curation manuelle d'une banque de consensus d'ET de manière distribuée entre plusieurs curateurs et étant donné qu'une telle tâche nécessite de combiner en une même interface des données hétérogènes, il devient de plus en plus nécessaire de mettre en place des outils ergonomiques spécifiques des ET. A ce sujet, la technologie wiki, sites web dont les pages sont modifiables par les utilisateurs, semble parfaitement correspondre aux besoins et est déjà utilisée pour la curation de l'annotation fonctionnelle des gènes codants (Huss et coll. 2008). Cette technologie permet également de conserver l'historique précis des modifications effectuées ainsi que leur justification, facilitant grandement le transfert d'expertise entre curateurs. De telles interfaces seraient donc très bénéfiques à l'annotation des ET dans les génomes séquencés et constitueraient un premier répertoire de connaissance concernant la diversification des familles d'ET dans différents génomes.

L'annotation du contenu en ET d'un génome, quel qu'il soit, s'accompagne toujours d'une phase de curation manuelle. Afin de faciliter celle-ci et de tirer profit des diverses informations réunies par les outils TEdenovo et TEannot, il serait judicieux de développer de nouvelles interfaces, pour une fois spécifiques des ET.

3.3 PERSPECTIVES FONDAMENTALES ET APPLIQUÉES

L'INRA se présente comme un organisme de recherche finalisée (Houllier 2009). Pour reprendre les mots de Michel Sebillotte (colloque 2007) :

> *La recherche finalisée met en relation la production des connaissances et les problèmes des sociétés. Le but de la recherche finalisée est de comprendre pour pouvoir mieux agir dans un monde en constante évolution. En tant qu'institution de recherche finalisée et chercheurs de cette institution, nous avons la responsabilité, non seulement*

> *de produire des connaissances, mais aussi de mettre au point des méthodes et d'inventer des techniques qui seront autant indispensables à l'action que la production des connaissances proprement dites.*

Etant particulièrement intéressé par les interactions entre science et société et ayant réalisé mon doctorat à l'INRA, au sein du département *Génétique et Amélioration des Plantes*, dans le cadre d'une cotutelle entre un laboratoire de bioinformatique (URGI) et un laboratoire d'analyse génomique d'une plante d'intérêt agronomique (GDEC), j'ai voulu réfléchir à la façon dont mes travaux pouvaient s'intégrer dans une démarche de recherche finalisée. Parmi les nombreux thèmes possibles, j'en ai choisi deux très différents l'un de l'autre mais qui, je l'espère, représentent bien la diversité des implications de l'annotation des ET dans les génomes pour des études plus appliquées.

3.3.1 *Liens entre éléments transposables et amélioration variétale*

L'agriculture est née avec la domestication de certaines espèces selon un processus de sélection artificielle réalisé inconsciemment à l'époque (Diamond 2002). Comme l'expliquait Darwin (Darwin 1992) :

> *La nature fournit les variations successives, l'homme les accumule dans certaines directions qui lui sont utiles. Ce procédé consiste à cultiver toujours les meilleures variétés connues, à en semer les graines et, quand une variété un peu meilleure vient à se produire, à la cultiver préférablement à tout autre.*

Et c'est en s'inspirant de ce procédé que Darwin a construit sa théorie de la sélection naturelle afin d'expliquer l'origine des espèces ainsi que leur évolution.

Pendant longtemps, les paysans se sont basés sur des mesures agro-morphologiques pour sélectionner les meilleures variétés. Cependant, avec l'avènement de la génétique et de la biologie moléculaire au vingtième siècle, les techniques se sont considérablement améliorées et le développement des marqueurs moléculaires issus des séquences a permis de réaliser une sélection indirecte par association. Le principe repose sur le fait que le phénotype d'intérêt est dû, pour partie, à une ou plusieurs composantes génétiques et que la sélection se fait au bénéfice des individus qui possèdent la composante (allèle) recherchée. Pour cela, si l'on identifie que tous les individus possédant la caracté-

ristique recherchée montrent également la présence d'un même allèle pour un marqueur (généralement une séquence d'ADN située à proximité physique de la composante), alors de manière résumée, le procédé consiste à sélectionner les individus sur la base de la présence de l'allèle du marqueur en question plutôt que sur la caractéristique elle-même, souvent beaucoup plus complexe et longue à analyser. Ce procédé qui consiste à faire de la sélection assistée par marqueurs (SAM), permet donc de raccourcir le travail de sélection et de le rendre plus précis dans le cas où les marqueurs sont très proches du caractère ciblé.

Différent types de marqueurs peuvent être utilisés, les meilleurs présentant plusieurs avantages, par exemple une forte proximité avec le locus cible, un haut niveau de polymorphisme et une simplicité de détection. Le développement des marqueurs basé sur la détection de fragments d'ADN et l'amélioration des techniques de séquençage suggèrent que la sélection assistée par marqueurs va devenir de plus en plus importante dans les programmes d'amélioration variétale (Collard et Mackill 2008). Elle est déjà en place sur des espèces à fort potentiel économique comme le maïs et a motivé son programme de séquençage soutenu au départ par les producteurs eux-mêmes afin d'accélérer les programmes de sélection.

Cependant, la réussite d'un tel procédé dépend, entre autres, de la quantité de marqueurs disponibles dans la lignée d'intérêt ainsi que de leur distribution le long des chromosomes. A ce titre, les ET sont des marqueurs de choix, notamment dans les espèces en possédant beaucoup, tel le blé. Ainsi, l'équipe dans laquelle j'ai effectué ma thèse en cotutelle à l'INRA de Clermont-Ferrand a mis au point un type de marqueurs basé sur les jonctions entre ET et séquences flanquantes (Paux et coll. 2006) : les marqueurs ISBP (*insertion site-based polymorphism*).

Dans un premier temps, les ISBP ont été détectés par annotation manuelle des séquences d'extrémités de clones BAC de la banque spécifique du chromosome 3B de la variété de référence *Chinese Spring* du blé tendre. En repérant les jonctions entre séquences d'ET et séquences flanquantes grâce à des comparaisons avec la banque de référence d'ET des génomes de blé et d'orge (TREP, *the Triticeae repeat sequence database*), l'équipe de Clermont-Ferrand a ainsi pu définir environ 2500 marqueurs ISBP sur le chromosome 3B. Ces marqueurs ont permis de lier la carte physique du chromosome aux cartes génétiques et ont été validés quant à leur intérêt pour la sélection sur du matériel de sélectionneurs français et australien (Paux et coll. 2010). Un marqueur

ISBP est déjà utilisé en routine pour la SAM d'un gène de résistance à une rouille noire (Sr2, McNeil et coll. 2008). Plus récemment, l'annotation de séquences de plusieurs mégabases du chromosome 3B (Choulet et coll. 2010) a permis de montrer le potentiel énorme de la séquence complète du génome de blé pour le développement des ISBP et leur utilisation pour saturer le génome. Ainsi, l'annotation de 3 Mb d'un des contigs, a permis de developper 1000 marqueurs potentiels et plusieurs centaines ont été développés pour ce locus précis. L'extrapolation des données indique un potentiel de presque 4 millions de marqueurs ISBP dans le génome du blé, avec une densité d'un marqueur environ tous les 3 kb. Les ISBP peuvent être définis soit à partir des séquences d'extrémités de clones BAC comme cité précédemment, soit à partir de *sequence survey* de chromosomes par la technique Roche 454 qui permet des lectures de plus de 400 bp à l'heure actuelle. Cependant, l'efficacité du design des marqueurs ISBP est grandement améliorée lorsque l'on a accès à la séquence entière de la région génomique en question et à une annotation optimale des ET. Un outil de design automatique a été mis au point récemment dans l'équipe de Clermont-Ferrand (ISBPFinder, Paux et coll. 2010). Il permet d'identifier les jonctions et de définir les amorces pour les réactions PCR automatiquement et sur la base de critères optimisés. Un groupe américain a repris la même idée en développant des marqueurs similaires et ont mis en ligne un outil de design d'amorces pour ces marqueurs (RJPrimers, You et coll. 2010).

Dans un avenir proche, des séquences du génome du blé seront disponibles, notamment concernant son plus long chromosome, le 3B. Celui-ci faisant 1 Gb et étant composé d'au moins 75% d'ET, disposer d'outils performants tels TEdenovo et TEannot devient indispensable pour réaliser une annotation rapidement, première étape avant la construction de marqueurs ISBP. Le premier point visible de cet effort est l'intégration de l'outil TEannot dans l'outil d'annotation automatique TriAnnot développé à l'INRA de Clermont-Ferrand en collaboration avec l'URGI (`https://gpi.versailles.inra.fr/triannot`) pour les séquences génomiques de Triticeae.

D'autres applications directes concernent la lutte contre les ravageurs des cultures. Les modèles de co-évolution entre hôtes et parasites montrent, dans certaines conditions, qu'un parasite peut subir une pression de sélection très forte l'amenant à se spécialiser sur le génotype hôte le plus fréquent autour de lui. A l'inverse, l'hôte subit une pression de sélection diversifiante.

De plus, ces interactions sont dynamiques dans le temps et l'espace, générant ce que certains appellent une mosaïque géographique (Thompson 2005). Dans le contexte actuel où l'agriculture conventionnelle tente de lutter contre les ravageurs en diminuant l'usage de pesticides, la compréhension des mécanismes d'émergence des résistances et de suppression des gènes d'avirulence est fondamentale. Plusieurs études ont montré que les gènes d'avirulence sont fréquemment localisés dans des régions à forte recombinaison, souvent proches des télomères, et à proximité d'ET. Ceux-ci peuvent représenter une source de variabilité importante, tant au niveau structural *via* des insertions/délétions et comme déclencheur de réarrangements, qu'au niveau fonctionnel en interférant avec l'expression des gènes voisins, ce qui a comme conséquence de permettre au pathogène d'échapper à la vigilance des systèmes de défense de la plante (Kang et coll. 2001, Khang et coll. 2008). Quant à la résistance aux pesticides, plusieurs études ont montré que *D. melanogaster* pouvait développer une résistance au DDT grâce à des mutations causées par des ET (Aminetzach et coll. 2005, Schmidt et coll. 2010).

L'annotation des ET est le pré-requis de la mise au point des marqueurs ISBP, ceux-ci ayant une importance considérable dans la mise au point de cartes génétiques des génomes très répétés. L'étude de la dynamique des ET, quant à elle, peut permettre de mieux comprendre comment certains champignons ou autres dévoreurs des cultures échappent aussi efficacement aux défenses des plantes. En plus de ces aspects, mon travail me permet également de proposer une innovation en terme de démarche de recherche sur la dynamique des ET.

3.3.2 *Place de la simulation dans l'étude des éléments transposables et l'évolution des génomes*

Depuis Aristote, la biologie est en prise avec le concept de causalité. C'est notamment ce qui a amené Mayr à distinguer, à tort ou à raison, la biologie fonctionnelle de la biologie évolutive (Mayr 1961). Pour lui, un biologiste fonctionnel sera intéressé par le *comment*, décrira les mécanismes en oeuvre dans les organismes vivant et cherchera les causes premières, alors qu'un biologiste évolutif s'intéressera au *pourquoi*, décrira les processus historiques de sélection naturelle et cherchera les causes ultimes. Bien sûr, ces deux aspects sont liés, mais dans la démarche concrète et quotidienne de recherche, certaines différences fon-

damentales se font ressentir. Dans mon cas, cela s'est matérialisé par une anecdote récurrente, la difficulté à faire comprendre à mes proches que je ne faisais ni expériences ni modélisation, mais plutôt de l'analyse de données, lesquelles sont, de surcroît, issues d'un processus évolutif s'étalant sur des millions d'années, au bas mot. Bien qu'il soit possible de montrer expérimentalement que la diminution drastique de la méthylation de l'ADN chez *A. thaliana* permet aux ET de transposer (Tsukahara et coll. 2009), il serait absurde d'imaginer *rejouer* expérimentalement la dynamique des ET dans cette espèce. Il est bien sûr possible de faire de la sélection artificielle, mais seul le temps de génération court des microorganismes autorise l'évolution expérimentale en condition contrôlée (Barrick et coll. 2009).

Ainsi, ce n'est a priori pas par le biais d'approches expérimentales que je pourrai directement répondre aux questions soulevées précédemment à propos des variants structuraux. Par contre, celles-ci, de part la description des mécanismes moléculaires régissant la dynamique des ET, apportent des connaissances indispensables au cadre conceptuel des recherches sur les variants. Mais pour aller plus loin, une autre voie est possible, et ceci grâce à la puissance des ordinateurs actuels, la simulation, c'est-à-dire la modélisation informatique du système étudié et l'analyse de son évolution *in silico*. Ce type d'approches n'est pas nouvelle en biologie, les modèles mathématiques de génétique des populations ayant très tôt été simulés sur ordinateur, notamment concernant les ET (Charlesworth et Charlesworth 1983). Ces modèles visaient à calculer la fréquence d'un ET dans une population, la proportion de sites polymorphes, les conditions d'invasion du génome, d'équilibre ou de disparition (Quesneville et Anxolabéhère 1998, Le Rouzic et Capy 2005, Le Rouzic et coll. 2007). Mais dans ces simulations, la structure des ET n'est pas explicitement représentée, tout au plus les copies sont caractérisées par un potentiel d'autonomie plus ou moins important.

C'est en étudiant les conditions d'émergence des transposons à ADN à partir d'un gène d'endonucléase (Quesneville et Anxolabéhère 2001) que la structure des ET a été explicitement représentée pour la première fois : répétitions terminales et transposase faite d'un domaine de liaison à l'ADN et d'un domaine catalytique. Cependant, ces simulations ne représentent l'évolution des génomes que vue de la part des ET, le reste du génome étant occulté. De plus, le devenir des copies, et plus précisément l'émergence de variants structuraux par l'intermédiaire de délétions internes, n'a pas été analysé en détail. D'autres travaux,

concernant des génomes circulaires de bactéries, ont été menés, cette fois l'ensemble du génome étant représenté et les individus étant sélectionnés sur leur capacité à produire des protéines leur permettant d'optimiser une fonction donnée (Knibbe et coll. 2007). La question était ici de comprendre le lien entre l'architecture du génome et la relation génotype-phénotype, il était donc indispensable de représenter l'ensemble du génome, ceci ayant aboutit à montrer que le taux de mutation conditionne la quantité de séquences non-codantes. Mais cette fois-ci ce sont les ET qui sont absent des simulations. Et dans un sens, à juste titre, car les paramètres indispensables à leur intégration dans des simulations *génome-complets* ne sont pas connus. Typiquement, on ne sait pas à l'heure actuelle déterminer précisément l'intervalle de valeurs adéquat du taux de délétion ni de la taille des délétions, bien que celles-ci soient un déterminant majeur de la persistance des ET dans les génomes (Petrov 2002).

Ainsi, dans ce contexte, il peut sembler nécessaire de commencer par analyser les délétions dans les copies d'ET dans les génomes réels, avant de songer à améliorer les simulations. Dans cette démarche, la simulation représente en quelque sorte un but ultime vers lequel tendre. Une simulation, tout d'abord qualitative puis quantitative, permettant de retrouver les grandeurs et dynamiques caractéristiques de l'évolution de la structure des génomes, serait en quelque sorte un indice fort montrant que les déterminants majeurs de cette évolution sont compris. Par exemple, disposer d'un tel outil poserait les premières bases dans la construction d'idéotypes à partir de données impliquant tous les composants des génomes, l'un des axes de recherche du département *Génétique et Amélioration des Plantes* de l'INRA.

Cependant, et c'est là que le titre de ce paragraphe prend tout son sens, la simulation ne devrait pas être uniquement vue comme une fin en soi, un but vers lequel tendre. Dans d'autres disciplines, la simulation accompagne les chercheurs au quotidien dans leurs tentatives de décrypter les phénomènes naturels, que ce soit à l'aide d'expériences ou à travers l'analyses de données historiques. Par exemple, les techniques actuelles de microscopie à fluorescence permettent de suivre en temps réel l'expression d'un gène dans chaque bactérie d'une colonie (Elowitz et Leibler 2000). Le but ici est d'analyser quantitativement les oscillations dues à la quantité variable de protéines au sein de chaque cellule et ce d'une génération à l'autre. Nous sommes bien loin de comprendre l'ensemble des phénomènes en jeu tels qu'observés sur les films montrant les bactéries fluorescentes. Et

pourtant il est nécessaire d'en extraire automatiquement l'intensité de la fluorescence entre cellules dans l'espace et le temps. Afin de mettre au point un algorithme remplissant cette tâche, il est utile de simuler le phénomène observé, même s'il n'est pas entièrement compris (Sbalzarini et Koumoutsakos 2005). L'intérêt principal réside dans le fait que ces simulations permettent de tester efficacement l'algorithme. Si celui-ci ne parvient pas à correctement manipuler les résultats de simulation, il semble ambitieux, pour ne pas dire vain, de vouloir l'utiliser directement sur des données réelles.

Dans cette logique il semble donc important, en plus des autres pistes évoquées, de mettre au point un outil de simulation permettant de tester indépendamment la première et la deuxième étape de l'outil TEdenovo, c'est-à-dire la détection des répétitions par comparaison du génome avec lui-même, et le regroupement de ces répétitions. Concernant la première étape, des outils performants ont été développés pour simuler l'évolution moléculaire de séquences d'ADN incluant des modèles mathématiques sophistiqués de substitutions et d'indels (Cartwright 2005). Il ne serait donc pas difficile de choisir une séquence d'ET connue, la faire évoluer selon différents scénarios pour obtenir une collection de copies divergentes, replacer ces copies dans un génome aléatoire et tester les résultats de la comparaison d'un tel génome avec lui-même. Cet exemple que l'on peut aisément qualifier de simpliste peut très logiquement être étendu, petit à petit, à des cas plus compliqués, faisant intervenir plusieurs familles d'ET, d'autres types de répétitions, des motifs d'insertions plus ou moins compliqués de copies d'ET les unes dans les autres, etc. La clé est que l'on maîtrise et connait ce que l'on doit retrouver. De la même manière, afin de tester la deuxième étape, il est simple d'imaginer des listes de coordonnées génomiques reflétant différents cas de répétitions alignées les unes avec les autres, ainsi que, plus important, différents motifs de fragmentation des copies et donc de disposition d'indels. A la question, *comment distinguer de manière pertinente une copie tronquée d'un variant structural ?*, cette démarche offre un angle d'attaque vers une meilleure caractérisation des variants structuraux dans les génomes réels.

Alors que simuler un génome de manière la plus réaliste possible tout en incorporant des ET semble hors de portée à l'heure actuelle, les valeurs biologiques des paramètres n'étant pas connues et pouvant donc difficilement être initialisées, simuler les données nécessaires à l'outil TEdenovo semble bien plus

envisageable. Encore une fois, une telle démarche ne vise pas à étudier la distribution des indels dans de vraies copies d'ET trouvées dans des génomes réels. Elle a plutôt pour but de simuler des configurations d'alignements deux-à-deux de répétitions par exemple, données qui permettent de tester de manière sûre les algorithmes de regroupement. Une fois qu'un tel simulateur existe et que les algorithmes utilisés dans l'outil TEdenovo parviennent à reconstruire les scénarios évolutifs simulés, il devient possible de les appliquer à des génomes réels. De plus, les simulations utilisées dans la mise au point de l'algorithme permettent non seulement de le tester, mais aussi d'interpréter les configurations observées dans les génomes séquencés. La phase de simulation et la phase d'analyse de génomes réels sont donc complémentaires, la simulation faisant partie intégrante de la mise en place des algorithmes.

Concernant les perspectives à long terme sur la compréhension de la diversification des ET, je propose une nouvelle démarche de recherche basée sur une méthodologie visant à dépasser les limites actuelles, et ce grâce à la complémentarité entre simulations pragmatiques et design d'algorithmes.

3.4 CONCLUSIONS

Le nombre de génomes complets déjà séquencés ou en cours de l'être augmente de plus en plus vite grâce à l'amélioration des technologies de séquençage. Afin de pouvoir étudier l'impact des ET sur la structure, la fonction et l'évolution de ces génomes, il est indispensable de disposer d'annotations précises des ET. A cette fin, j'ai développé un outil performant, intégrant plusieurs méthodes *de novo*, et permettant d'identifier la majorité des familles d'ET présentes dans un génome séquencé. Concernant la validation de cet outil, j'ai également mis au point un protocole de test à partir des génomes des espèces modèles *D. melanogaster* et *A. thaliana*.

Disposer d'un tel outil a permis de commencer l'analyse de nombreux génomes appartenant à des espèces différentes et pour lesquelles on disposait de peu d'informations préalables concernant les ET. Ces travaux ont notamment abouti à la rédaction d'une feuille de route visant à détailler la marche à suivre pour annoter le contenu en ET d'un génome de la manière la plus efficace possible.

Cependant, la diversification des familles d'ET et l'apparition de variants structuraux sont des processus très mal connus. Je me suis donc attaché, dans la mise au point de l'outil décrit précédemment, à accorder une attention particulière à la détection de ces variants, notamment dans le but de préciser leur importance qualitative et quantitative. Mais la mise au point d'une approche *de novo* analysant le contenu des génomes est particulièrement difficile. En effet, sur le plan épistémologique, il s'agit en pratique de valider une approche *de novo* sans que l'on connaisse précisément la qualité des données de référence.

Faut-il donc se contenter de l'outil mis au point et tenter d'identifier les variants structuraux au cas par cas, dès qu'une annotation, même partielle, est obtenue ? Non, l'histoire nous montre qu'étudier les ET a souvent abouti à une meilleure compréhension de phénomènes biologiques plus larges. Là encore, on peut supposer qu'étudier en détail la diversification des ET et tenter de quantifier l'importance des variants structuraux, affinera notre connaissance de la dynamique des génomes. En pratique, cela consiste à repenser le design d'un algorithme dont le but serait de faciliter l'identification des variants structuraux.

La piste que je propose fait appel à la simulation, à comprendre ici dans un sens proche de ce qu'est le *test-driven development* pour la mise au point de logiciels. La simulation serait vue comme le complémentaire indispensable de la mise au point d'un algorithme visant à identifier les variants structuraux. Simuler dans cette optique est même la première étape et commence modestement, en construisant des scénarios évolutifs simples, puis en testant l'outil avec, améliorant son algorithme et complexifiant les simulations au fur et à mesure, puis en explorant les génomes réels, et ainsi de suite.

Comme l'a dit le physicien Richard Feynman, « *what I cannot create, I do not understand* ». Ainsi, c'est vraisemblablement en s'inspirant des pratiques d'ingéniérie logicielle et en les adaptant à un contexte de recherche que les algorithmes d'annotation *de novo* d'ET, et plus globalement les outils d'analyse des génomes, pourront être améliorés, dans le but de générer des connaissances fondamentales sur la diversification des ET, contribuant ainsi à l'étude de la dynamique des génomes.

P. Abad, J. Gouzy, J.-M. M. Aury, P. Castagnone-Sereno, E. G. J. G. Danchin, E. Deleury, L. Perfus-Barbeoch, V. Anthouard, F. Artiguenave, V. C. C. Blok, M.-C. C. Caillaud, P. M. M. Coutinho, C. Dasilva, F. De Luca, F. Deau, M. Esquibet, T. Flutre, J. V. V. Goldstone, N. Hamamouch, T. Hewezi, O. Jaillon, C. Jubin, P. Leonetti, M. Magliano, T. R. R. Maier, G. V. V. Markov, P. McVeigh, G. Pesole, J. Poulain, M. Robinson-Rechavi, E. Sallet, B. Ségurens, D. Steinbach, T. Tytgat, E. Ugarte, C. van Ghelder, P. Veronico, T. J. J. Baum, M. Blaxter, T. Bleve-Zacheo, E. L. L. Davis, J. J. J. Ewbank, B. Favery, E. Grenier, B. Henrissat, J. T. T. Jones, V. Laudet, A. G. G. Maule, H. Quesneville, M.-N. N. Rosso, T. Schiex, G. Smant, J. Weissenbach, et P. Wincker. 2008. Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. *Nature Biotechnology*, 26(8):909–915.

M. D. Adams, S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, R. A. George, S. E. Lewis, S. Richards, M. Ashburner, S. N. Henderson, G. G. Sutton, J. R. Wortman, M. D. Yandell, Q. Zhang, L. X. Chen, R. C. Brandon, Y.-H. C. Rogers, R. G. Blazej, M. Champe, B. D. Pfeiffer, K. H. Wan, C. Doyle, E. G. Baxter, G. Helt, C. R. Nelson, G. L. Gabor Miklos, J. F. Abril, A. Agbayani, H.-J. An, C. Andrews-Pfannkoch, D. Baldwin, R. M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E. M. Beasley, K. Y. Beeson, P. V. Benos, B. P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M. R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K. C. Burtis, D. A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J. M. Cherry, S. Cawley, C. Dahlke, L. B. Davenport, P. Davies, B. Pablos, A. Delcher, Z. Deng, A. D. Mays, I. Dew, S. M. Dietz, K. Dodson, L. E. Doup, M. Downes, S. Dugan-Rocha, B. C. Dunkov, P. Dunn, K. J. Durbin, C. C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A. E. Gabrielian, N. S. Garg, W. M. Gelbart, K. Glasser, A. Glodek, F. Gong, J. H. Gorrell, Z. Gu, P. Guan, M. Harris, N. L. Harris, D. Harvey, T. J. Heiman, J. R. Hernandez, J. Houck, D. Hostin, K. A. Houston, T. J. Howland, M.-H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G. H. Karpen, Z. Ke, J. A. Kennison, K. A. Ketchum, B. E. Kimmel, C. D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A. A. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T. C. McIntosh, M. P. McLeod, D. McPherson, G. Merkulov, N. V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S. M. Mount, M. Moy, B. Murphy, L. Murphy, D. M. Muzny, D. L. Nelson, D. R. Nelson, K. A. Nelson, K. Nixon, D. R. Nusskern, J. M. Pacleb, M. Palazzolo, G. S. Pittman, S. Pan, J. Pollard, V. Puri, M. G. Reese, K. Reinert, K. Remington, R. D. C. Saunders, F. Scheeler, H. Shen, B. C. Shue, I. Sidén-Kiamos, M. Simpson, M. P. Skupski, T. Smith, E. Spier, A. C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A. H. Wang, X. Wang, Z.-Y. Wang, D. A. Wassarman, G. M. Weinstock, J. Weissenbach, S. M. Williams, T. Woodage, K. C. Worley, D. Wu, S. Yang, Q. A. Yao, J. Ye, R.-F. Yeh, J. S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X. H. Zheng, F. N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H. O. Smith, R. A. Gibbs, E. W. Myers, G. M. Rubin, et J. C. Venter. 2000. The genome sequence of *Drosophila melanogaster*. *Science*, 287 (5461):2185–2195.

Y. T. Aminetzach, J. M. Macpherson, et D. A. Petrov. 2005. Pesticide resistance via transposition-mediated adaptive gene truncation in *Drosophila*. *Science*, 309(5735):764–767.

D. Anxolabéhère, M. G. Kidwell, et G. Periquet. 1988. Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of *Drosophila melanogaster* by mobile P elements. *Molecular Biology and Evolution*, 5(3):252–269.

A. A. Aravin, G. J. Hannon, et J. Brennecke. 2007. The piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science*, 318(5851):761–764.

O. T. Avery, C. M. Macleod, et M. McCarty. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* type III. *Journal of Experimental Medecine*, 79(2):137–158.

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

T. M. Barnes, Y. Kohara, A. Coulson, et S. Hekimi. 1995. Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. *Genetics*, 141(1):159–179.

J. E. Barrick, D. S. Yu, S. H. Yoon, H. Jeong, T. K. Oh, D. Schneider, R. E. Lenski, et J. F. Kim. 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature*, 461(7268):1243–1247.

R. S. Baucom, J. C. Estill, C. Chaparro, N. Upshaw, N. Jogi, J. M. Deragon, R. P. Westerman, P. J. SanMiguel, et J. L. Bennetzen. 2009. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genetics*, 5(11):e1000732+.

K. Beck. 1999. *Extreme Programming explained : embrace change*. Addison-Wesley Professional.

C. Bergman, H. Quesneville, D. Anxolabéhère, et M. Ashburner. 2006. Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome Biology*, 7(11):R112+.

C. M. M. Bergman et H. Quesneville. 2007. Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics*, 8(6):382–392.

T. Bestor. 1999. Sex brings transposons and genomes into conflict. *Genetica*, 107(1):289–295.

J. Boeke, D. J. Garfinkel, C. A. Styles, et G. R. Fink. 1985. Ty elements transpose through an RNA intermediate. *Cell*, 40(3):491–500.

G. Bourque, B. Leong, V. B. Vega, X. Chen, Y. L. Lee, K. G. Srinivasan, J.-L. Chew, Y. Ruan, C.-L. Wei, H. H. Ng, et E. T. Liu. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Research*, 18(11):1752–1762.

R. J. Britten et E. H. Davidson. 1969. Gene regulation for higher cells : a theory. *Science*, 165(891) : 349–357.

R. J. Britten et E. H. Davidson. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *The Quarterly Review of Biology*, 46(2) : 111–138.

N. Buisine, H. Quesneville, et V. Colot. 2008. Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. *Genomics*, 91(5) : 467–475.

R. A. Cartwright. 2005. DNA assembly with gaps (Dawg) : simulating sequence evolution. *Bioinformatics*, 21(Suppl_3) : iii31–38.

B. Charlesworth et D. Charlesworth. 1983. The population dynamics of transposable elements. *Genetical Research*, 42(1) : 1–27.

F. Choulet, T. Wicker, C. Rustenholz, C. Paux, J. Salse, P. Leroy, S. Schlub, M. C. Le Paslier, G. Magdelenat, C. Gonthier, A. Couloux, H. Budak, J. Breen, M. Pumphrey, S. Liu, X. Kong, J. Jia, M. Gut, D. Brunel, J. A. Anderson, B. S. Gill, R. Appels, B. Keller, et C. Feuillet. 2010. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *The Plant Cell*, 22(6) : 1686–1701.

M. M. Cock, L. Sterck, P. Rouzé, D. Scornet, A. E. Allen, G. Amoutzias, V. Anthouard, F. Artiguenave, J.-M. M. Aury, J. H. Badger, B. Beszteri, K. Billiau, E. Bonnet, J. H. Bothwell, C. Bowler, C. Boyen, C. Brownlee, C. J. Carrano, B. Charrier, G. Youn, S. M. Coelho, J. Collén, E. Corre, C. Da Silva, L. Delage, N. Delaroque, S. M. Dittami, S. Doulbeau, M. Elias, G. Farnham, C. M. Gachon, B. Gschloessl, S. Heesch, K. Jabbari, C. Jubin, H. Kawai, K. Kimura, B. Kloareg, F. C. Küpper, D. Lang, A. Le Bail, C. Leblanc, P. Lerouge, M. Lohr, P. J. Lopez, C. Martens, F. Maumus, G. Michel, D. Miranda-Saavedra, J. Morales, H. Moreau, T. Motomura, C. Nagasato, C. A. Napoli, D. R. Nelson, P. Nyvall-Collén, A. F. Peters, C. Pommier, P. Potin, J. Poulain, H. Quesneville, B. Read, S. A. Rensing, A. Ritter, S. Rousvoal, M. Samanta, G. Samson, D. C. Schroeder, B. Ségurens, M. Strittmatter, T. Tonon, J. W. Tregear, K. Valentin, P. von Dassow, T. Yamagishi, Y. Van de Peer, et P. Wincker. 2010. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature*, 465(7298) : 617–621.

B. C. Collard et D. J. Mackill. 2008. Marker-assisted selection : an approach for precision plant breeding in the twenty-first century. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1491) : 557–572.

G. Coop et M. Przeworski. 2006. An evolutionary view of human recombination. *Nature Reviews Genetics*, 8(1) : 23–34.

E. d'Alençon, H. Sezutsu, F. Legeai, E. Permal, S. Bernard-Samain, S. Gimenez, C. Gagneur, F. Cousserans, M. Shimomura, A. Brun-Barale, T. Flutre, A. Couloux, P. East, K. Gordon, K. Mita, H. Quesneville, P. Fournier, et R. Feyereisen. 2010. Extensive synteny conservation of holocentric chromosomes in Lepidoptera despite high rates of local genome rearrangements. *Proceedings of the National Academy of Sciences*, 107(17) : 7680–7685.

C. Darwin. 1992. *L'Origine des espèces*. Flammarion, Paris.

R. Dawkins. 1982. *The extended phenotype : the long reach of the gene*. Oxford University Press, USA, revised edition.

A. F. Dernburg. 2001. Here, there, and everywhere : kinetochore function on holocentric chromosomes. *Journal of Cell Biology*, 153 : F33–F38.

J. Diamond. 2002. Evolution, consequences and future of plant and animal domestication. *Nature*, 418(6898) : 700–707.

T. Dobzhansky et A. H. Sturtevant. 1938. Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics*, 23(1) : 28–64.

B. A. Dombroski, S. L. Mathias, E. Nanthakumar, A. F. Scott, et H. H. Kazazian. 1991. Isolation of an active human transposable element. *Science*, 254(5039) : 1805–1808.

I. Dondoshansky. *Blastclust (NCBI software development toolkit)*, 6.1 edition, 2002.

J. Drouaud, C. Camilleri, P. Y. Bourguignon, A. Canaguier, A. Bérard, D. Vezon, S. Giancola, D. Brunel, V. Colot, B. Prum, H. Quesneville, et C. Mézard. 2006. Variation in crossing-over rates across chromosome 4 of *Arabidopsis thaliana* reveals the presence of meiotic recombination "hot spots". *Genome Research*, 16(1) : 106–114.

R. C. Edgar et E. W. Myers. 2005. PILER : identification and classification of genomic repeats. *Bioinformatics*, 21(suppl_1) : i152–158.

E. E. Eichler et D. Sankoff. 2003. Structural dynamics of eukaryotic chromosome evolution. *Science*, 301(5634) : 793–797.

D. Ellinghaus, S. Kurtz, et U. Willhoeft. 2008. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics*, 9(1) : 18+.

M. B. Elowitz et S. Leibler. 2000. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767) : 335–338.

C. Feschotte et E. J. Pritham. 2005. Non-mammalian c-integrases are encoded by giant transposable elements. *Trends in Genetics*, 21(10) : 551–552.

D. J. Finnegan. 1989. Eukaryotic transposable elements and genome evolution. *Trends in Genetics*, 5(4) : 103–107.

A. S. Fiston-Lavier, D. Anxolabéhère, et H. Quesneville. 2007. A model of segmental duplication formation in *Drosophila melanogaster*. *Genome Research*.

J. M. Flowers et M. D. Purugganan. 2008. The evolution of plant genomes : scaling up from a population perspective. *Current Opinion in Genetics & Development*, 18(6) : 565–570.

S. Foissac, J. Gouzy, S. Rombauts, C. Mathe, J. Amselem, L. Sterck, Y. Van de Peer, P. Rouze, et T. Schiex. 2008. Genome annotation in plants and fungi : EuGene as a model platform. *Current Bioinformatics*, 3(2) : 87–97.

A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, et S. G. Oliver. 1996. Life with 6000 genes. *Science*, 274(5287) : 546–567.

S. J. Gould et R. C. Lewontin. 1979. The spandrels of San Marco and the Panglossian paradigm : a critique of the adaptationist programme. *Proceedings of the Royal Society of London B*, 205 : 581–598.

S. J. Gould et E. S. Vrba. 1982. Exaptation - a missing term in the science of form. *Paleobiology*, 8(1) : 4–15.

Y. H. Gray. 2000. It takes two transposons to tango : transposable-element-mediated chromosomal rearrangements. *Trends in Genetics*, 16(10) : 461–468.

W. D. Hamilton. 1963. The evolution of altruistic behavior. *The American Naturalist*, 97(896) : 354–356.

K. Hanada, V. Vallejo, K. Nobuta, R. K. Slotkin, D. Lisch, B. C. Meyers, S. H. Shiu, et N. Jiang. 2009. The functional role of Pack-MULEs in rice inferred from purifying selection and expression profile. *Plant Cell*, 21(1) : 25–38.

W. G. Hill et A. Robertson. 1966. The effect of linkage on limits to artificial selection. *Genetical Research*, 8(3) : 269–294.

F. Houllier. Mission de réflexion sur l'organisation et le pilotage de l'INRA : internationalisation, efficacité, attractivité, 2009.

S. Huang, R. Li, Z. Zhang, L. Li, X. Gu, W. Fan, W. J. Lucas, X. Wang, B. Xie, P. Ni, Y. Ren, H. Zhu, J. Li, K. Lin, W. Jin, Z. Fei, G. Li, J. Staub, A. Kilian, E. A. G. van der Vossen, Y. Wu, J. Guo, J. He, Z. Jia, Y. Ren, G. Tian, Y. Lu, J. Ruan, W. Qian, M. Wang, Q. Huang, B. Li, Z. Xuan, J. Cao, Asan, Z. Wu, J. Zhang, Q. Cai, Y. Bai, B. Zhao, Y. Han, Y. Li, X. Li, S. Wang, Q. Shi, S. Liu, W. K. Cho, J.-Y. Kim, Y. Xu, K. Heller-Uszynska, H. Miao, Z. Cheng, S. Zhang, J. Wu, Y. Yang, H. Kang, M. Li, H. Liang, X. Ren, Z. Shi, M. Wen, M. Jian, H. Yang, G. Zhang, Z. Yang, R. Chen, S. Liu, J. Li, L. Ma, H. Liu, Y. Zhou, J. Zhao, X. Fang, G. Li, L. Fang, Y. Li, D. Liu, H. Zheng, Y. Zhang, N. Qin, Z. Li, G. Yang, S. Yang, L. Bolund, K. Kristiansen, H. Zheng, S. Li, X. Zhang, H. Yang, J. Wang, R. Sun, B. Zhang, S. Jiang, J. Wang, Y. Du, et S. Li. 2009. The genome of the cucumber, *Cucumis sativus L. Nature Genetics*, 41(12) : 1275–1281.

J. W. Huss, C. Orozco, J. Goodale, C. Wu, S. Batalov, T. J. Vickers, F. Valafar, et A. I. Su. 2008. A gene wiki for community annotation of gene function. *PLoS Biology*, 6(7) : e175+.

Z. Ivics, P. B. Hackett, R. H. Plasterk, et Z. Izsvák. 1997. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell*, 91(4) : 501–510.

Z. Izsvák, P. B. Hackett, L. J. Cooper, et Z. Ivics. 2010. Translating Sleeping Beauty transposition into cellular therapies : victories and challenges. *BioEssays*.

F. Jacob. 1977. Evolution and tinkering. *Science*, 196(4295) : 1161–1166.

F. Jacob et J. Monod. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3) : 318–356.

N. Jiang, Z. Bao, X. Zhang, S. R. Eddy, et S. R. Wessler. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, 431(7008) : 569–573.

Z. Jiang, H. Tang, V. Ventura, M. F. Cardone, T. Marques-Bonet, X. She, P. A. Pevzner, et E. E. Eichler. 2007. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nature Genetics*, 39(11) : 1361–1368.

N. Jones. 2005. McClintock's controlling elements : the full story. *Cytogenetics and Genome Research*, 109 : 90–103.

R. Kalendar, C. M. Vicient, O. Peleg, K. Anamthawat-Jonsson, A. Bolshoy, et A. H. Schulman. 2004. Large retrotransposon derivatives : abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics*, 166(3) : 1437–1450.

R. Kalendar, J. Tanskanen, W. Chang, K. Antonius, H. Sela, O. Peleg, et A. H. Schulman. 2008. Cassandra retrotransposons carry independently transcribed 5S RNA. *Proceedings of the National Academy of Sciences*, 105(15) : 5833–5838.

A. Kalyanaraman et S. Aluru. 2006. Efficient algorithms and software for detection of full-length LTR retrotransposons. *J Bioinform Comput Biol*, 4(2) : 197–216.

J. Kaminker, C. Bergman, B. Kronmiller, J. Carlson, R. Svirskas, S. Patel, E. Frise, D. Wheeler, S. Lewis, G. Rubin, M. Ashburner, et S. Celniker. 2002. The transposable elements of the *Drosophila melanogaster* euchromatin : a genomics perspective. *Genome Biology*, 3(12).

S. Kang, M. H. Lebrun, L. Farrall, et B. Valent. 2001. Gain of virulence caused by insertion of a Pot3 transposon in a *Magnaporthe grisea* avirulence gene. *Molecular Plant-Microbe Interactions*, 14(5) : 671–674.

V. V. Kapitonov et J. Jurka. 2001. Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences*, 98(15) : 8714–8719.

V. V. Kapitonov et J. Jurka. 2006. Self-synthesizing DNA transposons in eukaryotes. *Proceedings of the National Academy of Sciences*, 103(12) : 4540–4545.

N. Kashtan, E. Noor, et U. Alon. 2007. Varying environments can speed up evolution. *Proceedings of the National Academy of Sciences*, 104(34) : 13711–13716.

J. J. Kent. 2002. BLAT - the BLAST-like alignment tool. *Genome Research*, 12(4) : 656–664.

C. H. Khang, S. Y. Park, Y. H. Lee, B. Valent, et S. Kang. 2008. Genome organization and evolution of the AVR-Pita avirulence gene family in the *Magnaporthe grisea* species complex. *Molecular Plant-Microbe Interactions*, 21 (5) : 658–670.

J. M. Kidd, G. M. Cooper, W. F. Donahue, H. S. Hayden, N. Sampas, T. Graves, N. Hansen, B. Teague, C. Alkan, F. Antonacci, E. Haugen, T. Zerr, N. A. Yamada, P. Tsang, T. L. Newman, E. Tuzun, Z. Cheng, H. M. Ebling, N. Tusneem, R. David, W. Gillett, K. A. Phelps, M. Weaver, D. Saranga, A. Brand, W. Tao, E. Gustafson, K. McKernan, L. Chen, M. Malig, J. D. Smith, J. M. Korn, S. A. McCarroll, D. A. Altshuler, D. A. Peiffer, M. Dorschner, J. Stamatoyannopoulos, D. Schwartz, D. A. Nickerson, J. C. Mullikin, R. K. Wilson, L. Bruhn, M. V. Olson, R. Kaul, D. R. Smith, et E. E. Eichler. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191) : 56–64.

M. G. Kidwell, J. F. Kidwell, et J. A. Sved. 1977. Hybrid dysgenesis in *Drosophila melanogaster* : a syndrome of aberrant traits including mutation, sterility and male recombination. *Genetics*, 86(4) : 813–833.

C. Knibbe, O. Mazet, F. Chaudier, J. Fayard, et G. Beslon. 2007. Evolutionary coupling between the deleteriousness of gene mutations and the amount of non-coding sequences. *Journal of Theoretical Biology*, 244(4) : 621–630.

A. Le Rouzic et P. Capy. 2005. The first steps of transposable elements invasion : parasitic strategy versus genetic drift. *Genetics*, 169(2) : 1033–1043.

A. Le Rouzic, T. S. Boutin, et P. Capy. 2007. Long-term evolution of transposable elements. *Proceedings of the National Academy of Sciences*, 104(49) : 19375–19380.

E. Lerat. 2009. Identifying repeats and transposable elements in sequenced genomes : how to find your way through the dense forest of programs. *Heredity*, 104(6) : 520–533.

M. F. Lin, A. N. Deoras, M. D. Rasmussen, et M. Kellis. 2008. Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. *PLoS Computational Biology*, 4(4) : e1000067+.

F. Mallet, O. Bouton, S. Prudhomme, V. Cheynet, G. Oriol, B. Bonnaud, G. Lucotte, L. Duret, et B. Mandrand. 2004. The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. *Proceedings of the National Academy of Sciences*, 101(6) : 1731–1736.

E. Mayr. 1961. Cause and effect in biology. *Science*, 134 : 1501–1506.

E. M. McCarthy et J. F. McDonald. 2003. LTR_STRUC : a novel search and identification program for LTR retrotransposons. *Bioinformatics*, 19(3) : 362–367.

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

B. McClintock. 1956. Controlling elements and the gene. *Cold Spring Harbor Symposia on Quantitative Biology*, 21 : 197–216.

B. McClintock. 1961. Some parallels between gene control systems in maize and in bacteria. *The American Naturalist*, 95(884) : 265–277.

M. McNeil, R. Kota, E. Paux, D. Dunn, R. McLean, C. Feuillet, D. Li, X. Kong, E. Lagudah, J. Zhang, J. Jia, W. Spielmeyer, M. Bellgard, et R. Appels. 2008. Bac-derived markers for assaying the stem rust resistance gene, Sr2, in wheat breeding programs. *Molecular Breeding*, 22(1) : 15–24.

M. L. Metzker. 2009. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1) : 31–46.

R. Ming, S. Hou, Y. Feng, Q. Yu, A. Dionne-Laporte, J. H. Saw, P. Senin, W. Wang, B. V. Ly, K. L. T. Lewis, S. L. Salzberg, L. Feng, M. R. Jones, R. L. Skelton, J. E. Murray, C. Chen, W. Qian, J. Shen, P. Du, M. Eustice, E. Tong, H. Tang, E. Lyons, R. E. Paull, T. P. Michael, K. Wall, D. W. Rice, H. Albert, M.-L. Wang, Y. J. Zhu, M. Schatz, N. Nagarajan, R. A. Acob, P. Guan, A. Blas, C. M. Wai, C. M. Ackerman, Y. Ren, C. Liu, J. Wang, J. Wang, J.-K. Na, E. V. Shakirov, B. Haas, J. Thimmapuram, D. Nelson, X. Wang, J. E. Bowers, A. R.

Gschwend, A. L. Delcher, R. Singh, J. Y. Suzuki, S. Tripathi, K. Neupane, H. Wei, B. Irikura, M. Paidi, N. Jiang, W. Zhang, G. Presting, A. Windsor, R. Navajas-Perez, M. J. Torres, F. A. Feltus, B. Porter, Y. Li, A. M. Burroughs, M.-C. Luo, L. Liu, D. A. Christopher, S. M. Mount, P. H. Moore, T. Sugimura, J. Jiang, M. A. Schuler, V. Friedman, T. Mitchell-Olds, D. E. Shippen, C. W. dePamphilis, J. D. Palmer, M. Freeling, A. H. Paterson, D. Gonsalves, L. Wang, et M. Alam. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya Linnaeus*). *Nature*, 452(7190) : 991–996.

J. V. Moran, R. J. DeBerardinis, et H. H. Kazazian. 1999. Exon shuffling by L1 retrotransposition. *Science*, 283(5407) : 1530–1534.

M. Morange. 2000. *A history of molecular biology*. Harvard University Press.

M. Morgante, S. Brunner, G. Pea, K. Fengler, A. Zuccolo, et A. Rafalski. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature Genetics*.

J. Nicolas, P. Durand, G. Ranchy, S. Tempel, et A. S. Valin. 2005. Suffix-tree analyser (STAN) : looking for nucleotidic and peptidic patterns in chromosomes. *Bioinformatics*, 21(24) : 4408–4410.

L. Noé et G. Kucherov. 2004. Improved hit criteria for DNA local alignment. *BMC Bioinformatics*, 5(1) : 149–158.

L. Noé et G. Kucherov. 2005. YASS : enhancing the sensitivity of DNA similarity search. *Nucleic Acids Research*, 33(suppl_2) : W540–543.

K. O'Hare et G. M. Rubin. 1983. Structures of P transposable elements and their sites of insertion and excision in the *Drosophila melanogaster* genome. *Cell*, 34(1) : 25–35.

S. Ohno. 1972. So much "junk" DNA in our genome. *Brookhaven symposia in biology*, 23 : 366–370.

L. E. Orgel et F. H. Crick. 1980. Selfish DNA : the ultimate parasite. *Nature*, 284(5757) : 604–607.

P. J. Park. 2009. ChIP–seq : advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10) : 669–680.

A. H. Paterson, J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, G. Haberer, U. Hellsten, T. Mitros, A. Poliakov, J. Schmutz, M. Spannagl, H. Tang, X. Wang, T. Wicker, A. K. Bharti, J. Chapman, F. A. Feltus, U. Gowik, I. V. Grigoriev, E. Lyons, C. A. Maher, M. Martis, A. Narechania, R. P. Otillar, B. W. Penning, A. A. Salamov, Y. Wang, L. Zhang, N. C. Carpita, M. Freeling, A. R. Gingle, C. T. Hash, B. Keller, P. Klein, S. Kresovich, M. C. McCann, R. Ming, D. G. Peterson, M. ur Rahman, D. Ware, P. Westhoff, K. F. X. Mayer, J. Messing, et D. S. Rokhsar. 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, 457(7229) : 551–556.

E. Paux, D. Roger, E. Badaeva, G. Gay, M. Bernard, P. Sourdille, et C. Feuillet. 2006. Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. *The Plant Journal*, 48(3) : 463–474.

E. Paux, P. Sourdille, J. Salse, C. Saintenac, F. Choulet, P. Leroy, A. Korol, M. Michalak, S. Kianian, W. Spielmeyer, E. Lagudah, D. Somers, A. Kilian, M. Alaux, S. Vautrin, H. Berges, K. Eversole, R. Appels, J. Safar, H. Simkova, J. Dolezel, M. Bernard, et C. Feuillet. 2008. A physical map of the 1-gigabase bread wheat chromosome 3B. *Science*, 322(5898):101–104.

E. Paux, S. Faure, F. Choulet, D. Roger, V. Gauthier, J. P. Martinant, P. Sourdille, F. Balfourier, M. C. Le Paslier, A. Chauveau, M. Cakir, B. Gandon, et C. Feuillet. 2010. Insertion site-based polymorphism markers open new perspectives for genome saturation and marker-assisted selection in wheat. *Plant Biotechnology Journal*, 8(2):196–210.

D. A. Petrov. 2002. Mutational equilibrium model of genome size evolution. *Theoretical Population Biology*, 61(4):531–544.

P. Pevzner et G. Tesler. 2003. Genome rearrangements in mammalian evolution : lessons from human and mouse genomes. *Genome Research*, 13(1): 37–45.

H. Quesneville et D. Anxolabéhère. 1998. Dynamics of transposable elements in metapopulations : a model of P element invasion in Drosophila. *Theoretical Population Biology*, 54(2):175–193.

H. Quesneville et D. Anxolabéhère. 2001. Genetic algorithm-based model of evolutionary dynamics of class II transposable elements. *Journal of Theoretical Biology*, 213(1):21–30.

H. Quesneville, D. Nouaud, et D. Anxolabéhère. 2003. Detection of new transposable element families in *Drosophila melanogaster* and *Anopheles gambiae genomes*. *Journal of Molecular Evolution*, 57 Suppl 1.

H. Quesneville, C. M. Bergman, O. Andrieu, D. Autard, D. Nouaud, M. Ashburner, et D. Anxolabéhère. 2005. Combined evidence annotation of transposable elements in genome sequences. *PLoS Computational Biology*, 1(2): 166–175.

H. Quesneville, D. Nouaud, et D. Anxolabéhère. 2006. P elements and MITE relatives in the whole genome sequence of *Anopheles gambiae*. *BMC genomics*, 7(1):214+.

M. Rho, J. H. Choi, S. Kim, M. Lynch, et H. Tang. 2007. *De novo* identification of LTR retrotransposons in eukaryotic genomes. *BMC Genomics*, 8:90+.

C. Rizzon, G. Marais, M. Gouy, et C. Biemont. 2002. Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome. *Genome Research*, 12(3):400–407.

D. Roze et N. H. Barton. 2006. The Hill-Robertson effect and the evolution of recombination. *Genetics*, 173(3):1793–1811.

G. M. Rubin et A. C. Spradling. 1982. Genetic transformation of *Drosophila* with transposable element vectors. *Science*, 218(4570):348–353.

F. Sabot et A. H. Schulman. 2006. Parasitism and the retrotransposon life cycle in plants : a hitchhiker's guide to the genome. *Heredity*, 97(6):381–388.

S. Saha, S. Bridges, Z. V. Magbanua, et D. G. Peterson. 2008. Empirical comparison of ab initio repeat finding programs. *Nucleic Acids Research*, 36(7): 2284–2294.

P. SanMiguel, B. S. Gaut, A. Tikhonov, Y. Nakajima, et J. L. Bennetzen. 1998. The paleontology of intergene retrotransposons of maize. *Nature Genetics*, 20(1) : 43–45.

I. Sbalzarini et P. Koumoutsakos. 2005. Feature point tracking and trajectory analysis for video imaging in cell biology. *Journal of Structural Biology*, 151 (2) : 182–195.

J. M. Schmidt, R. T. Good, B. Appleton, J. Sherrard, G. C. Raymant, M. R. Bogwitz, J. Martin, P. J. Daborn, M. E. Goddard, P. Batterham, et C. Robin. 2010. Copy number variation and transposable elements feature in recent, ongoing adaptation at the Cyp6g1 locus. *PLoS Genetics*, 6(6) : e1000998+.

P. S. Schnable, D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. Denise, L. Courtney, S. S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S. M. Rock, E. Belter, F. Du, K. Kim, R. M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S. M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges, E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S. Kumari, B. Faga, M. J. Levy, L. McMahan, P. Van Buren, M. W. Vaughn, K. Ying, C.-T. T. Yeh, S. J. Emrich, Y. Jia, A. Kalyanaraman, A.-P. P. Hsia, B. B. Barbazuk, R. S. Baucom, T. P. Brutnell, N. C. Carpita, C. Chaparro, J.-M. M. Chia, J.-M. M. Deragon, J. C. Estill, Y. Fu, J. A. Jeddeloh, Y. Han, H. Lee, P. Li, D. R. Lisch, S. Liu, Z. Liu, D. Holligan, M. C. McCann, P. SanMiguel, A. M. Myers, D. Nettleton, J. Nguyen, B. W. Penning, L. Ponnala, K. L. Schneider, D. C. Schwartz, A. Sharma, C. Soderlund, N. M. Springer, Q. Sun, H. Wang, M. Waterman, R. Westerman, T. K. Wolfgruber, L. Yang, Y. Yu, L. Zhang, S. Zhou, Q. Zhu, J. L. Bennetzen, K. K. Dawe, J. Jiang, N. Jiang, G. G. Presting, S. R. Wessler, S. Aluru, R. A. Martienssen, S. W. Clifton, R. R. McCombie, R. A. Wing, et R. K. Wilson. 2009. The B73 maize genome : complexity, diversity, and dynamics. *Science*, 326(5956) : 1112–1115.

E. U. Selker. 1990. Premeiotic instability of repeated sequences in *Neurospora crassa*. *Annual Review of Genetics*, 24(1) : 579–613.

A. F. A. Smit. 1993. Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Research*, 21(8) : 1863–1872.

T. Tamura, C. Thibert, C. Royer, T. Kanda, A. Eappen, M. Kamba, N. Komoto, J. L. Thomas, B. Mauchamp, G. Chavancy, P. Shirk, M. Fraser, J. C. Prudhomme, et P. Couble. 2000. Germline transformation of the silkworm *Bombyx mori L.* using a piggyBac transposon-derived vector. *Nature Biotechnology*, 18(1) : 81–84.

S. Tempel, J. Nicolas, A. El Amrani, et I. Couée. 2007. Model-based identification of Helitrons results in a new classification of their families in *Arabidopsis thaliana*. *Gene*, 403(1-2) : 18–28.

the Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814):796–815.

the French–Italian Public Consortium for Grapevine Genome Characterization. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161):463–467.

the International Aphid Genomics Consortium. 2010. Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biology*, 8(2):e1000313+.

the International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

the Welcome Trust. Sharing data from large-scale biological research projects :a system of tripartite responsibility. Technical report, Wellcome Trust, 2003.

S. T. Thibault, M. A. Singer, W. Y. Miyazaki, B. Milash, N. A. Dompe, C. M. Singh, R. Buchholz, M. Demsky, R. Fawcett, H. L. Francis-Lang, L. Ryner, L. M. Cheung, A. Chong, C. Erickson, W. W. Fisher, K. Greer, S. R. Hartouni, E. Howie, L. Jakkula, D. Joo, K. Killpack, A. Laufer, J. Mazzotta, R. D. Smith, L. M. Stevens, C. Stuber, L. R. Tan, R. Ventura, A. Woo, I. Zakrajsek, L. Zhao, F. Chen, C. Swimmer, C. Kopczynski, G. Duyk, M. L. Winberg, et J. Margolis. 2004. A complementary transposon tool kit for *Drosophila melanogaster* using P and piggyBac. *Nature Genetics*, 36(3):283–287.

J. N. Thompson. 2005. *The geographic mosaic of coevolution*. University Of Chicago Press, 1 edition.

S. Tsukahara, A. Kobayashi, A. Kawabe, O. Mathieu, A. Miura, et T. Kakutani. 2009. Bursts of retrotransposition reproduced in Arabidopsis. *Nature*, 461 (7262):423–426.

Z. Wang, M. Gerstein, et M. Snyder. 2009. RNA-Seq : a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

J. D. Watson et F. H. C. Crick. 1953. Molecular structure of nucleic acids : a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, et A. H. Schulman. 2007. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12):973–982.

G. C. Williams. 1966. *Adaptation and natural selection*. Princeton University Press.

C. P. Witte, Q. H. Le, T. Bureau, et A. Kumar. 2001. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proceedings of the National Academy of Sciences*, 98(24):13778–13783.

S. I. Wright, N. Agrawal, et T. E. Bureau. 2003. Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. *Genome Research*, 13(8):1897–1903.

J. Xing, H. Wang, V. P. Belancio, R. Cordaux, P. L. Deininger, et M. A. Batzer. 2006. Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proceedings of the National Academy of Sciences*, 103(47):17608–17613.

Z. Xu et H. Wang. 2007. LTR_FINDER : an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, 35(suppl_2) : W265–268.

G. Yang, D. H. Nagel, C. Feschotte, C. N. Hancock, et S. R. Wessler. 2009. Tuned for transposition : molecular determinants underlying the hyperactivity of a Stowaway MITE. *Science*, 325(5946) : 1391–1394.

L. Yang et J. L. Bennetzen. 2009. Structure-based discovery and description of plant and animal Helitrons. *Proceedings of the National Academy of Sciences*, 106(31) : 12832–12837.

F. M. You, H. Wanjugi, N. Huo, G. R. Lazo, M. C. Luo, O. D. Anderson, J. Dvorak, et Y. Q. Gu. 2010. RJPrimers : unique transposable element insertion junction discovery and PCR primer design for marker development. *Nucleic Acids Research*, 38(suppl_2) : W313–320.

L. Zhou, R. Mitra, P. W. Atkinson, A. Burgess Hickman, F. Dyda, et N. L. Craig. 2004. Transposition of hAT elements links transposable elements and V(D)J recombination. *Nature*, 432(7020) : 995–1001.

Y. Zhou et B. Mishra. 2005. Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. *Proceedings of the National Academy of Sciences*, 102(11) : 4051–4056.

A

ANNEXES

A.1 ARTICLE "GENOME SEQUENCE OF THE METAZOAN PLANT-PARASITIC NEMATODE MELOIDOGYNE INCOGNITA"

Seuls les suppléments correspondant à l'analyse des éléments transposables sont inclus.

ARTICLES

nature biotechnology

Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*

Pierre Abad[1–3], Jérôme Gouzy[4], Jean-Marc Aury[5–7], Philippe Castagnone-Sereno[1–3], Etienne G J Danchin[1–3], Emeline Deleury[1–3], Laetitia Perfus-Barbeoch[1–3], Véronique Anthouard[5–7], François Artiguenave[5–7], Vivian C Blok[8], Marie-Cécile Caillaud[1–3], Pedro M Coutinho[9], Corinne Dasilva[5–7], Francesca De Luca[10], Florence Deau[1–3], Magali Esquibet[11], Timothé Flutre[12], Jared V Goldstone[13], Noureddine Hamamouch[14], Tarek Hewezi[15], Olivier Jaillon[5–7], Claire Jubin[5–7], Paola Leonetti[10], Marc Magliano[1–3], Tom R Maier[15], Gabriel V Markov[16,17], Paul McVeigh[18], Graziano Pesole[19,20], Julie Poulain[5–7], Marc Robinson-Rechavi[21,22], Erika Sallet[23,24], Béatrice Ségurens[5–7], Delphine Steinbach[12], Tom Tytgat[25], Edgardo Ugarte[5–7], Cyril van Ghelder[1–3], Pasqua Veronico[10], Thomas J Baum[15], Mark Blaxter[26], Teresa Bleve-Zacheo[10], Eric L Davis[14], Jonathan J Ewbank[27], Bruno Favery[1–3], Eric Grenier[11], Bernard Henrissat[9], John T Jones[8], Vincent Laudet[16], Aaron G Maule[18], Hadi Quesneville[12], Marie-Noëlle Rosso[1–3], Thomas Schiex[24], Geert Smant[25], Jean Weissenbach[5–7] & Patrick Wincker[5–7]

Plant-parasitic nematodes are major agricultural pests worldwide and novel approaches to control them are sorely needed. We report the draft genome sequence of the root-knot nematode *Meloidogyne incognita*, a biotrophic parasite of many crops, including tomato, cotton and coffee. Most of the assembled sequence of this asexually reproducing nematode, totaling 86 Mb, exists in pairs of homologous but divergent segments. This suggests that ancient allelic regions in *M. incognita* are evolving toward effective haploidy, permitting new mechanisms of adaptation. The number and diversity of plant cell wall–degrading enzymes in *M. incognita* is unprecedented in any animal for which a genome sequence is available, and may derive from multiple horizontal gene transfers from bacterial sources. Our results provide insights into the adaptations required by metazoans to successfully parasitize immunocompetent plants, and open the way for discovering new antiparasitic strategies.

Plant-parasitic nematodes are responsible for global agricultural losses amounting to an estimated $157 billion annually. Although chemical nematicides are the most reliable means of controlling root-knot nematodes, they are increasingly being withdrawn owing to their toxicity to humans and the environment. Novel and specific targets are thus needed to develop new strategies against these pests.

The Southern root-knot nematode *Meloidogyne incognita* is able to infect the roots of almost all cultivated plants, making it perhaps the

[1]INRA, UMR 1301, 400 route des Chappes, F-06903 Sophia-Antipolis, France. [2]CNRS, UMR 6243, 400 route des Chappes, F-06903 Sophia-Antipolis, France. [3]UNSA, UMR 1301, 400 route des Chappes, F-06903 Sophia-Antipolis, France. [4]Laboratoire Interactions Plantes Micro-organismes, UMR441/2594, INRA/CNRS, Chemin de Borde Rouge, BP 52627, F-31320 Castanet Tolosan, France. [5]Genoscope (CEA), 2 rue Gaston Crémieux, CP5706, F-91057 Evry, France. [6]CNRS, UMR 8030, 2 rue Gaston Crémieux, CP5706, F-91057 Evry, France. [7]Université d'Evry, F-91057 Evry, France. [8]Plant Pathology Programme, SCRI, Invergowrie, Dundee DD2 5DA, UK. [9]CNRS, UMR 6098 CNRS and Universités d'Aix-Marseille I & II, Case 932, 163 Av. de Luminy, F-13288 Marseille, France. [10]Istituto per la Protezione delle Piante, Consiglio Nazionale delle Ricerche, Via G. Amendola 165/a, 70126 Bari, Italy. [11]INRA, Agrocampus Rennes, Univ. Rennes I, UMR1099 BiO3P, Domaine de la Motte, F-35653 Le Rheu Cedex, France. [12]INRA, UR1164 Unité de Recherche en Génomique et Informatique (URGI), 523 place des terrasses de l'Agora, F-91034 Evry, France. [13]Biology Department, Woods Hole Oceanographic Institution, Co-op Building, MS #16, Woods Hole, Massachusetts 02543, USA. [14]Department of Plant Pathology, North Carolina State University, 840 Method Road, Unit 4, Box 7903 Raleigh, North Carolina 27607,USA. [15]Department of Plant Pathology, Iowa State University, 351 Bessey Hall, Ames, Iowa 50011, USA. [16]Université de Lyon, Institut de Génomique Fonctionnelle de Lyon, Molecular Zoology team, Ecole Normale Supérieure de Lyon, Université Lyon 1, CNRS, INRA, Institut Fédératif 128 Biosciences Gerland, Lyon Sud, 46 allée d'Italie, F-69364 Lyon Cedex 07, France. [17]USM 501, Evolution des Régulations Endocriniennes, Muséum National d'Histoire Naturelle, 7 rue Cuvier, F-75005 Paris, France. [18]Biomolecular Processes: Parasitology, School of Biological Sciences, Medical Biology Centre, 97 Lisburn Road, Queen's University Belfast, Belfast BT9 7BL, UK. [19]Dipartimento di Biochimica e Biologia Molecolare "E. Quagliariello", University of Bari, Via Orabona 4, 70126 Bari, Italy. [20]Istituto Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Via G. Amendola, 122/D, 70126 Bari, Italy. [21]Department of Ecology and Evolution, University of Lausanne, UNIL-Sorge, Le Biophore, CH-1015 Lausanne, Switzerland. [22]Swiss Institute of Bioinformatics, quartier sorge, Bâtiment Genopode, CH-1015 Lausanne, Switzerland. [23]Plateforme Bioinformatique du Génopole Toulouse Midi-Pyrénées, GIS Toulouse Genopole, 24 Chemin de Borde Rouge, BP 52627, F-31320 Castanet Tolosan, France. [24]Unité de Biométrie et d'Intelligence Artificielle UR875, INRA, Chemin de Borde Rouge, BP 52627, F-31320 Castanet Tolosan, France. [25]Laboratory of Nematology, Wageningen University, Binnenhaven 5, 6709PD Wageningen, The Netherlands. [26]Institute of Evolutionary Biology, University of Edinburgh, Kings Buildings, Ashworth Laboratories, West Mains Road, Edinburgh EH9 3JT, UK. [27]INSERM/CNRS/Université de la Méditerranée, Centre d'Immunologie de Marseille-Luminy, 163 av. de Luminy, Case 906, F-13288, Marseille cedex 09, France. Correspondence should be addressed to P.A. (pierre.abad@sophia.inra.fr).

Received 30 April; accepted 25 June; published online 27 July 2008; doi:10.1038/nbt.1482

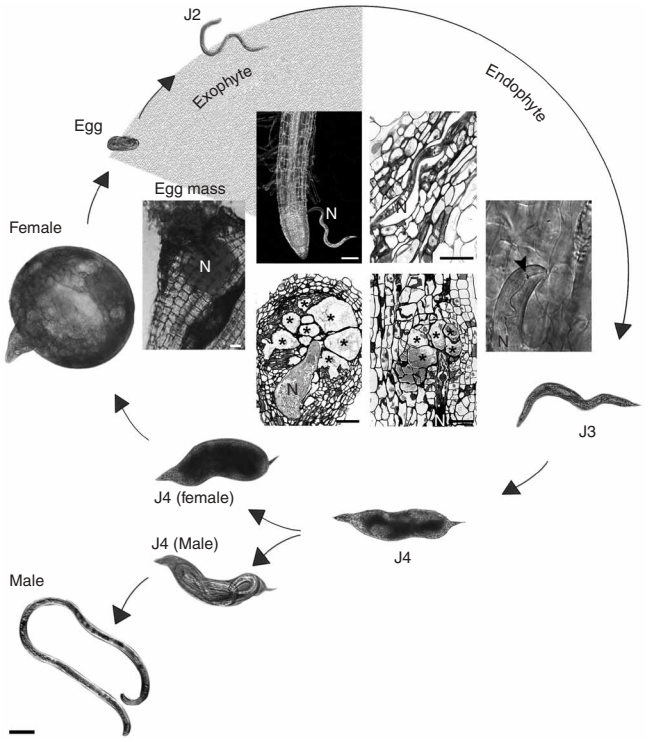


Figure 1 The parasitic life cycle of *Meloidogyne incognita*. Infective second-stage juveniles (J2) penetrate the root and migrate between cells to reach the plant vascular cylinder. The stylet (arrowhead) connected to the esophagus is used to pierce plant cell walls, to release esophageal secretions and to take up nutrients. Each J2 induces the dedifferentiation of five to seven root cells into multinucleate and hypertrophied feeding cells (*). These giant cells supply nutrients to the nematode (N). The nematode becomes sedentary and goes through three molts (J3, J4, adult). Occasionally, males develop and migrate out of the roots. However, it is believed that they play no role in reproduction. The pear-shaped female produces eggs that are released on the root surface. Embryogenesis within the egg is followed by the first molt, generating second-stage juveniles (J2). Scale bars, 50 μm.

most damaging of all crop pathogens[1]. *M. incognita* is an obligatory sedentary parasite that reproduces by mitotic parthenogenesis[2]. Root-knot nematodes have an intimate interaction with their hosts. Within the host root, adult females induce the redifferentiation of root cells into specialized 'giant' cells, upon which they feed continuously (**Fig. 1**). *M. incognita* can infect *Arabidopsis thaliana*, making this nematode a key model system for the understanding of metazoan adaptations to plant parasitism[3,4] (**Supplementary Data**, section 1 online).

The phylum Nematoda comprises >25,000 described species, many of which are parasites of animals or plants[2]. As many as 10 million species may have yet to be described. Although the model free-living nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* have been the subjects of intensive study[5,6], little is known about the other members of this diverse phylum. These two free-living models will likely not illuminate the biology of nematode parasitism (**Supplementary Fig. 1** online), as shown by the substantial differences between their genome sequences and that of the human parasite *Brugia malayi*[7].

The genome sequence of *M. incognita* presented here provides insights into the adaptations required by metazoans to successfully parasitize and counter defenses of immunocompetent plants, and suggests new antiparasitic strategies.

RESULTS

General features of the *M. incognita* genome

The *M. incognita* genome was sequenced using whole-genome shotgun strategy. Assembly with Arachne[8] yielded 2,817 supercontigs, totaling 86 Mb (**Table 1**; **Supplementary Data**, section 2; **Supplementary Fig. 2**; **Supplementary Table 1** online)—almost twice the estimated genome size (47- to 51-Mb haploid genome)[9]. All-against-all comparison of supercontigs revealed that 648 of the longest (covering ~55 Mb) consist of homologous but diverged segment pairs (**Fig. 2**) that might represent former alleles (**Supplementary Data**, section 2; **Supplementary Figs. 3** and **4** online). About 3.35 Mb of the assembly constitutes a third partial copy aligning with these supercontig pairs. Average sequence divergence between the aligned regions is ~8% (**Fig. 3**). A combination of different processes may explain the observed pattern in *M. incognita*, including polyploidy, polysomy, aneuploidy and hybridization[10,11]; all are frequently associated with asexual reproduction. These observations are consistent with a strictly mitotic parthenogenetic reproductive mode, which can permit homologous chromosomes to diverge considerably, as hypothesized for bdelloid rotifers[12] (**Supplementary Data**, section 2.2). No DNA attributable to bacterial endosymbiont genome(s) was identified.

Noncoding DNA repeats and transposable elements represent 36% of the *M. incognita* genome (**Supplementary Data**, section 3; **Supplementary Figs. 5** and **6** and **Supplementary Tables 2** and **3** online). One repeat family with 283 members on 46 contigs encoded the nematode *trans*-spliced leader (SL) exon, SL1, of which 258 members were found associated with a satellite DNA[13] (**Supplementary Fig. 7** online). In nematodes, many mature mRNAs share this 5′ SL exon, and *trans*-splicing is also associated with resolution of polycistronic pre-mRNAs derived from operons. We identified 1,585 candidate

Table 1 General features of the *Meloidogyne incognita* genome in comparison with the genomes of *B. malayi*[7] and *C. elegans*[5]

Features	*M. incognita*	*B. malayi*	*C. elegans*
Overall			
Estimated size of genome (Mb)	47–51[a]	90–95[a]	100[a]
Total size of assembled sequence (Mb)	86	88	100
Number of scaffolds and/or chromosomes (chr.)	2,817	8,180	6 chr.
G + C content (%)	31.4	30.5	35.4
Protein-coding regions			
Number of protein-coding gene models	19,212	11,515	20,072
Protein-coding sequence (% of genome)	25.3	17.8	25.5
Maximum/average protein length (amino acids)	5,970/354	9,420/343	18,562/440
Mean length of intergenic region (bp)	1,402	3,783	2,218
Gene density (genes per Mb)	223	162	228
Operon number	1,585	926	1,118
Percent of genes present in operon	19	18	14

For *B. malayi* a gene count ranging from 14,500 to 17,800 was inferred after inclusion of genes in the unannotated portion of the genome[7]. For *C. elegans* the gene and protein count is according to Wormpep database (WS183 release).
[a]*M. incognita*: flow cytometry[9]; *B. malayi*: flow cytometry and clone-based[7]; *C. elegans* genome has been completely sequenced telomere to telomere (no gaps) and is exactly 100,291,840 bp[45].

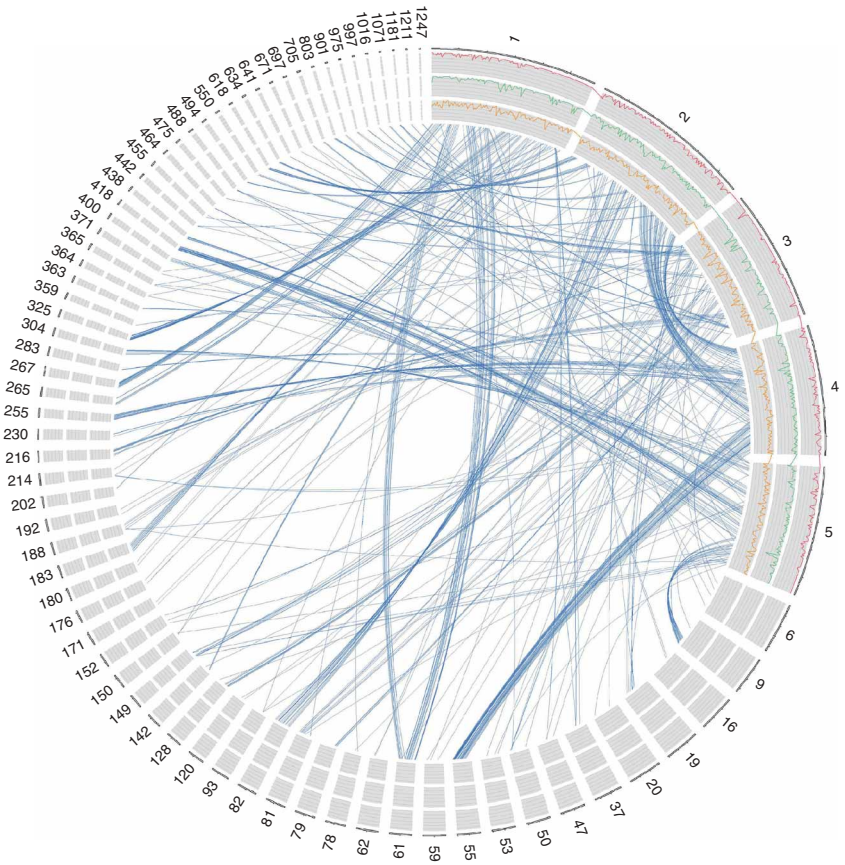


Figure 2 Allelic-like relationships for the five largest supercontigs of the *M. incognita* assembly. The five largest supercontigs are shown with plots of gene density (orange curve), conservation with *C. elegans* at amino acid level (green curve) and EST density (pink curve). Blue lines represent most similar matches at the protein level between each predicted gene on these five supercontigs and 70 matching supercontigs.

M. incognita operons containing a total of 3,966 genes. The two longest operons contained ten genes each and are not allelic copies (**Supplementary Table 4** online). Operons are a dynamic component of nematode genome architecture, as different sets of genes were operonic in *M. incognita*, *C. elegans* and *B. malayi*, and only one operon was found to be strictly conserved between the three nematodes (**Supplementary Data**, section 4; **Supplementary Figs. 8** and **9**; **Supplementary Table 5** online).

The gene content of a plant-parasitic nematode

The genome sequence was annotated using the integrative gene prediction platform EuGene[14], specifically trained for *M. incognita* (**Supplementary Data**, section 5; **Supplementary Table 6** online). We identified 19,212 protein-coding genes (**Table 1**). Due to the high variation between allelic-like copies (**Fig. 3**) potentially allowing functional divergence, all copies were considered to be different genes. Indeed, 69% of protein sequences were <95% identical to any other (**Supplementary Table 7** and **Supplementary Fig. 10** online). The protein-coding genes occupy 25.3% of the sequence at an average density of 223 genes Mb^{-1}, and 36% are supported by expressed sequence tags (ESTs). InterPro protein domains were identified in 55% of proteins and 22% were predicted to be secreted. Comparison of domain occurrence in *M. incognita* with that in *C. elegans* identified an increased abundance of 'pectate lyase', glycoside hydrolase family GH5 and peptidase C48 (SUMO) domains, and fewer chemoreceptor domains. We compared the domain content of the *M. incognita* protein set to those of *C. elegans*, *B. malayi*, *Drosophila melanogaster* and three fungi, of which two are plant pathogens. Thirty-two domains were detected only in *M. incognita*, and two additional domains were only shared between the two plant-pathogenic fungi and *M. incognita*. Functions assigned to the 34 domains specific to plant pathogens encompassed plant cell-wall degradation and chorismate mutase activity (see below). OrthoMCL[15] clustering of the same eight proteomes suggested that 52% of *M. incognita* predicted proteins had no ortholog in the other species. Among them, 1,819 proteins (of which 338 were supported by ESTs) are secreted and lack any known domain (**Supplementary Data**, section 6; **Supplementary Figs. 11** and **12**; **Supplementary Tables 8–10** online). The core complement of proteins in the phylum Nematoda is relatively small: ~23% of the ortholog groups were shared by *M. incognita*, *C. elegans* and *B. malayi* (**Supplementary Fig. 12b**).

Identifying plant parasitism genes

Nematode proteins produced in and secreted from specialized gland cells into the host are likely to be important effectors of plant parasitism[4,16]. We identified gene products that might be involved in parasitic interaction, particularly those that might modify plant cell walls.

M. incognita has an unprecedented set of 61 plant cell wall–degrading, carbohydrate-active enzymes (CAZymes). Although a few such individual CAZymes had been identified previously in some plant-parasitic nematodes and in two insect species[4,16,17], they are absent from all other metazoans studied to date (**Table 2**; **Supplementary Data**, section 7.1; **Supplementary Tables 11–14** online). We identified 21 cellulases and six xylanases from family GH5, two polygalacturonases from family GH28 and 30 pectate lyases from family PL3. We also identified CAZymes not previously reported from metazoans, including two additional plant cell wall–degrading arabinases (family GH43) and two invertases (family GH32). Invertases catalyze the conversion of sucrose (an abundant disaccharide in plants) into glucose and fructose, which can be used by *M. incognita* as a carbon source. We also identified a total of 20 candidate expansins in *M. incognita*, which may disrupt noncovalent bonds in plant cell walls, making the components more accessible to plant cell wall–degrading enzymes[18]. This suite of plant cell wall–degrading CAZymes, expansins and associated invertases was probably acquired by horizontal gene transfer (HGT), as the most similar proteins (outside plant-parasitic nematodes) were bacterial homologs (**Supplementary Table 12**). *M. incognita* also has four secreted chorismate mutases[19], which most closely resemble bacterial enzymes. Chorismate mutase is a key enzyme in biosynthesis of aromatic amino acids and related products, and *M. incognita* may subvert host tyrosine-dependant lignification or defense responses.

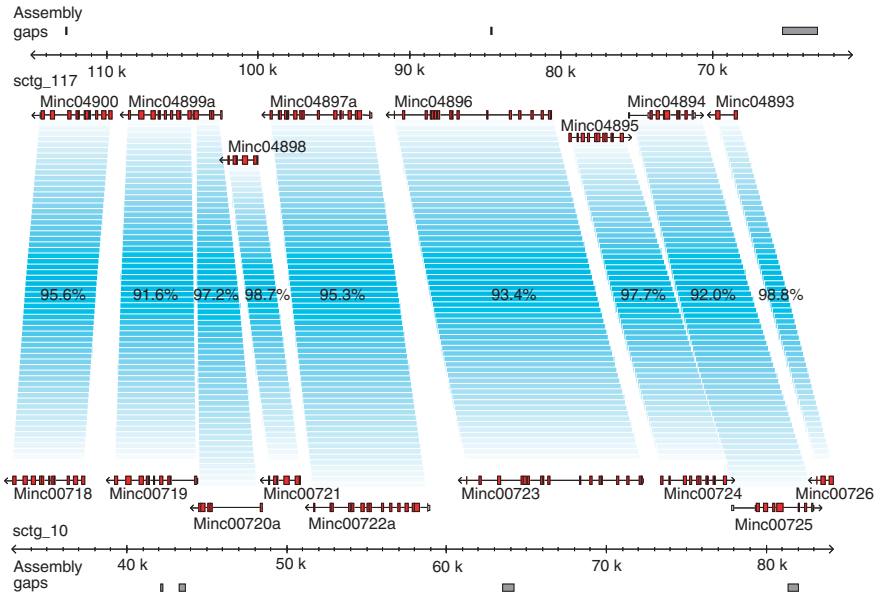


Figure 3 Example of two allelic-like regions in the *Meloidogyne incognita* assembly. Exons are represented by red boxes and are linked together to form genes (arrows indicate the direction of transcription). Gray boxes show assembly gaps. Highly diverged allelic genes are linked together using blue boxes. Gene order is well conserved between the two allelic-like regions, with only minor differences in predicted gene structure. Percentages of sequence identity at the protein level between the two allelic-like regions are indicated.

Overall, these genes suggest a critical role of HGT events in the evolution of plant parasitism within root-knot nematodes.

Apart from genes restricted to *M. incognita*, we also identified gene families showing substantial expansion compared to *C. elegans*. Among the most notable idiosyncrasies in *M. incognita*, we identified more than 20 cysteine proteases of the C48 SUMO (small ubiquitin-like modifier) deconjugating enzyme family—four times the number in *C. elegans* (**Supplementary Data**, section 7.2; **Supplementary Table 15** online). As some phytopathogenic bacterial virulence factors are SUMO proteases[20], the proteolysis of sumoylated host substrates may be a general strategy used by pathogens to manipulate host plant signal transduction. The *M. incognita* genome also encodes nine serine proteases from the S16 sub-family (Lon proteases), whereas only three are identified in *C. elegans*. These proteases regulate type III protein secretion in phytopathogenic bacteria[21] and may have analogous roles in *M. incognita*.

We identified orthologs to other known candidate plant-parasitic nematode parasitism genes in the genome of *M. incognita*. As most of these gene families are also present in animal-parasitic nematodes and *C. elegans*, *M. incognita* members putatively involved in parasitism were probably recruited from ancestral nematode families (**Supplementary Data**, section 7.3; **Supplementary Table 16** online). Twenty-seven previously described *M. incognita*–restricted pioneer genes expressed in esophageal glands[22] were retrieved in the genome. Eleven additional copies were identified; all remain *Meloidogyne spp.* specific (**Supplementary Data**, section 7.4; **Supplementary Table 17** online). These secreted proteins of as-yet-unknown function are likely targets for novel intervention strategies, and warrant deeper investigation.

Protection against environmental stresses

One aspect of plant defense responses is the production of cytotoxic oxygen radicals. However, *M. incognita* has fewer genes encoding superoxide dismutases and glutathione peroxidases than *C. elegans* (**Supplementary Data**, section 7.5; **Supplementary Table 18** online). More striking still was the reduction in glutathione *S*-transferases (GSTs) and cytochromes P450 (CYPs), enzymes involved in xenobiotic metabolism and protection against peroxidative damage. Whereas *C. elegans* has 44 GSTs, including representatives from the Omega, Sigma and Zeta classes[23], *M. incognita* possesses only 5 GSTs, all from the Sigma class. Sigma class GSTs are involved in protection against oxidants rather than xenobiotics. A comparable reduction in *gst* genes was observed in *B. malayi*[7]. Similarly, whereas *C. elegans* has 80 different *cyp* genes from 16 families[24], only 27 full or partial *cyp* genes, from 8 families, were identified in *M. incognita*. CYP35 and other families of xenobiotic-metabolizing P450s are absent from *M. incognita* (**Supplementary Data**, section 7.5; **Supplementary Table 18**).

We identified *M. incognita* orthologs of all genes of the innate immunity signaling pathways of *C. elegans*[25] except *trf-1*, which is part of the Toll pathway (**Supplementary Data**, section 7.5; **Supplementary Table 19** online). However, immune effectors such as lysozymes, C-type lectins and chitinases were much less abundant in *M. incognita* than in *C. elegans*. As previously observed in *B. malayi*[7], entire classes of immune effectors known from *C. elegans* were absent from *M. incognita*, including antibacterial genes such as *abf* and *spp*[26] and antifungal genes of several classes (*nlp*, *cnc*, *fip*, *fipr*)[25] (**Supplementary Data**, section 7.5; **Supplementary Table 19**). As plant parasites embedded in root tissues are protected from a variety of biotic and abiotic stresses, we speculate that the reduction and specialization of chemical and immune defense genes is a result of life in this privileged environment.

C. elegans has a broad range of unusual fucosylated N-glycan structures compared to other metazoans[27]. *M. incognita* has almost twice as many candidate fucosyltransferases as *C. elegans* (**Supplementary Data**, section 7.1; **Supplementary Table 14**). As suggested for animal-parasitic nematodes, multi-fucosylated structures on the surface of the nematode cuticle could help *M. incognita* to evade recognition[27].

Table 2 *Meloidogyne incognita* enzymes with predicted plant cell wall–degrading activities, compared with those in *C. elegans* and *D. melanogaster*

Substrate	Cellulose	Xylan	Arabinan	Pectin		Other	
Family	GH5 (cel)	GH5 (xyl)	GH43	GH28	PL3	EXPN	Total
M. incognita	21	6	2	2	30	20	81
C. elegans	0	0	0	0	0	0	0
D. melanogaster	0	0	0	0	0	0	0

Number of genes encoding enzymes with candidate activity on different substrate is listed in the three selected species. GH, glycoside hydrolases; PL, polysaccharide lyases; EXPN, expansin-like proteins, following the CAZy nomenclature (http://www.cazy.org/). A total of nine and two cellulose-binding modules of family CBM2 (bacterial type) were found appended to candidate expansins and cellulases, respectively.

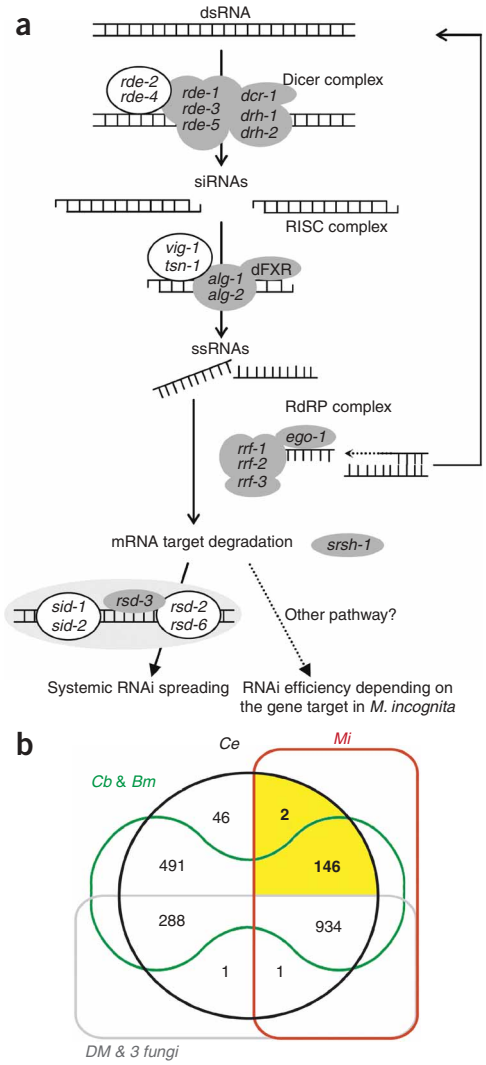


Figure 4 RNAi pathway and lethal targets. (**a**) Comparison of the RNAi pathway genes of *C. elegans* and *M. incognita*. A gray background indicates that at least one homologous gene was found in *M. incognita*, and a white background indicates that no homologous gene was found in *M. incognita*. (**b**) Distribution of orthologs to *C. elegans* lethal RNAi genes (Ce, black) between *M. incognita* (Mi, red), *C. briggsae* and *B. malayi* (Cb & Bm, green), *D. melanogaster* and three fungi, *N. crassa*, *G. zea* and *M. grisea* (Dm & 3 fungi, gray) using OrthoMCL. A yellow background indicates 148 nematode-only gene clusters.

Core biological processes

Nuclear receptors, kinases, G-protein coupled receptors (GPCRs) and neuropeptides encompass some of the gene products most extensively involved in core physiological, developmental and regulatory processes.

C. elegans has a surprisingly large number of nuclear receptors, but curiously lacks orthologs of many nuclear receptor types conserved in other animals[28]. Some of these conserved nuclear receptors are present in *B. malayi*[7]. Among the 92 predicted nuclear receptors in *M. incognita*, we identified orthologs of several known nematode nuclear receptors, although many of the nuclear receptors present in *B. malayi* and absent in *C. elegans* were also absent in *M. incognita* (**Supplementary Data**, section 7.6; **Supplementary Table 20** online). Many *C. elegans* nuclear receptors are classified as supplementary nuclear receptors (SupNRs), likely derived from a hepatocyte nuclear factor-4-like ancestor[29]. Orthologs of SupNRs were found in *M. incognita*, including a 41-member, *M. incognita*-specific expansion. Fourteen SupNRs are one-to-one orthologs between *B. malayi*, *M. incognita* and *C. elegans*, or conserved only between *M. incognita* and *C. elegans*, with secondary losses in *B. malayi* (**Supplementary Data**, section 7.6; **Supplementary Fig. 13** online). Thus the expansion of SupNRs started before the *Brugia-Meloidogyne-Caenorhabditis* split and has proceeded independently in *C. elegans* and *M. incognita*.

M. incognita has 499 predicted kinases compared to 411 in *C. elegans*[30] and 215 in *B. malayi*[7]. The kinases were grouped into 232 OrthoMCL clusters, 24 of which contained only nematode members, suggesting that they have nematode-specific functions. Four kinase families contained only *M. incognita* and *B. malayi* members, suggesting potential roles for these genes in parasitism. Finally, 66 kinase families, containing 122 genes, appear to be *M. incognita*-specific (**Supplementary Data**, section 7.7; **Supplementary Table 21** online). Seven percent (1,280) of all *C. elegans* genes are predicted to encode GPCRs that play crucial roles in chemosensation. These *C. elegans* genes have been divided into three serpentine receptor superfamilies and five solo families[31]. *M. incognita* has only 108 GPCR genes and these derive from two of the three serpentine receptor superfamilies and one of the solo families. These *M. incognita* chemosensory genes are commonly found as duplicates clustered on the genome, as observed in *C. elegans* (**Supplementary Data**, section 7.8; **Supplementary Fig. 14**; **Supplementary Table 22** online).

Neuropeptide diversity is remarkably high in nematodes, given the structural simplicity of their nervous systems. *C. elegans* has 28 Phe-Met-Arg-Phe-amide-like peptide (*flp*) and 35 neuropeptide-like protein (*nlp*) genes encoding ~200 distinct neuropeptides[32]. The identified neuropeptide complement of *M. incognita* is smaller: 19 *flp* genes and 21 *nlp* genes. However, two *flp* genes, *Mi-flp-30* and *Mi-flp-31*, encode neuropeptides that have not been identified in *C. elegans*, suggesting that they could fulfill functions specific to a phytoparasitic lifestyle (**Supplementary Data**, section 7.9; **Supplementary Table 23** online).

The XX-XO sex determination pathway in *C. elegans* is intimately linked to the dosage compensation pathway[33]. *M. incognita* reproduces exclusively by mitotic parthenogenesis, and males do not contribute genetically to production of offspring[11]. *M. incognita* also displays an environmental influence on sex determination: under less favorable environmental conditions far more males are produced. These males can arise due to sex reversal[34] and intersexual forms can be produced. *M. incognita* homologs of at least one member of each step of the *C. elegans* sex determination cascade were identified, including *sdc-1* from the dosage compensation pathway, *tra-1*, *tra-3* and *fem-2* from the sex determination pathway itself, and also downstream genes such as *mag-1* (which represses male-promoting genes) and *mab-23* (which controls male differentiation and behavior). In addition, a large family (~35 genes) of *M. incognita* secreted proteins, similar to the C2H2 zinc finger motif–containing *tra-1* from *C. elegans*, was identified (**Supplementary Data**, section 7.10; **Supplementary Table 24** online). It is therefore possible that *M. incognita* uses a similar genetic system for sex determination, but with the male pathway also modulated in response to environmental cues.

Taken together, these comparative analyses of genes, underpinning important traits, highlight the huge biodiversity in the phylum Nematoda. Idiosyncrasies identified in *M. incognita* may account for

its parasitic lifestyle and lead to the development of new control strategies directed against plant-parasitic nematodes.

RNA interference and lethal phenotypes

RNA interference (RNAi) is a promising technology for the functional analysis of parasitic nematode genes. RNAi can be induced in *M. incognita* by feeding, with variable silencing efficiencies depending on the gene target[35,36]. *M. incognita* has many genes of the *C. elegans* RNAi pathway, including components of the amplification complex (*ego-1*, *rrf-1*, *rrf-2* and *rrf-3*). However, we found no homologs of *sid-1*, *sid-2*, *rsd-2* and *rsd-6*, which are genes involved in systemic RNAi and double-stranded RNA spreading to surrounding cells (**Fig. 4**, **Supplementary Data**, section 7.11; **Supplementary Table 25** online). These genes are also absent from *B. malayi*[7] and *Haemonchus contortus*[37], suggesting that systematic RNAi may spread through the action of novel or poorly conserved factors. We retrieved 2,958 *C. elegans* genes having a lethal RNAi phenotype and searched for orthologs in *M. incognita*. Among the 1,083 OrthoMCL families identified, 148 (containing 344 *M. incognita* genes) appear to be nematode specific (**Supplementary Data**, section 7.12). Because of their lethal RNAi phenotype and distinctive sequence properties, these genes provide an attractive set of new antiparasite drug targets.

DISCUSSION

The genome of *M. incognita* has many traits that render it particularly attractive for studying the fundamentals of plant parasitism in the Nematoda. One remarkable feature is that most of the genome is composed of pairs of homologous segments that may denote former diverged alleles. This suggests that *M. incognita* is evolving without sex toward effective haploidy through the Meselson effect[38–40]. As the *M. incognita* genome is the first one sequenced and assembled for a strictly parthenogenetic species, we expect that its comparison with sexual nematode genomes will shed light on mechanisms leading to its peculiar structure. Functional divergence between ancient alleles of genes involved in the host-parasite interface could explain the extremely wide host range and geographic distribution of this polyphagous nematode. Analysis of the gene content of *M. incognita* revealed a suite of plant cell wall–degrading enzymes, which has no equivalent in any animal studied to date. The striking similarity of these enzymes to bacterial homologs suggests that these genes were acquired by multiple HGT events. Just as many instances of bacterial HGT involve sets of genes implicated in adaptations to new hosts or food sources, the candidate HGT events in *M. incognita* involve genes with potential roles in interactions with hosts. The alternative hypothesis—that these genes were acquired vertically from a common ancestor of bacteria and nematodes and lost in most eukaryote lineages—appears less parsimonious. Other singularities encompass *M. incognita*-restricted secreted proteins or lineage-specific expansions and/or reductions that may play roles in host-parasite interaction.

Transcriptional profiling, proteomic analysis and high throughput RNAi strategies are in progress and will lead to a deeper understanding of the processes by which a nematode causes plant disease. Combining such knowledge with functional genomic data from the model host plant *A. thaliana* should provide new insights into the intimate molecular dialog governing plant-nematode interactions and allow the further development of target-specific strategies to limit crop damage. Through the use of comparative genomics, the availability of free-living, animal- and plant-parasitic nematode genomes should provide new insights into parasitism and niche adaptation.

METHODS

Strain and DNA extraction. We used the *M. incognita* strain 'Morelos' from the root-knot nematode collection held at INRA (Institut National de la Recherche Agronomique) Sophia Antipolis, France. Nematode eggs were collected in a sterile manner from tomato roots and checked for the presence of plant material contaminants. DNA was extracted as described in **Supplementary Methods**, section 8.1 online.

Genome sequencing and assembly. We obtained paired-end sequences from plasmid and BAC libraries with the Sanger dideoxynucleotide technology on ABI3730xl DNA analyzers. The 1,000,873 individual reads were assembled in 2,817 supercontigs using Arachne[8] (**Supplementary Methods**, section 8.2; **Supplementary Table 26** online).

Genome structure, operons and noncoding elements. The assembled genome was searched for repetitive and non-coding elements. Scaffolds were aligned to determine pairs and triplets of allelic-like regions. Gene positions along scaffolds were used to predict clusters of genes forming putative operons (**Supplementary Methods**, section 8.3–8.7).

Prediction of protein coding genes. Gene predictions were performed using EuGene[14], optimized for *M. incognita* models and tested on a data set of 230 nonredundant, full-length cDNAs. Translation starts and splice sites were predicted by SpliceMachine[41]. Available *M. incognita* ESTs were aligned on the genome using GenomeThreader[42]. Similarities to *C. elegans* and other species' protein, genome and EST sequences were identified using BLAST[43]. Repetitive sequences were masked using RepeatMasker (http://repeatmasker.org/, **Supplementary Methods**, section 8.8; **Supplementary Fig. 15** online).

Automatic functional annotation. Protein domains were searched with InterproScan[44]. We also submitted proteins from seven additional species to the same InterproScan search. We included three other nematodes (*C. elegans*, *C. briggsae* and *B. malayi*), the fruitfly (*D. melanogaster*) and three fungi (*Magnaporthe grisea, Gibberella zea* and *Neurospora crassa*). To identify clusters of orthologous genes between *M. incognita* and the seven additional species, we used OrthoMCL[15] (**Supplementary Methods**, section 8.9).

Expert functional annotation. The collection of predicted protein coding genes was manually annotated by a consortium of laboratories. Each laboratory focused on a particular process or gene family relevant to the different aspects of *M. incognita* biology. Patterns of presence and/or absence and expansion and/or reduction in comparison to *C. elegans*, and other species were examined. The quality of predicted genes was manually checked and a functional annotation was proposed accordingly (**Supplementary Methods**, sections 8.10–8.20). A genome browser and additional information on the project are available from http://meloidogyne.toulouse.inra.fr/.

Accession codes. The 9,538 contigs resulting from the *Meloidogyne incognita* genome assembly and annotation were deposited in the EMBL/Genbank/DDBJ databases under accession numbers CABB01000001–CABB01009538.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

SCRI laboratory (V.C.B. and J.T.J.) received funding from the Scottish Government. This work benefited from links funded via COST Action 872. G.V.M. and V.L. are supported by ARC, CNRS, EMBO, MENRT and Region Rhone-Alpes. G.V.M., M.R.-R. and V.L. are also funded by the EU Cascade Network of Excellence and the integrated project Crescendo. M.-C.C. is supported by MENRT. We thank Philippe Lecomte for critical reading of the manuscript and all our collaborators from the "Plant-Nematode interaction" team of INRA Sophia Antipolis for technical help and support.

AUTHOR CONTRIBUTIONS

P.A. and J.G. contributed equally as first authors. J.-M.A., P.C.-S., E.G.J.D., E.D. and L.P.-B. contributed equally as second authors. T.J.B., M.B., T.B-Z., E.L.D., J.J.E., B.F., E.G., B.H., J.T.J., V.L., A.G.M., H.Q., M.-N.R., T.S., G.S., J.W. and P.W. contributed equally as senior authors. P.A., M.B., P.C.-S. and E.G.J.D. wrote the manuscript with input from J.T.J. and A.G.M. For biological material,

contributions were as follows. F.D., M.M. and L.P.-B. for strain growth, control and selection and DNA extraction. P.A., M.-C.C., F.D., E.D., B.F., M.-N.R. and L.P.-B. for cDNA libraries and EST data. For genome sequencing and assembly, contributions were as follows. B.S., E.U., J.P., V.A. for sequencing. C.J. for assembly. C.D. for cDNA clustering and library analyses. J.-M.A., O.J., C.J., F.A. for bioinformatics of allelism characterization. J.W. and P.W. supervision and coordination of the sequencing. For genome structure and organization, contributions were as follows. P.C.-S., T.F., H.Q. and D.S. for repetitive and transposable elements. J.G., E.S. for rRNAs, tRNAs, miRNAs. M.B. for operonic structures. M.-N.R., E.S. and C.V.G. for splice leaders (SL). For *in-silico* global genome analysis, contributions were as follows. E.D., J.G. and T.S. for gene predictions, automatic functional annotation, databases and bioinformatics. E.D. and B.F. for global protein set comparative analysis. Proteome expert annotation was as follows: P.M.C., E.G.J.D. and B.H., for Carbohydrate-Active enZymes. P.C.-S. and E.G. for proteases. M.-C.C., E.L.D., M.E., B.F., E.G.J.D., E.D., E.G., J.T.J., N.H., L.P.-B., G.S. and T.T. for candidate nematode parasitism and pioneer genes. P.A., T.B.-Z., E.G.J.D., E.D., J.J.E., J.V.G., G.P. and M.-N.R. for protection against plant defenses and immune system. V.L., G.V.M. and M.R.-R. for nuclear receptors. T.J.B., T.H. and T.R.M. for the kinome. E.G.J.D. and L.P.-B. for GPCRs. T.B.-Z., F.D.L., P.L. and P.V. for collagen. A.G.M. and P.M.V. for neuropeptides. J.T.J. for sex determination. V.C.B., E.G.J.D. and L.P.-B. for RNAi pathway and lethal RNAi phenotypes.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Trudgill, D.L. & Blok, V.C. Apomictic, polyphagous root-knot nematodes: exceptionally successful and damaging biotrophic root pathogens. *Annu. Rev. Phytopathol.* **39**, 53–77 (2001).
2. Blaxter, M.L. Nematoda: genes, genomes and the evolution of parasitism. *Adv. Parasitol.* **54**, 101–195 (2003).
3. Caillaud, M.C. *et al.* MAP65–3 Microtubule-associated protein is essential for nematode-induced giant cell ontogenesis in *Arabidopsis*. *Plant Cell* **20**, 423–437 (2008).
4. Caillaud, M.C. *et al.* Root-knot nematodes manipulate plant cell functions during a compatible interaction. *J. Plant Physiol.* **165**, 104–113 (2008).
5. The C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
6. Stein, L.D. *et al.* The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol.* **1**, E45 (2003).
7. Ghedin, E. *et al.* Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* **317**, 1756–1760 (2007).
8. Jaffe, D.B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–96 (2003).
9. Leroy, S., Duperray, C. & Morand, S. Flow cytometry for parasite nematode genome size measurement. *Mol. Biochem. Parasitol.* **128**, 91–93 (2003).
10. Triantaphyllou, A.C. in *An Advance Treatise on Meloidogyne* vol. 1 (eds. Sasser, J.N. & Carter, C.C.) 113–126, (North Carolina State University Graphics, Raleigh, USA, 1985).
11. Castagnone-Sereno, P. Genetic variability and adaptive evolution in parthenogenetic root-knot nematodes. *Heredity* **96**, 282–289 (2006).
12. Mark Welch, D.B., Cummings, M.P., Hillis, D.M. & Meselson, M. Divergent gene copies in the asexual class Bdelloidea (Rotifera) separated before the bdelloid radiation or within bdelloid families. *Proc. Natl. Acad. Sci. USA* **101**, 1622–1625 (2004).
13. Piotte, C., Castagnone-Sereno, P., Bongiovanni, M., Dalmasso, A. & Abad, P. Cloning and characterization of two satellite DNAs in the low-C-value genome of the nematode *Meloidogyne* spp. *Gene* **138**, 175–180 (1994).
14. Foissac, S. & Schiex, T. Integrating alternative splicing detection into gene prediction. *BMC Bioinformatics* **6**, 25 (2005).
15. Li, L., Stoeckert, C.J. Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
16. Davis, E.L., Hussey, R.S. & Baum, T.J. Getting to the roots of parasitism by nematodes. *Trends Parasitol.* **20**, 134–141 (2004).
17. Wei, Y.D. *et al.* Molecular cloning, expression, and enzymatic activity of a novel endogenous cellulase from the mulberry longicorn beetle, *Apriona germari*. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **145**, 220–229 (2006).
18. Qin, L. *et al.* Plant degradation: a nematode expansin acting on plants. *Nature* **427**, 30 (2004).
19. Lambert, K.N., Allen, K.D. & Sussex, I.M. Cloning and characterization of an esophageal-gland-specific chorismate mutase from the phytoparasitic nematode *Meloidogyne javanica*. *Mol. Plant Microbe Interact.* **12**, 328–336 (1999).
20. Hotson, A. & Mudgett, M.B. Cysteine proteases in phytopathogenic bacteria: identification of plant targets and activation of innate immunity. *Curr. Opin. Plant Biol.* **7**, 384–390 (2004).
21. Tang, X., Xiao, Y. & Zhou, J.M. Regulation of the type III secretion system in phytopathogenic bacteria. *Mol. Plant Microbe Interact.* **19**, 1159–1166 (2006).
22. Huang, G. *et al.* A profile of putative parasitism genes expressed in the esophageal gland cells of the root-knot nematode *Meloidogyne incognita*. *Mol. Plant Microbe Interact.* **16**, 376–381 (2003).
23. Lindblom, T.H. & Dodd, A.K. Xenobiotic detoxification in the nematode *Caenorhabditis elegans*. *J. Exp. Zoolog. A Comp. Exp. Biol.* **305**, 720–730 (2006).
24. Menzel, R., Bogaert, T. & Achazi, R. A systematic gene expression screen of *Caenorhabditis elegans* cytochrome P450 genes reveals CYP35 as strongly xenobiotic inducible. *Arch. Biochem. Biophys.* **395**, 158–168 (2001).
25. Ewbank, J.J. Signaling in the immune response. *WormBook* doi/10.1895/wormbook.1.83.1, <http://www.wormbook.org/> (2006).
26. Alegado, R.A. & Tan, M.W. Resistance to antimicrobial peptides contributes to persistence of Salmonella typhimurium in the *C. elegans* intestine. *Cell Microbiol.* **10**, 1259–1273(2008).
27. Paschinger, K., Gutternigg, M., Rendic, D. & Wilson, I.B. The N-glycosylation pattern of *Caenorhabditis elegans*. *Carbohydr. Res.* **343**, 2041–2049 (2007).
28. Bertrand, S. *et al.* Evolutionary genomics of nuclear receptors: from 25 ancestral genes to derived endocrine systems. *Mol. Biol. Evol.* **21**, 1923–1937 (2004).
29. Robinson-Rechavi, M., Maina, C.V., Gissendanner, C.R., Laudet, V. & Sluder, A. Explosive lineage-specific expansion of the orphan nuclear receptor HNF4 in nematodes. *J. Mol. Evol.* **60**, 577–586 (2005).
30. Plowman, G.D., Sudarsanam, S., Bingham, J., Whyte, D. & Hunter, T. The protein kinases of *Caenorhabditis elegans*: a model for signal transduction in multicellular organisms. *Proc. Natl. Acad. Sci. USA* **96**, 13603–13610 (1999).
31. Robertson, H.M. & Thomas, J.H. The putative chemoreceptor families of *C. elegans*. *WormBook* doi/10.1895/wormbook.1.66.1, <http://www.wormbook.org/> (2006).
32. Marks, N.J. & Maule, A.G. in *Neuropeptide Systems as Targets for Parasite and Pest Control* (eds. Geary, T.G. & Maule, A.G.) (Landes Bioscience/Eurekah.com, Austin, TX, USA, 2008).
33. Zarkower, D. Somatic sex determination. *WormBook* doi/10.1895/wormbook.1.84.1, <http://www.wormbook.org/> (2006).
34. Papadopoulou, J. & Triantaphyllou, A.C. Sex-determinantion in *Meloidogyne incognita* and anatomical evidence of sexual reversal. *J. Nematol.* **14**, 549–566 (1982).
35. Rosso, M.N., Dubrana, M.P., Cimbolini, N., Jaubert, S. & Abad, P. Application of RNA interference to root-knot nematode genes encoding esophageal gland proteins. *Mol. Plant Microbe Interact.* **18**, 615–620 (2005).
36. Huang, G., Allen, R., Davis, E.L., Baum, T.J. & Hussey, R.S. Engineering broad root-knot resistance in transgenic plants by RNAi silencing of a conserved and essential root-knot nematode parasitism gene. *Proc. Natl. Acad. Sci. USA* **103**, 14302–14306 (2006).
37. Zawadzki, J.L., Presidente, P.J., Meeusen, E.N. & De Veer, M.J. RNAi in Haemonchus contortus: a potential method for target validation. *Trends Parasitol.* **22**, 495–499 (2006).
38. Birky, C.W. Jr. Bdelloid rotifers revisited. *Proc. Natl. Acad. Sci. USA* **101**, 2651–2652 (2004).
39. Mark Welch, D. & Meselson, M. Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. *Science* **288**, 1211–1215 (2000).
40. Mark Welch, D.B., Mark Welch, J.L. & Meselson, M. Evidence for degenerate tetraploidy in bdelloid rotifers. *Proc. Natl. Acad. Sci. USA* **105**, 5145–5149 (2008).
41. Degroeve, S., Saeys, Y., De Baets, B., Rouze, P. & Van de Peer, Y. SpliceMachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics* **21**, 1332–1338 (2005).
42. Gremme, G., Brendel, V., Sparks, M.E. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* **47**, 965–978 (2005).
43. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
44. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116–W120 (2005).
45. Hillier, L.W. *et al.* Genomics in *C. elegans*: so many genes, such a little worm. *Genome Res.* **15**, 1651–1660 (2005).

Supplementary Data and Methods for: Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*

Pierre Abad[1,2,3], Jérôme Gouzy[4], Jean-Marc Aury[5,6,7], Philippe Castagnone-Sereno[1,2,3], Etienne G.J. Danchin[1,2,3], Emeline Deleury[1,2,3], Laetitia Perfus-Barbeoch[1,2,3], Véronique Anthouard[5,6,7], François Artiguenave[5,6,7], Vivian C. Blok[8], Marie-Cécile Caillaud[1,2,3], Pedro M. Coutinho[9], Corinne Dasilva[5,6,7], Francesca De Luca[10], Florence Deau[1,2,3], Magali Esquibet[11], Timothé Flutre[12], Jared V. Goldstone[13], Noureddine Hamamouch[14], Tarek Hewezi[15], Olivier Jaillon[5,6,7], Claire Jubin[5,6,7], Paola Leonetti[10], Marc Magliano[1,2,3], Tom R. Maier[15], Gabriel V. Markov[16,17], Paul McVeigh[18], Graziano Pesole[19,20], Julie Poulain[5,6,7], Marc Robinson-Rechavi[21,22], Erika Sallet[23,24], Béatrice Ségurens[5,6,7], Delphine Steinbach[12], Tom Tytgat[25], Edgardo Ugarte[5,6,7], Cyril van Ghelder [1,2,3], Pasqua Veronico[10], Thomas J. Baum[15], Mark Blaxter[26], Teresa Bleve-Zacheo[10], Eric L. Davis[14], Jonathan J. Ewbank[27], Bruno Favery[1,2,3], Eric Grenier[11], Bernard Henrissat[9], John T. Jones[8], Vincent Laudet[16], Aaron G. Maule[18], Hadi Quesneville[12], Marie-Noëlle Rosso[1,2,3], Thomas Schiex[24], Geert Smant[25], Jean Weissenbach[5,6,7], Patrick Wincker[5,6,7]

[1]INRA, UMR 1301, 400 route des Chappes, F-06903 Sophia-Antipolis, France. [2]CNRS, UMR 6243, 400 route des Chappes, F-06903 Sophia-Antipolis, France. [3]UNSA, UMR 1301, 400 route des Chappes, F-06903 Sophia-Antipolis, France. [4]Laboratoire Interactions Plantes Micro-organismes, UMR441/2594, INRA/CNRS, Chemin de Borde Rouge, BP 52627, F-31320 Castanet Tolosan, France. [5]Genoscope (CEA), 2 rue Gaston Crémieux, CP5706, F-91057 Evry, France. [6]CNRS, UMR 8030, 2 rue Gaston Crémieux, CP5706, F-91057 Evry, France. [7]Université d'Evry, F-91057 Evry, France. [8]Plant Pathology Programme, SCRI, Invergowrie, Dundee, DD2 5DA, UK. [9]CNRS, UMR 6098 CNRS and Universites of Aix-Marseille I & II, Case 932, 163 Av. de Luminy, F-13288 Marseille, France. [10]Istituto per la Protezione delle Piante, Consiglio Nazionale delle Ricerche, Via G. Amendola 165/a, 70126, Bari, Italy. [11]INRA, Agrocampus Rennes, Univ. Rennes I, UMR1099 BiO3P, Domaine de la Motte, F-35653 Le Rheu Cedex, France. [12]INRA, UR1164 Unité de Recherche en Génomique et Informatique (URGI), 523 place des terrasses de l'Agora, F-91034 Evry, France. [13]Biology Department, Woods Hole Oceanographic Institution,Co-op Building, MS #16, Woods Hole, Massachusetts 02543, USA. [14]Department of Plant Pathology North Carolina State University, 840 Method Road, Unit 4, Box 7903 Raleigh, North Carolina 27607,USA.

[15]Department of Plant Pathology, Iowa State University, 351 Bessey Hall, Ames, Iowa 50011, USA. [16]Université de Lyon, Institut de Génomique Fonctionnelle de Lyon, Molecular Zoology team, Ecole Normale Supérieure de Lyon, Université Lyon 1, CNRS, INRA, Institut Fédératif 128 Biosciences Gerland, Lyon Sud, 46 allée d'Italie, F-69364 Lyon Cedex 07, France. [17]USM 501 - Evolution des Régulations Endocriniennes, Muséum National d'Histoire Naturelle, 7 rue Cuvier, F-75005 Paris, France. [18]Biomolecular Processes: Parasitology, School of Biological Sciences, Medical Biology Centre, 97 Lisburn Road, Queen's University Belfast, Belfast, BT9 7BL, UK. [19]Dipartimento di Biochimica e Biologia Molecolare "E. Quagliariello", University of Bari, Via Orabona 4, 70126 Bari, Italy. [20]Istituto Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Via G. Amendola, 122/D – 70126 Bari, Italy. [21]Department of Ecology and Evolution, University of Lausanne, UNIL-Sorge, Le Biophore, CH - 1015 Lausanne, Switzerland. [22]Swiss Institute of Bioinformatics, quartier sorge, Bâtiment Genopode, CH - 1015 Lausanne, Switzerland. [23]Plateforme Bioinformatique du Genopole Toulouse Midi-Pyrénées, GIS Toulouse Genopole, 24 Chemin de Borde Rouge, BP 52627, F-31320 Castanet Tolosan, France. [24]Unité de Biométrie et d'Intelligence Artificielle UR875, INRA, Chemin de Borde Rouge, BP 52627, F-31320 Castanet Tolosan, France. [25]Laboratory of Nematology, Wageningen University, Binnenhaven 5, 6709PD Wageningen, The Netherlands. [26]Institute of Evolutionary Biology, University of Edinburgh, Kings Buildings, Ashworth Laboratories, West Mains Road, Edinburgh EH9 3JT, UK. [27]INSERM/CNRS/Université de la Méditerranée, Centre d'Immunologie de Marseille-Luminy, 163 av. de Luminy, Case 906, F-3288, Marseille cedex 09, France. Correspondence should be addressed to P.A. (pierre.abad@sophia.inra.fr).

3 Repetitive and non protein-coding sequences

3.1 Repeats and Transposable Elements

The BLASTER all-by-all comparison of the *M. incognita* genome (first step of the *denovo* pipeline, **Supplementary Methods**, section 8.4) indicated that repeats cover 19% of the genome. This is clearly an underestimate due to the high stringency of this search. At the end of the annotation pipeline, thirty-six percent of the *M. incognita* genome matched consensus sequences for repeats (**Table S2**).

In total, 4,041 different repeat families were detected, from which 3,066 had no visible TE features, and 135 had contradictory characteristics. Note that some of the families with no TE features and contradictory characteristics may correspond to satellite repeats.

Only 690 families had obvious TE features: 210 LTR retroelements, 29 LINEs-like, 13 SINEs-like, 430 TIR transposons and 8 Helitrons. **Table S2** summarizes the statistics for these categories plus the repeats such as SSRs and those without any TE features or contradictory characteristics.

Table S2 | Summary statistics for repeats in *M. incognita* genome

Repeats types	TE class	Number of families	Number of copies	Coverage (bp)	Coverage (% of genome)
LTR retroelement	Class I	210	2,625	2,238,615	2.37
LINE		29	381	268,357	0.31
SINE		13	148	75,296	0.09
TIR transposon	Class II	430	9,725	2,897,931	3.37
Helitron		8	108	201,474	0.23
SSR		150	2,100	541,345	0.63
no TE features		3,066	82,598	22,556,643	26.21
contradictory features		135	2,263	3,019,110	3.51
Total		4,041	99,843	31,601,493	36.72

The highest number of copies for a repeat family was 583. This family consensus sequence is 628bp long with no visible TE feature. Each family had on average 24 copies with a median of 12, indicating a skewed distribution of copy number among repeat families.

The highest number of complete copies, 43, was for another family with a 944 bp long consensus sequence and no obvious TE feature. These copies are considered as complete as their length is at least 95% of the consensus. **Figures S5** and **S6** show the copy numbers for different types of repeat.

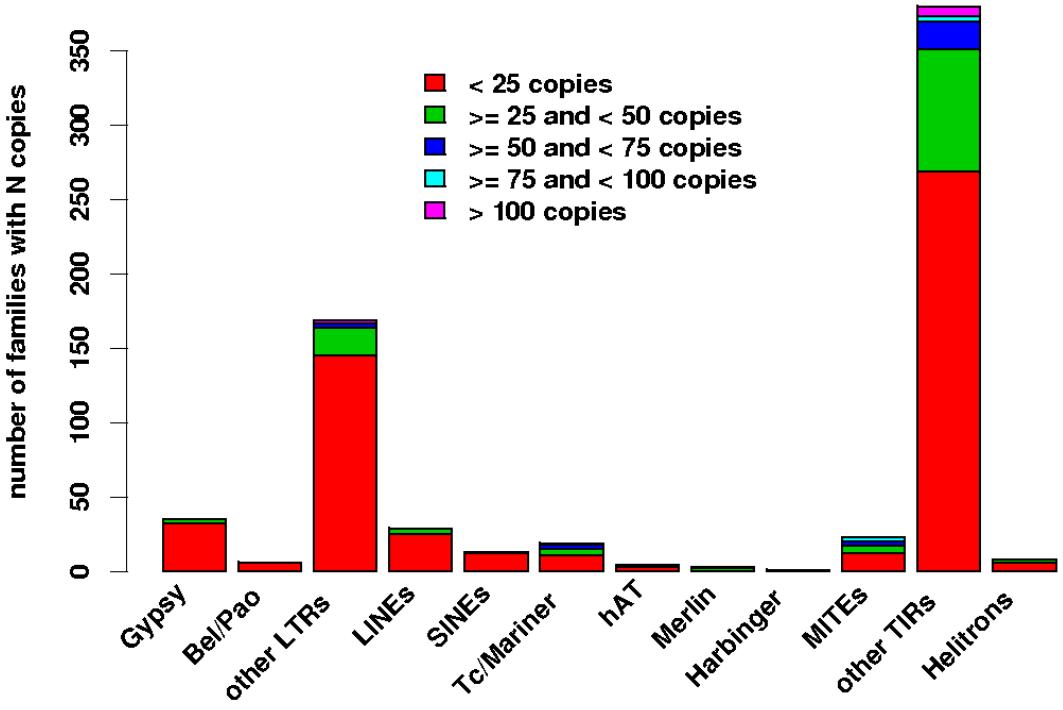


Figure S5 | Distribution of copy number for different super-families of TEs.

In terms of dynamics, one hundred and eleven different families have a minimum copy-to-family consensus identity percentage greater than 95% indicating recent families and possible current activity. All together, they represent 334 copies and cover 0.4% of the genome (**Fig. S6**).

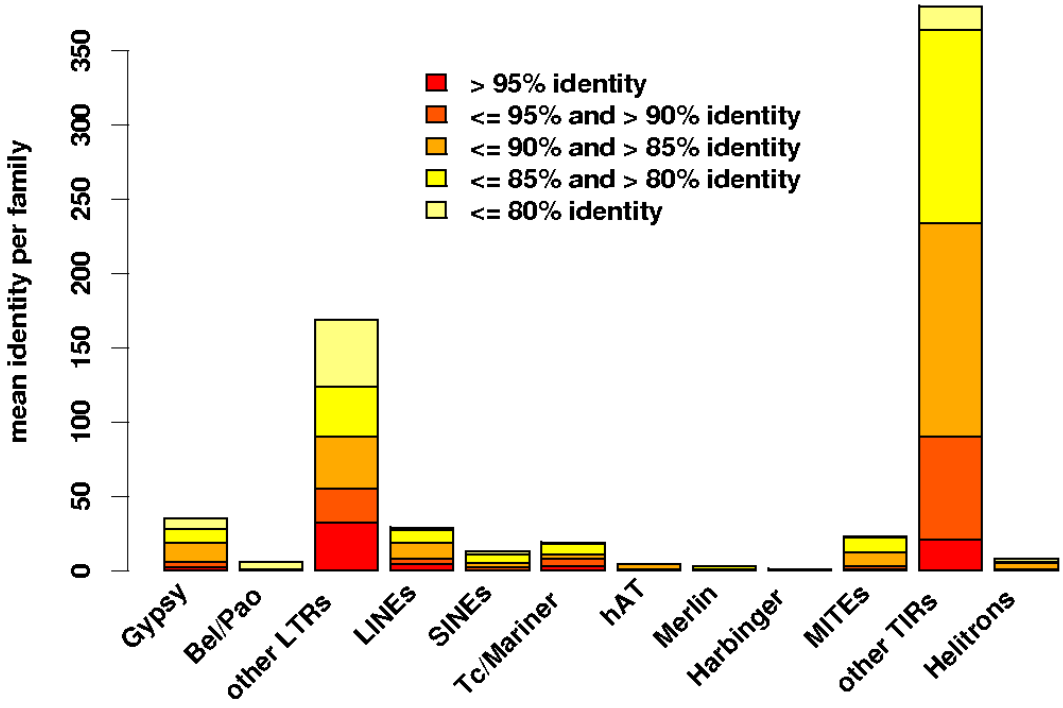


Figure S6 | Distribution of the mean identity between the copies and the consensus for different super-families of TEs.

A.2 ARTICLE "EXTENSIVE SYNTENY CONSERVATION OF HOLOCENTRIC CHROMOSOMES IN LEPIDOPTERA DESPITE HIGH RATES OF LOCAL GENOME REARRANGEMENTS"

Les suppléments ne sont pas inclus car ils ne sont pas indispensables à la compréhension de cet article dans le cadre de ce manuscrit de thèse.

Extensive synteny conservation of holocentric chromosomes in Lepidoptera despite high rates of local genome rearrangements

E. d'Alençon[a,1], H. Sezutsu[b,c,1], F. Legeai[d], E. Permal[e], S. Bernard-Samain[f], S. Gimenez[a], C. Gagneur[a], F. Cousserans[a], M. Shimomura[c], A. Brun-Barale[b], T. Flutre[e], A. Couloux[f], P. East[g], K. Gordon[g], K. Mita[c], H. Quesneville[e], P. Fournier[a], and R. Feyereisen[b,2]

[a]Unité Mixte de Recherche 1231, Institut National de la Recherche Agronomique, Université Montpellier II, 34095 Montpellier, France; [b]Unité Mixte de Recherche 1301, Institut National de la Recherche Agronomique, Centre National de la Recherche Scientifique, Université de Nice Sophia Antipolis, 06903 Sophia Antipolis, France; [c]National Institute of Agrobiological Sciences, Tsukuba 305-8634, Ibaraki, Japan; [d]Unité Mixte de Recherche 1099, Institut National de la Recherche Agronomique, AgroCampus, Institut National de Recherche en Informatique et en Automatique, 35042 Rennes, France; [e]UR1164, Institut National de la Recherche Agronomique Centre de Versailles, Versailles 78026, France; [f]Genoscope, Centre National de Séquençage, 91057 Evry, France; and [g]Commonwealth Scientific and Industrial Research Organisation, Division of Entomology, Canberra, ACT 2601, Australia

Edited* by May R. Berenbaum, University of Illinois, Urbana, IL, and approved March 16, 2010 (received for review September 11, 2009)

The recent assembly of the silkworm *Bombyx mori* genome with 432 Mb on 28 holocentric chromosomes has become a reference in the genomic analysis of the very diverse Order of Lepidoptera. We sequenced BACs from two major pests, the noctuid moths *Helicoverpa armigera* and *Spodoptera frugiperda*, corresponding to 15 regions distributed on 11 *B. mori* chromosomes, each BAC/region being anchored by known orthologous gene(s) to analyze syntenic relationships and genome rearrangements among the three species. Nearly 300 genes and numerous transposable elements were identified, with long interspersed nuclear elements and terminal inverted repeats the most abundant transposable element classes. There was a high degree of synteny conservation between *B. mori* and the two noctuid species. Conserved syntenic blocks of identified genes were very small, however, approximately 1.3 genes per block between *B. mori* and the two noctuid species and 2.0 genes per block between *S. frugiperda* and *H. armigera*. This corresponds to approximately two chromosome breaks per Mb DNA per My. This is a much higher evolution rate than among species of the *Drosophila* genus and may be related to the holocentric nature of the lepidopteran genomes. We report a large cluster of eight members of the aminopeptidase N gene family that we estimate to have been present since the Jurassic. In contrast, several clusters of cytochrome P450 genes showed multiple lineage-specific duplication events, in particular in the lepidopteran CYP9A subfamily. Our study highlights the value of the silkworm genome as a reference in lepidopteran comparative genomics.

comparative genomics | silkworm | Noctuidae | transposable elements | gene clusters

The insect order Lepidoptera is second only to Coleoptera as the most prolific in animal species number, with an estimated total of more than 160,000 species falling into more than 130 families. Assembly of the first lepidopteran genome, that of the domesticated silkworm *Bombyx mori,* has established a valuable reference for lepidopteran comparative genomics and genetics (1). The radiation of the major clades of Lepidoptera occurred in the late Jurassic less than 150 Mya (2, 3). The superfamily Noctuoidea contains approximately one fourth of all Lepidoptera and includes a very large number of major pest species of agriculture and forestry. It has a fossil record dating back to at least 75 Mya (4), which is close to the time of divergence of the superfamily Bombycoidea to which the silkworm belongs. The diversification and proliferation of lepidopteran species is therefore very recent (5). The insights gained from comparative genomic analyses using a reference genome from a model species such as the silkworm would greatly facilitate research on all Lepidoptera, and in particular on selective targets for innovative pest management at a time when competition for food between humans and insects is becoming a critical challenge for a rapidly growing human population. The 432 Mb genome of *B. mori* is the first fully sequenced lepidopteran genome (1), and detailed SNP and BAC-based chromosomal maps are available (6, 7). We wished to compare the silkworm genome at a finer scale with that of major lepidopteran pests from the family Noctuidae, *Helicoverpa armigera* and *Spodoptera frugiperda*. The Old World cotton bollworm *H. armigera* is a highly polyphagous pest and ranks as the world's worst pest of agriculture (8, 9). It is closely related to *Helicoverpa zea*, its New World relative from which it diverged approximately 1.5 Mya (10) and to *Heliothis virescens*, both major pest species in their own right. Similarly, the fall armyworm *S. frugiperda* is a major pest of maize and rice in the Americas and represents a genus comprising many pests worldwide. Conservation of synteny would allow a rapid identification of genes in these pest species from the knowledge of the *Bombyx* genome. The precise definition of conserved segments and of the degree of chromosome rearrangement is more difficult (11–15). Nonetheless, conservation of synteny, when it can be documented, is an extremely useful feature of comparative genomics that validates the use of a model organism and that defines a framework for a finer study of gene and genome evolution.

Overall conservation of synteny in Lepidoptera has been reported in several studies (16–20) suggesting that, over the 100 My separating the *Bombyx* and butterfly lineages, some degree of gene synteny has thus been maintained. However, a more detailed study was warranted to allow more general conclusions. In particular, we asked whether the holocentric nature of the multiple chromosomes in Lepidoptera ($n = 28$ in *Bombyx*) favors scrambling of gene order and masks microsyntenic relationships. The *Bombyx* genome has among the highest level of repetitive sequences [43.6% (1)] of all insect genomes studied to date [vs. *Apis mellifera* with just 1% (21)]. We therefore also asked whether

Author contributions: E.A., K.M., P.F., and R.F. designed research; E.A., S.B.-S., S.G., C.G., and A.C. performed research; F.L., M.S., A.B.-B., T.F., P.E., K.G., and K.M. contributed new reagents/analytic tools; E.A., H.S., F.L., E.P., S.B.-S., F.C., K.G., H.Q., P.F., and R.F. analyzed data; and E.A., H.S., H.Q., P.F., and R.F. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

Data deposition: The BAC sequences reported in this paper have been deposited in the GenBank database (accession nos. FP340404–FP340438; see Table S1 for a list of probes and accession numbers).

[1]E.A. and H.S. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: rfeyer@sophia.inra.fr.

This article contains supporting information online at www.pnas.org/cgi/content/full/0910413107/DCSupplemental.

the high level of repeated sequences would make it difficult to use *Bombyx* as a reference genome for lepidopteran pests.

To answer these questions, we studied the gene landscape and patterns of repetitive sequence distribution, as well as synteny and rearrangements at a very fine scale in 15 genomic regions anchored by orthologous genes in the two noctuid species, *H. armigera* and *S. frugiperda*, and in the silkworm *B. mori*. Our study is based on a detailed analysis of high-quality BAC sequences in the two noctuid species, and on the newly assembled complete genome sequence of the silkworm.

Results

Strategy of the Comparative Genome Analysis. We sequenced and annotated BACs representing 15 putatively homologous regions of the genome in comparison with the fully assembled genome of *B. mori*. We sought to cover genomic regions neighboring genes that vary widely in their evolutionary constraints (Table S1). These regions were covered by 18 BACs in *H. armigera* covering a total of 1.963 Mb of genomic DNA and by 17 BACs in *S. frugiperda* covering a total of 2.042 Mb of genomic DNA. This represents approximately 0.5% each of the two genomes. We estimated the total sequence overlap between the *H. armigera* and *S. frugiperda* sequences to cover 1.22 Mb, with 81.6 ± 8.1 kb of overlap for each pair. The 15 genomic regions were each compared to regions spanning 200 kb of the *B. mori* genome on 15 scaffolds that were distributed on 11 of the 28 chromosomes. The GC content in the three species was similar (32.7–36.3%). We identified 502, 201, and 274 genes in *B. mori*, *H. armigera*, and *S. frugiperda*, respectively (Table S2). We compared both the complete set of presumed orthologues and a limited set of additional genes of the three species (34–46 genes), which were analyzed in greater detail. These genes were selected because of the high degree of confidence in their annotation, i.e., 18 unique genes and 16 to 28 members of multigene families. Table S2 shows that gene length was greater in *B. mori* (6.5–8.0 kb) than in *S. frugiperda* (4.3–4.9 kb) and in *H. armigera* (3.1–4.9 kb). This was a result of a corresponding variation in intron size that was greatest in *B. mori*. The intergenic regions represent 41% of the *B. mori* genome compared with 55% in the noctuids.

Repetitive Sequences and Transposable Elements. The BACs covering the 15 regions of interest and 20, 20, and 55 additional BACs sequences from *H. armigera*, *S. frugiperda*, and *B. mori*, respectively (that are not included in the synteny study) were used for de novo repeated sequences detection using the REPET pipelines (22). The results (Dataset S1 (XLS)) show that the most abundant classes are long interspersed nuclear elements (LINEs) and terminal inverted repeats (TIRs) in the three lepidopteran subgenomes with a quasi-equal number of consensus for both classes in each species. Class II elements predominate in *H. armigera* and *S. frugiperda* with 13 and 11 consensus, respectively, compared with seven and six consensus of class I, whereas elements of class I are the most abundant in the 55 BACs of our *B. mori* subgenome (76 consensus of class I for 59 of class II).

The total proportions of repeated sequences in *H. armigera*, *S. frugiperda*, and *B. mori* were 16.2%, 8.0%, and 33.8%, respectively. This latter value is close to that found for the whole silkworm genome, i.e., 43.6% for repeated sequences and 35.1% for transposable elements (TEs), respectively (1, 23). Among repeated sequences of the TE type, retrotransposons cover 42.5%, 76.5%, and 68.4% of the total TE sequence in *H. armigera*, *S. frugiperda*, and *B. mori*, respectively, with an expansion of some LINE families in *S. frugiperda* and *B. mori*. The TE annotation is available online at the Lepido-DB (*http://www.inra.fr/lepidodb*), and the description of families and their respective distribution in the three species will be the object of a future study.

Conservation of Synteny. ***Quantifying the synteny conservation.*** Syntenic genes were detected and annotated (Fig. S1) in the 15 genomic regions as described below in *Materials and Methods*, and the resulting three-way comparisons are presented in Fig. S2. Syntenic genes can also be visualized online with the Cmap software at the Lepido-DB Web site. Fig. 1 shows an example of such three-way comparisons, with the CYP332A region with a very high degree of synteny conservation.

Taking *B. mori* scaffolds as a reference, we counted for each region the number of genes maintained in the overlapping portion of the *B. mori* scaffolds with either of the noctuid BACs (Table 1 and Dataset S2). Among the 270 genes analyzed, 141 (52.2%) were present in either of the noctuids BACs. Most of the analyzed genes (74.8%) were of "known class," that is, homologous to a gene of known function, having a match to an RNA sequence, or syntenic with a gene of known function; the remaining 25.2% were classified as "HP," i.e., encoding hypothetical proteins. Among the *B. mori* identified known genes, 64.4% were found in both of the noctuids BACs, whereas only 16.2% of the HP genes were syntenic. This

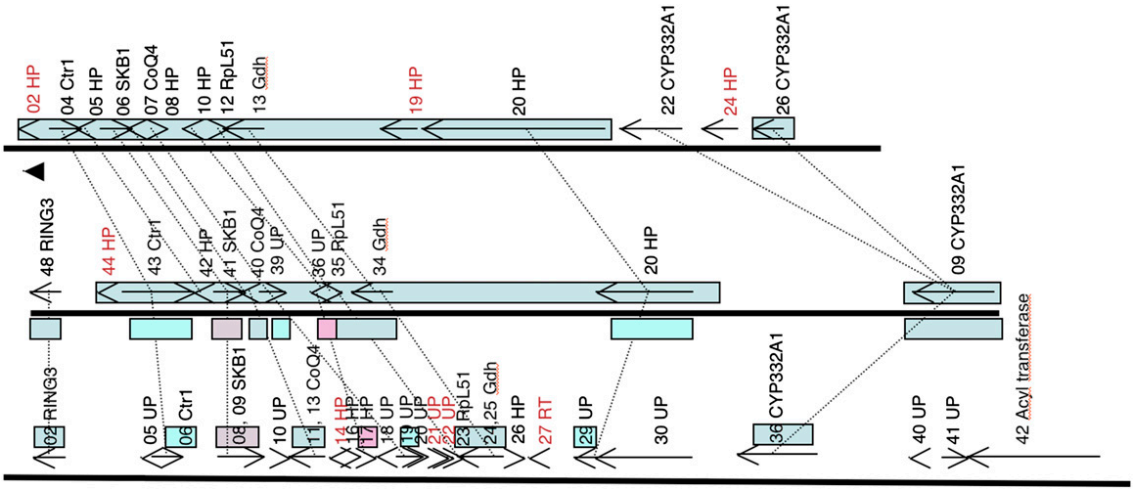


Fig. 1. Syntenic relationship at the CYP332A locus. Schemes at scale representing, from top to bottom, *H. armigera*, *S. frugiperda*, and *B. mori*. Arrows represent genes predicted by KAIKOGAAS. Only valid genes are shown (Fig. S1). Synteny links are shown with black dotted lines. In text boxes, gene ID, or for other genes, HP, unknown protein (UP; presence of a match to an EST with a threshold of 10^{-40} by BlastN). Synteny blocks (as defined in text) are shown as colored boxes spanning genes arrows in the case of *H. armigera* and S. *frugiperda*, below and above gene arrows in the case of *S. frugiperda* and *B. mori*, respectively. See Fig. S2 for gene name abbreviations.

Table 1. Gene categories with *B. mori* as reference

Type	Genes with known function (ID)	Genes with identified mRNA (UP)	Partial sum (ID + UP)	Genes encoding hypothetical protein (HP)	Total genes (ID + UP + HP)
Syntenic	113	17	130 (64.4%)	11 (16.2%)	141 (52.2%)
Nonsyntenic	2	70	72 (35.6%)	57 (83.8%)	129 (47.8%)

UP, unknown protein.

suggests that known genes, whose sequence is conserved between species, are located in more stable regions of the genome whereas unconserved genes, some of which may be species-specific, lie in more plastic regions. Twenty-six of 141 genes of *B. mori* (18.3%) that are found in the corresponding BACs from noctuids were in the reverse orientation.

We counted the number of presumed orthologues found in the corresponding *B. mori* region for each of the 15 BAC pairs, irrespective of their order or orientation. This is a restrictive view, because in some cases, genes present on the BACs were found outside the 200-kb *B. mori* region, but within the same scaffold. For example, in the Or83b genomic region of *B. mori* (Fig. S2), the FBPA genes were located 297 kb away from the Or83b gene on the same scaffold. The data (Dataset S3) show that the number of presumed orthologues was very variable, from a minimum of one (the anchor gene used to select the BAC) to 16 in the TATA binding protein (TBP) region of *S. frugiperda* and 13 in the CYP4M region of *H. armigera*. These variations were roughly correlated with gene density that was not evenly distributed. The TBP region of *S. frugiperda* was characterized by one gene every 4.5 kb, whereas the juvenile hormone acid methyl transferase region had only one gene every 24.6 kb in *H. armigera*. On average, 69.8% and 50.2% of the genes from *H. armigera* and *S. frugiperda,* respectively, were found to be in macrosyntenic conservation, with a median density of eight genes per BAC. In other words, in approximately half the cases, two genes in synteny over 50 kb in *B. mori* would be found within 34 kb of each other in the two noctuid species. This clearly indicates conservation of synteny at a macroscopic scale. We then analyzed the syntenic relationships at a finer scale.

Fine-scale microsynteny and size of synteny blocks. We measured the number and length of synteny blocks, i.e., conserved segment corresponding to any region in which gene content and order are conserved (24). Genes belonging to a TE were excluded. Only validated genes were considered (as detailed earlier). This analysis (Fig. 2) showed that synteny blocks are very small. They spanned 2.01 genes per block on average between the two related noctuids (maximum size of nine genes), and were smaller between *B. mori* and *S. frugiperda* or *H. armigera* (1.29 and 1.32 on average, respectively, with a maximum of four genes). Because of their limited size, synteny blocks were seldom interrupted by lack of sequence extent (i.e., BAC extremities). The average number of genes per BAC or genomic region analyzed was 11 for *H. armigera*, 16 for *S. frugiperda*, and 26 for *B. mori*, far greater than the average block size.

We then examined the recombination events responsible for the synteny breaks. Apart from inversions, we assumed that the presence of *B. mori* genes whose orthologues were not found in the two noctuid species has resulted from transpositions to or from unsequenced parts of these genomes (i.e., beyond the boundaries of the BACs). In addition to orthologues we also analyzed paralogues resulting from duplications in one of the species, taking into account the closest paralogue pairs based on phylogenetic analysis of the encoded proteins. We recorded a ratio of 1.0:8.0:3.4 of inversions, transpositions, and duplications, respectively, between the pest genomes (Table 2). This ratio was 1.0:8.7:3.4 between *B. mori* and *H. armigera* and 1.0:11.4:3.4 between *B. mori* and *S. frugiperda*. We then focused on the overall rate of rearrangements since speciation of the three lepidopteran species. An inversion event results from two double-strand breaks, whereas a transposition requires three (24). The duplications observed within gene families may have resulted from replication slippage or a break-and-join mechanism such as unequal crossing over or transposition and may thus involve zero, two, or three breaks, respectively, and we conservatively counted one break per duplication. The results (Table 3) showed a very high number of breakages. Several genomic regions in our analysis contained variable numbers of duplicated genes from large gene families that may have biased our results by inflating the number of breaks. We therefore corrected for this potential bias by counting only orthologues or best paralogue pairs (Table 3, column 2), by excluding all genes that had undergone duplication in one of the species (Table 3, column 3) or by excluding all BAC regions in which members of large gene families were present (Table 3, column 4). In all three cases, the calculated evolution rate remained very similar. The evolution rate is approximately two breaks per Mb per Mya and has perhaps accelerated within the Noctuidae.

Correlation Between Breakages and TE Density. TEs are often a source of genomic rearrangements or can insert at genomic breakages. We recorded the precise coordinates of synteny breaks between the two noctuid genomes, and counted the number of TE copies by a sliding window of 10 kb along the BACs. In many cases, a clear correlation was evidenced. In the case of the *CYP4M* region (Fig. 3*A*), a clear association of TE copy density was seen with the synteny breaks corresponding to one gene inversion and to the duplications of the *CYP4M* genes. Such an association was found also in the *H. armigera* BAC carrying the gene encoding the ecdysone receptor at the places where duplications of the gene encoding a putative multibinding protein genes have occurred (Fig. 3*B*).

Evolution of Gene Clusters. Several gene clusters were covered by our analysis. The aminopeptidase N (APN) of lepidopteran midgut has received wide attention because of its role in the mode of action

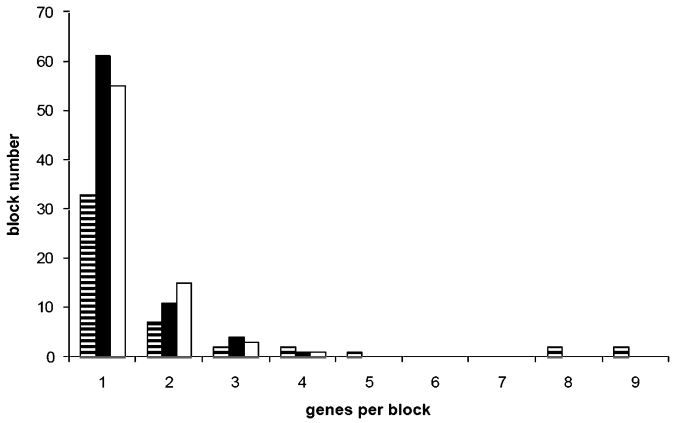


Fig. 2. Number of genes per synteny block. Black and white striped bars: between *H. armigera* and *S. frugiperda* (average size 2.04). Black bars: between *B. mori* and *S. frugiperda* (average size 1.29). White bars: between *B. mori* and *H. armigera* (average size 1.32).

Table 2. Number of rearrangements recorded between species

Species	Inversions	Transpositions	Duplications
Ha-Sf	5	40	17
Bm-Ha	10	87	34
Bm-Sf	9	103	31

of Cry toxins of *Bacillus thurigiensis* (25), and multiple APN cDNAs from a variety of species are available (26). We show here that APN genes are organized in a single, large cluster of nine genes on chromosome 9 of *B. mori*. Remarkably, this cluster is highly conserved in both order and orientation of the genes in the three species (Fig. S3). In addition to seven previously identified APNs (26), the cluster contained two additional genes: *Zn-m1*, a member of the more distantly related protease m1 zinc metalloproteases found in insects and vertebrates, and *APN-8*, related to a fat body–specific transcript previously reported from *Spodoptera litura* (27).

Our analysis also covered five clusters of P450 genes, one of the largest multigene family in insects (28). The *CYP* genes in the five clusters have evolved at a roughly similar rate, as seen by the similar range in overall protein identity of the encoded enzymes (51–68% between *B. mori* and the two noctuids). However, this evolution was punctuated by multiple duplication events in the two noctuid lineages, and possible gene loss in the silkworm lineage, so that we found 17 to 22 *CYP* genes in the noctuids and just nine in *B. mori*. The *CYP9A* cluster on chromosome 17 had the most complex evolutionary history (Fig. 4). We found four *CYP9A* genes in *B. mori*, and there are no additional *CYP9A* genes in the genome. In contrast, we found five *CYP9A* genes in *H. armigera* and nine in *S. frugiperda*. We searched the yet unassembled *H. armigera* genome for other *CYP9A* genes and found that the four *B. mori* genes are monophyletic with the only other *CYP9A* gene that is present in the *H. armigera* genome, but beyond the boundaries of our BAC coverage. This gene, *CYP9A14*, is most closely related to *CYP9A22* of *B. mori*, and to date a closest paralogue in *S. frugiperda* has not been found. Two gene duplications leading to *CYP9A19*, *-20*, and *-21* therefore occurred in the *Bombyx* lineage. The *CYP9A3* and *CYP9A32* genes of the noctuid species are on the opposite strand of the chromosomes respective to the rest of the cluster. They are probable orthologues and their expected third orthologue has probably been lost in the *Bombyx* lineage. The remainder of the *CYP9A* genes in the noctuids resulted from seven gene duplications, of which four occurred in the *Spodoptera* lineage. The rate of evolution varied over time in a lineage-specific manner, and this was probably accompanied by chromosomes rearrangements, as flanking genes have drastically changed in distance and orientation (Fig. 4).

Large-Scale Chromosomal Inversion. In the RpL5A genomic region, the *rpl5A* gene and the three genes immediately downstream of *rpl5A* are in the same orientation in the three species (Fig. S4). In the flanking regions, three known genes are reversed in *B. mori* compared with *S. frugiperda* on one side, and five genes are also reversed in *B. mori* on the other side. The colinearity all along the BACs from noctuids is shown at the nucleotide level by dot plots, which also show the large inversion between *B. mori* and the two noctuids (Fig. S4). This inversion dates before the divergence of the two noctuid species. It is possible that other large-scale inversions were missed in the present study because the size of the BACs may be equal or smaller than the size of such inversions.

Discussion

The assembled sequence for *Bombyx mori* amounts to a 432 Mb genome size, and this size was estimated to 405 Mb for *H. virescens* (29), a species closely related to *H. armigera*. Sequenced genomes show enormous variation in TE copy number, which can largely account for differences in genome size (30). Invasion of the silkworm genome by class I TEs may be responsible for its slightly increased size relative to noctuids. The most abundant classes of TE are LINEs and TIRs in the three genomes. In terms of genome coverage, LINEs predominate in *B. mori* and *S. frugiperda* and contribute equally with TIRs in *H. armigera*. The prevalence of non-LTR retrotransposons makes lepidopteran genomes different from that of *Drosophila melanogaster*, in which LTR retrotransposons are the most abundant (31). Both the rather high repeat coverage and the prevalence of non-LTR retrotransposons make lepidopteran genomes look like mammalian genomes (32, 33).

We chose to measure the evolution rate of chromosomes (12) to further compare the genomic plasticity of Lepidoptera with that of other invertebrates like Diptera and nematodes. Thus, we examined the border of each synteny block to identify the recombination events responsible for the synteny breaks, and deduced a number of breakages per Mb. We calculated 85 breakages per Mb between the pest genomes and (132/150) breakages per Mb between pests and *B. mori*. These values are higher than the 51 breakages per Mb determined in a comparison of 13% of the *Caenorhabditis briggsae* genome with the *Caenorhabditis elegans* genome (12), whose divergence time is estimated at 50 to 120 Mya. These initial values of 0.4 to 1 breakages per Mb per Mya (12) were later refined by the same method to 0.5 to 0.7 breakages per Mb per Mya in a whole-genome comparison (34), suggesting that the values we obtained for our genome sampling are likely to be robust. The evolution rate of lepidopteran genomes (approximately 2 breakages per Mb per Mya) is thus faster than that of nematodes, themselves evolving fourfold faster than *Drosophila* species (12), whose chromosomes rearrange two orders of magnitude faster than those of mammals and faster than plant chromosomes (35). This very high rate is clearly not correlated to generation time or effective population size (36), because these life history traits can be very similar between Lepidoptera and higher Diptera. Nematodes and Lepidoptera share a common feature, i.e., the holocentric organization of their chromosomes. The scattered organization of centromeric determinants may lead to a greater genomic plasticity as chromosome fragments resulting from double-strand breaks may be maintained and reintegrated elsewhere. A ratio of 1.0:2.3 of inversions to transpositions was described in the comparison of

Table 3. Number of chromosomal breakages per Mb DNA since divergence of the species

Species	Divergence time, MYa	No. of breakages/Mb (number/Mb/MY)		
		Overall values for 15 regions, counting only 1:1 orthologue or best paralogue pairs	Values for 15 regions, excluding gene duplications	Values for seven regions lacking gene families
Ha/Sf	20–40	85 (2.83)	75 (2.50)	99 (3.30)
Bm/Ha	60–100	132 (1.65)	114 (1.43)	109 (1.37)
Bm/Sf	60–100	150 (1.88)	136 (1.70)	165 (2.07)
Mean	—	122 (2.12)	108 (1.87)	125 (2.25)

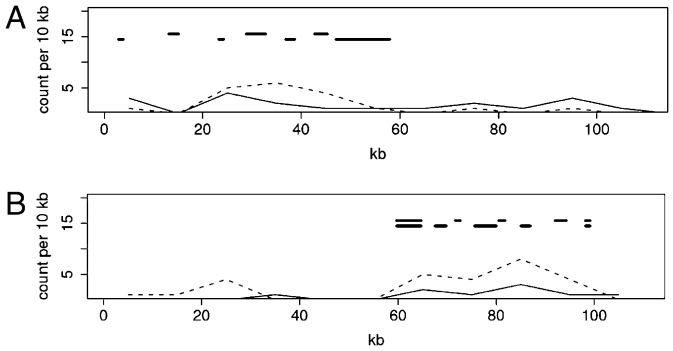


Fig. 3. Correlation between synteny breaks and TE copy density. TE copy density is shown as a function of BAC length. 3A, *S. frugiperda CYP4M* region; 3B, *H. armigera EcR* region; dotted lines, TE copy density. Thick lines, regions corresponding to synteny breaks; thin lines. gene density.

Caenorhabditis species (12). We find ratios of 1.0:8.0 (between noctuids) or 1.0:10.0 (average between noctuids and *Bombyx*). This higher rate of transposition events may result from a higher proportion of TE in *Bombyx* than in *C. briggsae* (45%/22.4%) and from the different nature of these elements (mainly retrotransposons in *Bombyx* vs. DNA transposons in *C. elegans*).

The synteny block size we measured is very small for species having diverged approximately 20 Mya (2.0 genes per block between the two noctuid species) or between 60 and 100 Mya (3, 5) (1.3 genes per block in our study for noctuids/*B. mori*). This is to be compared to synteny block size observed between *Drosophila pseudoobscura* with *D. melanogaster*, in which the average number of genes in syntenic blocks is 10 (37) and the divergence time approximately 55 Mya. Similarly, the gene-based microsynteny blocks as calculated with single copy orthologues in mosquitoes was 3.9 genes per block between *Aedes aegypti* and *Anopheles gambiae*, species that diverged 150 Mya (38). This figure decreases to 2.4 genes per block between *Aedes aegypti* and *D. melanogaster*, for which the divergence time is approximately 250 Mya.

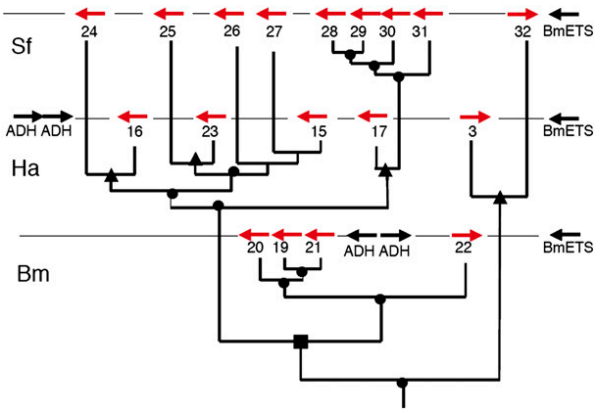


Fig. 4. Evolution of the *CYP9A* gene cluster. The *CYP9A* genes in the three species are shown in their correct orientation and order, but not relative distance, on the BACs or scaffold. From top to bottom: Sf, *S. frugiperda*; Ha, *H. armigera*; Bm, *B. mori*. The phylogenetic tree based on alignments of the CYP proteins is superimposed with its correct topology, but with branch lengths modified for clarity of the figure. Gene duplication events (●), *B. mori*/noctuid split (i.e., ancient speciation rather than duplication event) (■), and the *S. frugiperda*/*H. armigera* split (▲) are indicated. The sequence of events for the *CYP9A15, -26*, and *-27* genes is unresolved. The relative orientation of the genes indicates at least two inversions in addition to the duplication events. Recognized flanking genes [alcohol dehydrogenases (ADH) and BmETS transcription factor] are shown (see Fig. S2 for details).

The high genome fragmentation evidenced in our study is paradoxical, however, because the small size of the synteny blocks masks a higher order of synteny conservation, i.e., the "noise" of multiple insertions, deletions and inversions masks the "signal" that is apparent from many of our three-pair comparisons: In 11 of the 15 regions, there are six or more (as many as 16) genes found on the corresponding BAC/region of the other two species. Clearly there are constraints that prevent the total scrambling of gene order in these genomes. When estimating the relationship between protein sequence identity and gene order of orthologues in insect genomes, it was suggested that gene order would be lost below an average of 50% identity between orthologue pairs (13), but here synteny conservation was observed for many genes below that threshold.

The higher order of gene conservation was our initial expectation, based on previous studies in Lepidoptera (16). In a study of 72 orthologous loci between *B. mori* and *Heliconius melpomene*, a very high degree of conserved synteny was observed, with 21 linkage groups (LGs) of the butterfly mapped with one to seven genes in common with the silkworm. However, only one chromosomal inversion was noted, and it is probable that the density of markers per LG did not allow the detection of more rearrangements (17). A study on *Bicyclus anynana* with 462 markers documented a high degree of macrosyntenic relationships between the 28 chromosomes of *B. mori* and the 23 LGs of *B. anynana.* LG10 had 15 genes and LG 21 four genes in conserved order (19). The remainder of the chromosomes showed one or more rearrangements or inversions. Similarly, 124 of 131 orthologues retained the same order on the chromosomes of *B. mori* and *Manduca sexta* (20), and these authors noted the paradox of conserved synteny on holocentric chromosomes, in which increased chromosome rearrangements are more likely. With the higher resolution of BAC sequencing, high colinearity of genes was observed between two related species of *Heliconius*, *H. melpomene*, and *H. erato*, in two regions covering 180 kb and 280 kb of sequence (18). Just one gene was present in an indel block, although nine such indel blocks characterized by repetitive sequences were noted on a 180-kb span. A comparison of *H. erato* and *B. mori* over approximately 190 kb of sequence revealed 11 conserved gene, with one translocation and two inversions (18), thus foreshadowing our results. On 15 regions we observed both a high degree of gene conservation or macrosynteny and a high amount of local rearrangements.

Our observation of small synteny blocks in a background of high macrosynteny in Lepidoptera contrasts with Drosophila, in which paracentric inversions (i.e., within one chromosome arm or Muller element) are common and result in a lack of synteny (37, 39). This was evident in a direct comparison of two *Drosophila* genomes (37) and shown to explain most gene order shuffling in a study of 12 species (39). Furthermore, it was pointed out (37) that the fitness costs associated with such chromosomal rearrangements are reduced because there is no crossing over in the male, avoiding the generation of aneuploidy with dicentric/acentric chromosomes, whereas in the female such chromosomes are directed to polar bodies and not to gametes. The holocentric chromosomes of Lepidoptera appear to allow inversions only on much more limited chromosomal regions, possibly those between the multiple equivalents of centromeres. Perhaps holocentric chromosomes are then more resistant to large-scale rearrangements, thus explaining our paradoxical high degree of conserved synteny over the noise of the small chromosomal rearrangements. As it is the female that has no crossing over in Lepidoptera, this hypothesis suggests how negative fitness costs associated with chromosomal inversions might be reduced. The hypothesis needs to be tested on defined, larger, replication units. When the *H. armigera* genome sequence and those of other lepidopteran species become available, a larger dataset will become available for analysis by specific algorithms (e.g., ref .40) to answer these questions. Sequences functioning as centromeres would also need to be identified and mapped physically.

In conclusion, our microsynteny study provides insight into the extent of genome conservation that underlies macrosynteny and emphasizes the significance of the *B. mori* genome as a reference for Lepidoptera. It is logical to assume that the broad coverage of multiple chromosomes, and of genomic regions of varying degrees of gene richness, is sufficient justification to allow the extension of our conclusions to the whole genome. Despite a higher average size of genes in *B. mori*, and a different complement of repetitive sequences/TEs, the organization of the genomes shows a high degree of macrosynteny conservation. *Bombyx* can therefore be used as a first reference for genomic/genetic studies in Lepidoptera. Our study also highlights the very rapid evolution of the three genomes. Gene clusters from rapidly evolving multigene families such as the P450s show evidence of recent and multiple gene duplication events that are specific for both noctuid genomes; however, we report a high number of synteny breaks or small size of synteny blocks, evidence for more frequent rearrangements than in the *Drosophila* lineage for instance. Therefore, the view that emerges from our comparative genome analysis depends on the focus. From up close, the fragmentation is obvious, but on a larger scale conservation is maintained despite, or perhaps because of, the holocentric nature of the lepidopteran genomes. This higher order level of chromosome organization and the evolutionary constraints that have led to it in Lepidoptera remain a matter of conjecture.

Materials and Methods

BAC library construction, probe selection, and screening, BAC sequencing and annotation by Kaiko (Silkworm) Genome Automated Annotation System (KAIKOGAAS) followed by manual curation are described in detail in *SI Materials and Methods*. The fully annotated BACs and *B. mori* genome regions are available at the Lepido-DB Web site. To detect syntenic genes, each BAC sequence was compared by BlastX (threshold $<10^{-5}$) to the sequences of the peptides predicted by KAIKOGAAS on the corresponding BAC in the other noctuid species or on a 200-kb region of the *B. mori* genome. We determined reciprocal best hits that resulted from a comparison by BlastP between two sets of peptides at a threshold of 10^{-6} over 75% of the peptide length. Only a partial genomic sequence (i.e., our BAC collection) is available for the two noctuid species that were compared in the present study; therefore, the reciprocal best hits may not always represent strictly orthologous genes. In the case of unique genes in the *B. mori* genome, this assumption can reasonably be made, however. When a gene has duplicated in one species, the closest paralogue and thus most likely orthologue was defined using phylogenetic trees built with all paralogous proteins.

To measure microsynteny block length (24), all segments within which gene order and orientation was conserved between two genomic regions were identified and the number of genes they contain was counted. The pipeline for definition of the gene categories used in our analysis of synteny conservation is outlined and explained in *SI Materials and Methods* and Fig. S1.

ACKNOWLEDGMENTS. This work was supported by Genoscope project 2004/43 Génomique comparée des Lépidoptères; the Integrated Research Project for Plant, Insect and Animal Using Genome Technology (Insect Genome) by MAFF, Japan; Institut National de la Recherche Agronomique/Plant Health and Environment (H.S. and C.G.); and Agence Nationale de la Recherche Grants 06 BLAN 0346 (to R.F.), 07 GPLA 004, and 07 BLAN 0057 (to E.P.) including a postdoctoral fellowship. T.F. is the recipient of an Institut National de la Recherche Agronomique PhD fellowship. We thank Drs. D. G. Heckel, E. Jacquin-Joly, H. H. Rees, Y. Park, and T. Shinoda for kindly providing us with some hybridization probes, and reviewers for thoughtful comments.

1. The International Silkworm Genome Consortium (2008) The genome of a lepidopteran model insect, the silkworm Bombyx mori. *Insect Biochem Mol Biol* 38:1036–1045.
2. Labandeira CC, Dilcher DL, Davis DR, Wagner DL (1994) Ninety-seven million years of angiosperm-insect association: Paleobiological insights into the meaning of coevolution. *Proc Natl Acad Sci USA* 91:12278–12282.
3. Gaunt MW, Miles MA (2002) An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. *Mol Biol Evol* 19: 748–761.
4. Gall LF, Tiffney BH (1983) A Fossil Noctuid Moth Egg from the Late Cretaceous of Eastern North America. *Science* 219:507–509.
5. Grimaldi D, Engel MS (2005) *Evolution of the Insects* (Cambridge University Press, Cambridge), 755 pp.
6. Yamamoto K, et al. (2006) Construction of a single nucleotide polymorphism linkage map for the silkworm, Bombyx mori, based on bacterial artificial chromosome end sequences. *Genetics* 173:151–161.
7. Yamamoto K, et al. (2008) A BAC-based integrated linkage map of the silkworm Bombyx mori. *Genome Biol* 9:R21.
8. Fitt GP (1989) The ecology of Heliothis species in relation to agroecosystems. *Annu Rev Entomol* 34:17–52.
9. Sharma HC (2005) *Heliothis/Helicoverpa Management: Emerging Trends and Strategies for Future Research.* (Oxford & IBH Publishing, New Delhi).
10. Behere GT, et al. (2007) Mitochondrial DNA analysis of field populations of Helicoverpa armigera (Lepidoptera: Noctuidae) and of its relationship to H. zea. *BMC Evol Biol* 7:117.
11. Nadeau JH, Sankoff D (1998) Counting on comparative maps. *Trends Genet* 14:495–501.
12. Coghlan A, Wolfe KH (2002) Fourfold faster rate of genome rearrangement in nematodes than in Drosophila. *Genome Res* 12:857–867.
13. Zdobnov EM, Bork P (2007) Quantification of insect genome divergence. *Trends Genet* 23:16–20.
14. Zdobnov EM, et al. (2002) Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. *Science* 298:149–159.
15. Bourque G, Pevzner PA, Tesler G (2004) Reconstructing the genomic architecture of ancestral mammals: Lessons from human, mouse, and rat genomes. *Genome Res* 14: 507–516.
16. Sahara K, et al. (2007) Conserved synteny of genes between chromosome 15 of Bombyx mori and a chromosome of Manduca sexta shown by five-color BAC-FISH. *Genome* 50:1061–1065.
17. Pringle EG, et al. (2007) Synteny and chromosome evolution in the lepidoptera: Evidence from mapping in Heliconius melpomene. *Genetics* 177:417–426.
18. Papa R, et al. (2008) Highly conserved gene order and numerous novel repetitive elements in genomic regions linked to wing pattern variation in Heliconius butterflies. *BMC Genomics* 9:345.
19. Beldade P, Saenko SV, Pul N, Long AD (2009) A gene-based linkage map for Bicyclus anynana butterflies allows for a comprehensive analysis of synteny with the lepidopteran reference genome. *PLoS Genet* 5:e1000366.
20. Yasukochi Y, et al. (2009) Extensive conserved synteny of genes between the karyotypes of Manduca sexta and Bombyx mori revealed by BAC-FISH mapping. *PLoS One* 4:e7465.
21. Honeybee Genome Sequencing Consortium (2006) Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* 443:931–949.
22. Quesneville H, et al. (2005) Combined evidence annotation of transposable elements in genome sequences. *PLOS Comput Biol* 1:166–175.
23. Osanai-Futahashi M, Suetsugu Y, Mita K, Fujiwara H (2008) Genome-wide screening and characterization of transposable elements and their distribution analysis in the silkworm, Bombyx mori. *Insect Biochem Mol Biol* 38:1046–1057.
24. Sankoff D (1999) *Comparative mapping and genome rearrangement* (Iowa State Univ, Ames, IA), pp 124–134.
25. Soberon M, Gill SS, Bravo A (2009) Signaling versus punching hole: How do Bacillus thuringiensis toxins kill insect midgut cells? *Cell Mol Life Sci* 66:1337–1349.
26. Angelucci C, et al. (2008) Diversity of aminopeptidases, derived from four lepidopteran gene duplications, and polycalins expressed in the midgut of *Helicoverpa armigera*: Identification of proteins binding the delta-endotoxin, Cry1Ac of Bacillus thuringiensis. *Insect Biochem Mol Biol* 38:685–696.
27. Budatha M, Meur G, Kirti PB, Dutta-Gupta A (2007) Characterization of *Bacillus thuringiensis* Cry toxin binding novel GPI anchored aminopeptidase from fat body of the moth Spodoptera litura. *Biotechnol Lett* 29:1651–1657.
28. Feyereisen R (2006) Evolution of insect P450. *Biochem Soc Trans* 34:1252–1255.
29. Taylor M, Zawadzki J, Black B, Kreitman M (1993) Genome size and endopolyploidy in pyrethroid-resistant and susceptible strains of *Heliothis virescens* (Lepidoptera: Noctuidae). *J Econ Entomol* 86:1030–1034.
30. Gregory TR (2005) Synergy between sequence and size in large-scale genomics. *Nat Rev Genet* 6:699–708.
31. Clark AG, et al. (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 450:203–218.
32. Belancio VP, Hedges DJ, Deininger P (2008) Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health. *Genome Res* 18:343–358.
33. Cordaux R (2008) The human genome in the LINE of fire. *Proc Natl Acad Sci USA* 105: 19033–19034.
34. Stein LD, et al. (2003) The genome sequence of Caenorhabditis briggsae: A platform for comparative genomics. *PLoS Biol* 1:E45.
35. Ranz JM, Casals F, Ruiz A (2001) How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila. *Genome Res* 11:230–239.
36. Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L (2005) Chromosome evolution in eukaryotes: A multi-kingdom perspective. *Trends Genet* 21:673–682.
37. Richards S, et al. (2005) Comparative genome sequencing of Drosophila pseudoobscura: Chromosomal, gene, and cis-element evolution. *Genome Res* 15:1–18.
38. Nene V, et al. (2007) Genome sequence of Aedes aegypti, a major arbovirus vector. *Science* 316:1718–1723.
39. Bhutkar AS, et al. (2008) Chromosomal rearrangement inferred from comparisons of 12 Drosophila genomes. *Genetics* 179:1657–1680.
40. Zhao H, Bourque G (2009) Recovering genome rearrangements in the mammalian phylogeny. *Genome Res* 19:934–942.