Annotation des génomes de paramécies

Olivier Arnaiz

Thèse de doctorat

NNT : 2020UPASL046

université PARIS-SACLAY

Annotation de génomes de paramécies

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°577 Structure et dynamique des systèmes vivants (SDSV)
Spécialité de doctorat: Sciences de la vie et de la santé
Unité de recherche : Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France
Référent : Université de Versailles -Saint-Quentin-en-Yvelines

Thèse présentée et soutenue à Gif-sur-Yvette, le 27 novembre 2020, par

Olivier ARNAIZ

Composition du Jury

Gaëlle LELANDAIS Professeur, Université Paris-Saclay	Présidente
Hugues ROEST CROLLIUS Directeur de Recherche, IBENS	Rapporteur
Stéphanie SIDIBE BOCS Chargée de Recherche, CIRAD	Rapportrice
Gilles FISCHER Directeur de Recherche, LCQB	Examinateur
Olivier JAILLON Chargé de Recherche, Institut Jacob - Genoscope	Examinateur
Hadi QUESNEVILLE Directeur de Recherche, URGI	Examinateur
Linda SPERLING Directrice de Recherche, I2BC	Directrice de thèse



Université Paris-Saclay
École doctorale 577: Structure et dynamique des systèmes vivants
Institut de Biologie Intégrative de la Cellule

Thèse de Doctorat

Spécialité « Sciences de la vie et de la santé »

par

Olivier ARNAIZ

Annotation de génomes de paramécies

Thèse soutenue le 27 novembre 2020 devant le jury composé de :

Mme	Gaëlle Lelandais	Université Paris-Saclay - Gif-sur-Yvette	(Présidente)
M.	Hugues Roest Crollius	IBENS - Paris	(Rapporteur)
Mme	Stéphanie Sidibe Bocs	CIRAD - Montpellier	(Rapportrice)
M.	Gilles Fischer	LCQB - Paris	(Examinateur)
M.	Olivier Jaillon	Institut Jacob, Génoscope - Evry	(Examinateur)
M.	Hadi Quesneville	URGI - Versailles	(Examinateur)
Mme	Linda Sperling	I2BC - Gif-sur-Yvette	(*Accompagnatrice* VAE)

À mon papa

Remerciements

Par où commencer ? Comment remercier toutes les personnes que j'ai croisées ou avec qui j'ai collaboré pendant ces 15 dernières années. Je vais très probablement en oublier donc je voudrais m'excuser de ne pas les avoir citées dans ces quelques lignes, et les remercier.

Avant tout, un grand merci aux membres du jury d'avoir accepté de lire ce manuscrit et d'évaluer mon travail de thèse. Un merci particulier à mes rapporteurs Stéphanie Sidibe Bocs et Hugues Roest Crollius.

Rien n'aurait été possible sans le soutien indéfectible de Linda. Je souhaite te remercier plus que chaleureusement pour tout ce que tu m'as apporté. Tu m'as fait confiance. Tu m'as encouragé. Parfois, tu m'as même protégé. Je pense que tu m'as fait progresser à tous les niveaux et je t'en suis très reconnaissant. Je trouve que nous formons une très bonne équipe. En effet, nous sommes très différents mais nous nous complémentons : tu es pessimiste, je suis trop naïf. Tu es stressée, je suis insouciant. Tu aimes regarder les données de près, je pars tout de suite sur des analyses globales sans réfléchir. Tu aimes écrire, je n'aim... en face de ce document, on comprend ma *douleur*. Merci Linda d'avoir relu (à plusieurs reprises) ce manuscrit. Je sais que tu as rongé ton frein de ne pas corriger chaque phrase.

Après mes études, je me suis toujours dit que je ne ferai jamais de thèse. C'était pas mon truc ! Et me voilà, à l'aube de mes 40 ans, à redevenir étudiant et à vouloir devenir docteur. Mireille, je souhaite grandement te remercier pour ton soutien et tes encouragements à relever ce défi doctoral. Je vous revois toutes les deux, avec Linda, me regardant : ... alors tu y vas !!!?!!! En effet, ma vision de la thèse a évolué ces dernières années. Cyril, je te remercie, pas seulement, mais notamment pour ça. En acceptant ce challenge, je comptais apprendre à écrire et apprendre à aimer écrire. Je ne suis pas persuadé d'avoir atteint ces deux objectifs mais il est vrai que plus le manuscrit avançait, moins c'était compliqué... J'ai même parfois apprécié. Toujours est-il, je suis très content de l'avoir fait ! Linda, Mireille encore merci à toutes les deux.

Merci à tous les membres de mon équipe actuelle (Coralie, Mélanie, Vinciane, Julien et Marc), mais également à tous les anciens membres de l'équipe Bétermier, que je ne citerai pas mais qui se reconnaitront. Depuis que je participe aux réunions de labo de

l'équipe de Mireille, à parler tampon, perméabilisation, CPHG et autre injection, j'ai sincèrement l'impression de mieux comprendre les problèmes de paillasse (et aussi pourquoi je fais de l'informatique) et même la biologie de la paramécie.

Également dans mes remerciements, l'ensemble de mes collaborateurs et plus particulièrement Sandra, Eric et Laurent (ainsi que leur équipe). Merci pour ces heures de conversation téléphonique passionnante et enrichissante. Petit clin d'œil à mes collaborateurs plus éloignés que sont Jacek et Jeff.

Comment ne pas remercier tous les paraméciologues de l'équipe d'Anne-Marie. Le plaisir de venir travailler est intact depuis toutes ces années grâce à la bonne ambiance qui règne, et qui a toujours régné, dans cette aile du bâtiment 26. Jean, merci de m'avoir fait profiter de ton harcèlement légendaire. France, j'espère que tu n'es pas trop déçue que je ne sois finalement pas parti dans le privé. Khaled, merci de supporter mes heures de téléphone avec Linda. Une pensée également aux autres membres de l'équipe (Janine, Pierrick, Manon, Michel, ...) mais aussi à tous les anciens paraméciologues du CGM.

Il est essentiel pour moi de remercier les personnels du service informatique et de la plateforme de séquençage pour leur conseil et assistance. Plus généralement, un grand merci à tous les services supports de l'I2BC, nous aidant dans l'ombre à faire de la bonne science.

Enfin, merci à ma famille et à mes amis pour leur soutien, et notamment à Sabine, ma partenaire de/à vie qui m'a également poussé à faire cette thèse. Une dédicace particulière à mes parents qui m'ont transmis la passion des sciences et ont participé à forger ma curiosité, combinaison utile dans la recherche. À mon papa ... , je suis persuadé que tu aurais été fier de moi ...

Avant propos

L'objectif de cet *avant propos* est de me présenter et de retracer mon parcours professionnel et scientifique. Cette *biographie* me permettra d'introduire les grandes thématiques biologiques que j'ai pu aborder durant ma carrière. Elle me donnera également l'occasion de citer l'ensemble des travaux que j'ai co-signés. Certaines problématiques scientifiques mentionnées dans cette section seront décrites avec plus de détails dans mon introduction. Enfin, je donnerai quelques explications au sujet de la construction de ce manuscrit.

.1 Contexte actuel de travail

Je travaille à l'I2BC (Institut de Biologie Intégrative de la Cellule). J'ai récemment intégré l'équipe de recherche dirigée par Mireille Bétermier "Réarrangement programmées du génome" (dans le département "Biologie des génomes"). Mon équipe s'intéresse aux mécanismes programmés d'élimination d'ADN chez un cilié : la paramécie. Ces réarrangements programmés de génome impliquent des éléments transposables (ET) et leur contrôle épigénétique mais également des acteurs de voies de réparation de cassures de l'ADN.

Tout d'abord, un mot de mon activité professionnelle. Je suis ingénieur d'étude en calcul scientifique dans un laboratoire de biologie, autrement dit bioinformaticien. Ma préoccupation principale est de répondre à des questions biologiques en utilisant des méthodes informatiques. J'essaye au maximum d'utiliser les méthodes existantes. Quotidiennement, je développe de nouvelles méthodes sans pour autant me considérer comme un algorithmicien. Je ne suis ni un biologiste, ni un informaticien mais je peux "communiquer" avec les deux. Le terme "bioinformaticien" est vague et regroupe plusieurs métiers et plusieurs champs disciplinaires (génomique, biologie structurale, modélisation, statistique, évolution, construction et gestion de bases de données, ...). Personnellement, mon domaine concerne plutôt l'étude des génomes et de leur dynamique même si mes premières amours ont porté sur le développement de bases de données. La communauté "paramécie" utilisant des approches de biologie moléculaire est relativement petite. Ces équipes étudient des thématiques variées comme la génomique, l'évolution de génomes, l'étude des réseaux cytosquelettiques, les cils, la symbiose ou l'éco-toxicologie... Seulement 3 500 publications traitent de la paramécie depuis le début des années 1900 (nombre

de références bibliographiques en janvier 2020 avec le mot clé *Paramecium* sur PubMed, en comparaison à 7 000, 73 000, 107 000 pour *Tetrahymena*, *Arabidopsis* et *Drosophila* respectivement). Avec l'essor des nouvelles technologies de séquençage, la quantité de données a augmenté de façon exponentielle et en conséquence les besoins en bioinformatique. J'ai la chance de jouer un rôle central au sein des consortiums français et européens, me permettant de participer à des études variées.

Une vision schématique divise mon activité en quatre grandes thématiques. Toutes en lien avec l'étude de la paramécie, il est bien évident qu'elles sont largement interconnectées.

— **Développement et maintien de systèmes d'information**. Vous verrez, dans la suite de cet avant propos, que ce travail a occupé une grosse partie de mon activité, notamment au début de ma carrière.
— **Assemblage et annotation de génomes**. Au cœur de mon activité, c'est la thématique que j'ai choisi d'approfondir dans ce document de thèse. Des aspects plus évolutifs sont étudiés en collaboration avec les laboratoires de L. Duret (LBBE de Lyon) et M. Lynch (*Indiana University* puis *Arizona State University* aux USA).
— **Étude des mécanismes épigénétiques impliqués dans la reconnaissance des séquences à éliminer pendant les réarrangements**. Ce sujet est traité en étroite collaboration avec les équipes de S. Duharcourt (IJM Paris), E. Meyer (ENS Paris) et JK Nowak (IBB en Pologne), et fait l'objet de quelques paragraphes dans l'introduction.
— **Étude des mécanismes impliqués dans les coupures d'ADN et leur réparation**. J'explore cette thématique au sein de l'équipe de M. Bétermier.

.2 Mon parcours professionnel et scientifique

Après des études de biologie cellulaire et physiologie à Jussieu (Paris 6), je me suis orienté vers un DESS *Informatique Appliquée à la Biologie* (2004). Le peu d'appétence que j'avais pour l'expérimentation "humide" et cette année de cours intensifs en informatique m'ont conforté dans mon choix de travailler en bioinformatique. L'objectif de mon stage de DESS était de développer une interface Web pour visualiser des informations d'interactions protéine-protéine au sein de complexes (Benoit et al. 2008) [1]. Heureux de cette expérience, j'ai voulu continuer dans le domaine du développement de systèmes d'information. En 2005, Linda Sperling propose un CDD d'un an pour la conception d'un système d'information afin d'exploiter le génome et l'annotation d'un eucaryote unicellulaire appelé la paramécie. Lors de cette année, j'ai mis en place la première mouture de ParameciumDB et découvert progressivement la paramécie et sa biologie (Arnaiz et al. 2007). Ce contrat marque également le début d'une collaboration fructueuse et enrichissante avec Linda (et la paramécie).

En 2007, j'ai été recruté au CNRS en tant qu'ingénieur d'étude en systèmes d'informa-

1. Je ne suis pas auteur de l'article.

tion (Branche d'Activité Professionnelle E). Jusqu'en 2009, mon activité était consacrée au développement de bases de données. Progressivement une activité de recherche a pris de plus en plus d'importance. Au-delà de ma volonté d'évoluer, j'imagine qu'il y a plusieurs raisons à cette mutation. La création de ParameciumDB a été fortement connectée au séquençage et à l'annotation du génome somatique de *Paramecium tetraurelia* (Aury et al. 2006). Avoir accès au génome entier et l'arrivée des nouvelles technologies de séquençage ont profondément changé les méthodologies des "paraméciologues". En effet, les besoins en bioinformatique sont devenus de plus en plus criants pour le consortium paramécie et mes compétences ont été sollicitées pour les projets scientifiques.

.2.0.1 Développement de systèmes d'information

Le système d'information (SI) ParameciumDB (`https://paramecium.i2bc.paris-saclay.fr/`) a été inauguré en août 2005 (Arnaiz et al. 2007). Il a subi une mise à jour en 2010 (Arnaiz and Sperling 2011), puis en 2018 (Arnaiz et al. 2019). La vocation de ParameciumDB est de stocker des informations liées à l'étude de la paramécie. En effet, des données génomiques, génétiques, phénotypiques, transcriptomiques, protéomiques, bibliographiques et des informations sur les souches y sont référencées. Avec Linda, et dès le début du projet, nous avons adhéré à la philosophie, et aux composants, du projet GMOD (*Generic Model Organism Database*). GMOD distribue une *collection d'outils génériques et interopérables pour gérer, visualiser, stocker et disséminer des données génétiques et génomiques*. Afin de ne pas réinventer le fil à couper le beurre, le consortium GMOD donne accès à des outils libres et matures aux biologistes désirant exploiter leurs données. De petites communautés, sans grosse infrastructure informatique ou moyens financiers, peuvent profiter d'années de développement de logiciels. Pour Scott Cain (coordinateur du projet GMOD) venu nous rendre visite en 2005, la communauté paramécie et le projet ParameciumDB étaient exactement le public visé par GMOD. Au-delà des outils incontournables comme *GBrowse* ou *chado* mis en place dans ParameciumDB, nous avons été les premiers à déployer l'environnement (*framework*) GMODweb (O'Connor et al. 2008).

Aux laboratoires de Gif-sur-Yvette, la paramécie est étudiée pour deux aspects : la génomique et les réarrangements de génome, qui sont mes champs de prédilection, et l'étude des cils et du cytosquelette. La paramécie est un cilié, et comme son nom l'indique, elle est tapissée de cils. La plupart des organismes présentent des cellules avec un ou plusieurs cils. Des défauts dans la biogenèse ou le fonctionnement de ces cils entrainent des maladies humaines appelées ciliopathies. Le cil étant une structure très conservée au cours de l'évolution, la paramécie avec 4000 cils à sa surface est un modèle très pertinent pour l'étude de la biogénèse et du fonctionnement du cil (Gogendeau et al. 2008, Shi et al. 2017, Gogendeau et al. 2019). L'équipe dirigée par Jean Cohen, puis reprise par Anne-Marie Tassin, étudie cette thématique à la fois chez la paramécie et chez la souris. Avec France Koll et Jean Cohen, nous avons constaté le besoin de la communauté scientifique

d'une base de données regroupant, et mettant en relation, un ensemble de résultats hétérogènes issus d'études ciliaires menées sur différents organismes. De cette observation est née le SI Cildb (`http://cildb.i2bc.paris-saclay.fr/`). Pour construire Cildb, j'ai profité du logiciel BioMart (outil GMOD) popularisé par le portail EnsEMBL (Guberman et al. 2011, Smedley et al. 2015). Aujourd'hui, Cildb jouit d'une très bonne réputation et est considéré comme un outil important pour les chercheurs étudiant les ciliopathies. Ce travail a été valorisé par deux publications (Arnaiz et al. 2009; 2014). La conception de cette base de données m'a permis d'être distingué par le cristal du CNRS en 2011.

.2.0.2 Étude du génome de la paramécie

Avant d'aborder les études menées sur la paramécie, je dois introduire quelques notions sur sa biologie quelque peu exotique (plus de détails dans la **section III** p.51). Comme les animaux, les ciliés séparent la lignée germinale (comme le spermatozoïde ou l'ovule chez l'humain) de la lignée somatique (cellules formant le corps de l'organisme) mais au sein d'une seule cellule. La **figure 1A** montre une paramécie avec ses deux types de noyaux dans son cytoplasme. Un noyau germinal diploïde (micronoyau ou *micronucleus* MIC) utilisé pour transmettre l'information génétique à la prochaine génération sexuelle, et un noyau somatique (macronoyau ou *macronucleus* MAC) polyploïde (800n) optimisé pour l'expression des gènes (voir **Figure III.1** p.52). A chaque cycle sexuel (la **figure 1B** montre deux paramécies en conjugaison), l'ancien MAC est progressivement détruit et un nouveau MAC est généré à partir d'une copie du MIC ayant subi des réarrangements programmés de génome. Ces réarrangements comprennent une amplification d'ADN (passage d'une ploïdie de 2n à 800n) et une élimination de matériel génétique comprenant des séquences répétées (satellites et éléments transposables) et de petites séquences non codantes en copie unique appelées IES (*Internal Eliminated Sequence*).

Ma première réelle expérience de recherche a porté sur le séquençage et l'annotation du génome MAC de *Paramecium tetraurelia* (Aury et al. 2006). Ce génome, d'une complexité d'environ 72 Mb, est relativement atypique pour un eucaryote car il est très riche et très dense en gènes (∼40 000 gènes codant pour des protéines et 78% du génome est codant). Sa haute teneur en gènes peut être expliquée par des duplications globales de génome (WGD *Whole Genome Duplication*). En effet, le génome de *Paramecium tetraurelia* a subi aux moins trois WGD successives. La WGD la plus récente est encore très visible au niveau nucléotidique, et coïncide avec une explosion d'événements de spéciation donnant naissance au groupe *aurelia* contenant au moins 15 espèces de paramécie morphologiquement identiques. Nous avons également montré qu'après une WGD, beaucoup de gènes sont maintenus en deux copies (la moitié des gènes dupliqués suite à la WGD la plus récente est encore présente en deux copies) du fait de contraintes de dosage notamment sur les complexes protéiques. En effet, le taux de rétention des paralogues de WGD est relié au niveau d'expression des gènes. Ces résultats ont été étudiés en utilisant une approche de puces à ADN (Arnaiz et al. 2010) et une approche ARN-seq (Arnaiz et al. 2017).

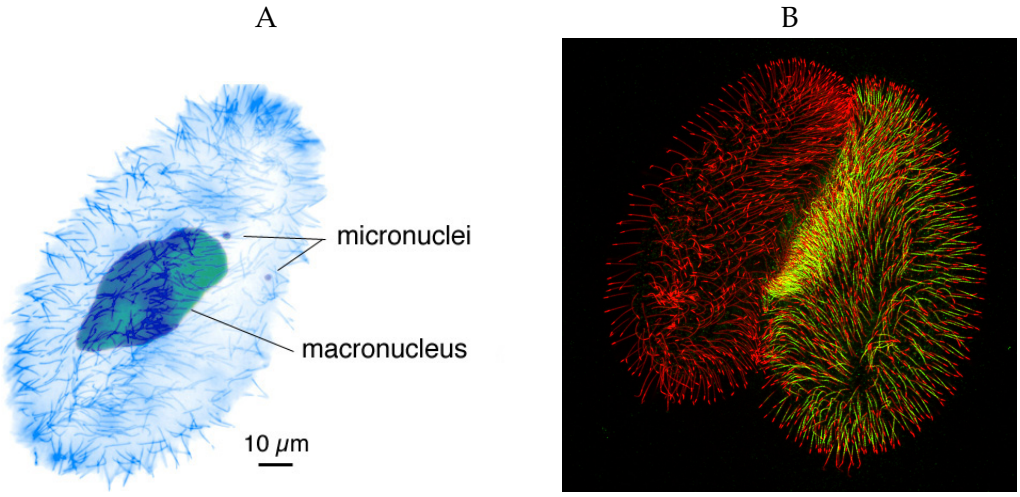


FIGURE 1 – *Images de paramécies*

A Image d'immunofluorescence réalisée par Janine Beisson montrant les cils couvrant la surface de la paramécie ainsi que les noyaux germinaux et somatiques. **B** Image d'immunofluorescence en utilisant un microscope confocale réalisée par A. Aubusson-Fleury montrant un couple de paramécies pendant la conjugaison. En jaune le marquage fluorescent d'une protéine ciliaire pour l'un des partenaires, et en rouge l'immunomarquage d'une modification post-traductionnelle d'une tubuline décorant les cils.

Parallèlement ces études transcriptomiques ont permis d'identifier des gènes différentiellement exprimés pendant le cycle sexuel de la paramécie. Comme nous le verrons dans ce manuscrit, une partie de ces gènes sont impliqués dans les processus de réarrangement de génome (NOWACKI ET AL. 2009, BHULLAR ET AL. 2018, NEKRASOVA ET AL. 2019).

Les IES peuvent être intragéniques et doivent donc s'exciser précisément, pendant la maturation du nouveau MAC, pour reformer des gènes fonctionnels. Même si ce mécanisme d'excision est très efficace, un certain taux d'erreur d'excision des IES persiste parmi les 800n que compte le noyau somatique. En recherchant des incohérences entre le génome MAC et les lectures de séquençage, nous avons pu identifier des centaines de nouvelles IES putatives, venant s'ajouter aux 42 IES connues à l'époque (DURET ET AL. 2008). L'identification de toutes les IES du génome était compliquée par le fait que l'ADN des micronoyaux ne représente que 0.5% de l'ADN de la cellule et que la purification de ces noyaux n'était pas maitrisée. Mais ces verrous techniques ont été levés par l'identification de l'endonucléase impliquée dans l'élimination d'ADN. Cette protéine, appelée Pgm (PiggyMac) est une transposase domestiquée d'un transposon de la famille piggyBac. Avec la paramécie, il est possible de réaliser de l'extinction génique par ARN interférence. L'ingestion de bactéries exprimant des ARN double brins homologues à la région cible, entraine la production de petits ARN (siARN). Des facteurs protéiques associés à ces siARN vont conduire à la dégradation des ARNm du gène visé (MARKER ET AL. 2014, CARRADEC ET AL. 2015).

Par bonheur, l'inactivation de PGM pendant les réarrangements prévient l'élimination des régions spécifiques du MIC mais n'empêche pas l'amplification de l'ADN. Ainsi il est

possible d'obtenir un ADN enrichi en séquences non réarrangées. J'ai établi les procédures d'identification et de caractérisation des IES à l'échelle du génome à partir de cet ADN (Arnaiz et al. 2012, Denby Wilkes et al. 2016). Environ 45 000 IES ont été identifiées. Nous avons montré qu'une fraction des IES avait une homologie avec des transposons Tc1/Mariner, ce qui validerait l'hypothèse que les IES seraient des reliques de transposons. Je me dois de nuancer légèrement cette observation, car nous avons montré qu'une IES peut être exaptée à partir d'une séquence cellulaire originellement destinée au MAC afin de créer une nouvelle fonction (Singh et al. 2014). L'excision ou non de cette IES conditionne l'expression d'un gène déterminant le type sexuel, apportant la preuve de principe qu'une IES peut avoir une fonction.

A partir de 2013, la connaissance du génome germinal a franchi une autre étape avec la mise au point d'une méthode de purification de noyaux micronucléaires par cytométrie de flux (FACS *Fluorescence Activated Cell Sorting*). Dans Guérin et al. (2017), nous avons pu accéder aux premières véritables séquences MIC. L'assemblage, certes incomplet, a permis d'annoter de nouveaux éléments transposables et de confirmer que toutes les IES dépendent de PGM pour leur excision. Fort de ce succès, le séquençage de génomes MIC (plusieurs articles en préparation) et MAC (Gout et al. 2019) de plusieurs espèces de paramécie a été initié dans le cadre d'un grand projet de séquençage France Génomique impliquant des équipes CNRS, INRA ainsi que le *Génoscope* (voir **discussion C** p.157).

.2.0.3 Étude des réarrangements de génome

Les IES sont invariablement bornées de deux 5'-TA-3' et présentent un consensus très faible aux bornes (5'-TAYAGYNR-3'). Cette information n'est pas suffisante pour reconnaître les IES, en particulier dans un génome riche en AT (28% G+C). La question du ciblage des régions à éliminer est une problématique cruciale. Cette reconnaissance passe par des petits ARN de 25nt (scnARN) comparables aux piRNA. Mais contrairement aux piRNA générés sur des "*clusters*" génomiques, les scnARN sont générés à partir de tout le génome germinal (Singh et al. 2014). Dans l'introduction je décrirai le modèle de "*scanning*" prônant que ces petits scnARN seraient produits à partir du génome MIC en méiose puis comparés, par appariement de séquence, à des transcrits non-codants produits dans l'ancien MAC (Maliszewska-Olejniczak et al. 2015, Gruchota et al. 2017). Les scnRNA correspondant aux séquences spécifiques du génome MIC, n'ayant pas d'équivalent dans le MAC seraient transportés dans le MAC en développement pour cibler spécifiquement les séquences homologues à éliminer grâce à un marquage épigénétique (Frapporti et al. 2019). Grâce à ce « système immunitaire », la cellule se protège contre l'invasion de tout parasite moléculaire. En effet, le MAC étant perdu à chaque cycle sexuel, son invasion n'aurait aucune conséquence sur le patrimoine génétique transmis à la descendance, tandis qu'un ADN étranger s'intégrant dans le génome MIC seraient éliminé du MAC par ce système de défense.

Afin d'éliminer les séquences, l'endonucléase Pgm, aidée par des co-facteurs, réalise

une cassure double brin (CDB) de l'ADN (Bischerour et al. 2018). Ces CDB ont lieu à chaque extrémité des IES et sont réparées par les acteurs de la voie NHEJ (*Non Homologous End Joining*) de réparation des CDB de l'ADN (Marmignon et al. 2014). Avec 45000 IES il faut imaginer 90000 CDB (par génome haploide) à introduire et à réparer dans une fenêtre de temps réduite. Ce mécanisme complexe doit être à la fois précis, efficace et rapide. Dans l'équipe de Mireille Bétermier, nous étudions les facteurs impliqués à la fois dans la coupure et dans la réparation de l'ADN.

.3 Pourquoi faire une thèse ?

Depuis presque 15 ans, je travaille en lien avec la recherche sur la paramécie, ce qui me permet de rencontrer et de collaborer avec de nombreuses personnalités scientifiques. Toutes ces années n'ont en rien érodé la passion que j'ai pour mon activité. Les membres de mon équipe et mes collaborateurs m'ont fait confiance et m'ont donné l'occasion d'évoluer dans mon métier et dans mes responsabilités. Cette dynamique me permet d'avoir une activité variée (recherche, encadrement, rédaction, ...) et d'apprendre tous les jours. De plus, la paramécie est un modèle d'étude palpitant nous réservant des surprises quotidiennement.

Le co-encadrement de la thèse de Cyril Denby Wilkes (2014) m'a donné envie de soutenir une thèse moi-même. Je pense avoir atteint un stade de ma carrière où j'ai besoin de prendre du recul sur mon activité de recherche. Une thèse est l'opportunité idéale d'avoir une réflexion sur mon parcours scientifique. Présenter une thèse de doctorat dans le cadre de la Valorisation des Acquis et Expérience (VAE) est donc l'occasion parfaite.

.4 Ma thématique et l'organisation du manuscrit

Mon métier de bioinformaticien m'a donné l'occasion de travailler sur beaucoup de thématiques scientifiques différentes. Présenter l'ensemble de mon travail dans un même manuscrit aurait été très compliqué. J'ai donc choisi de baser ce document et ma réflexion sur mes travaux directement reliés à l'annotation de génome de paramécie. J'ai œuvré à l'amélioration des chaines de procédures pour l'annotation des gènes d'un génome de paramécie (Arnaiz et al. 2017). J'ai également caractérisé et étudié les IES à l'échelle du génome pour l'espèce *Paramecium tetraurelia* (Arnaiz et al. 2012). J'ai co-développé le logiciel ParTIES pour quantifier l'efficacité d'excision des IES et détecter les erreurs d'excision (Denby Wilkes et al. 2016). En 2017, nous avons publié le premier assemblage MIC et ce génome a permis d'annoter manuellement de nouveaux ET (Guérin et al. 2017). Nous avons montré qu'il existait un lien évolutif entre les IES et les ET (Arnaiz et al. 2012). De plus, je participe à l'annotation des ET de génomes MIC de plusieurs espèces de paramécies qui va compléter notre connaissance et notre vision des IES et des ET (voir **discussion C** p.157).

Pour moi, le manuscrit d'une thèse en VAE ne peut pas être construit comme un manuscrit de thèse classique. Je pense notamment à l'élaboration de l'introduction. A mon sens, la différence vient probablement du fait que le travail s'étale sur une fenêtre de temps plus longue. La science avançant, ce qui est "résultat" aujourd'hui devient "introduction" le lendemain. C'est la raison pour laquelle, dans un but didactique, mon introduction présentera des éléments que vous retrouverez dans mes résultats, conduisant inévitablement à une certaine forme de redondance.

Étant bioinformaticien et ayant vécu l'évolution de la génomique de la paramécie liée à l'arrivée des technologies de séquençage à haut débit (ou NGS pour *Next Generation Sequencing*), j'ai choisi d'adopter une vision méthodologique pour mon introduction. Celle-ci est divisée en trois chapitres. Abordant des concepts biologiques assez généraux, le chapitre I (p.3) présente l'organisation et la composition des génomes eucaryotes, ainsi qu'une réflexion sur l'évolution des technologies de séquençage et l'impact significatif sur les méthodes pour obtenir les séquences de génomes entiers. Le chapitre II (p.31) est dédié aux méthodes d'annotation des gènes et des ET. Et enfin, le chapitre III (p.51) est consacré aux génomes des ciliés et plus particulièrement celui de la paramécie. Les résultats (p.89) reposent sur quatre publications relatives à l'annotation de génomes chez la paramécie (Arnaiz et al. 2012; 2017, Denby Wilkes et al. 2016, Guérin et al. 2017). Avant chaque article, je contextualiserai et résumerai les messages importants, et j'expliquerai pourquoi j'ai choisi de l'intégrer dans ce manuscrit. Dans la partie discussion et perspectives (p.157), j'exposerai des éléments de réflexion sur l'évolution de la génomique de la paramécie et j'énoncerai quelques uns de mes objectifs scientifiques pour les prochaines années.

Ma liste de publications

Arnaiz, O., Cain, S., Cohen, J., and Sperling, L. *ParameciumDB : a community resource that integrates the Paramecium tetraurelia genome sequence with genetic data*. Nucleic Acids Research, 35(Database issue) :D439–444 (2007). doi :10.1093/nar/gkl777. (Cité pages viii et ix.)

Arnaiz, O., Cohen, J., Tassin, A.-M., and Koll, F. *Remodeling Cildb, a popular database for cilia and links for ciliopathies*. Cilia, 3 :9 (2014). doi :10.1186/2046-2530-3-9. (Cité page x.)

Arnaiz, O., Goût, J.-F., Bétermier, M., Bouhouche, K., Cohen, J., Duret, L., Kapusta, A., Meyer, E., and Sperling, L. *Gene expression in a paleopolyploid : a transcriptome resource for the ciliate Paramecium tetraurelia*. BMC genomics, 11 :547 (2010). doi : 10.1186/1471-2164-11-547. (Cité page xi.)

Arnaiz, O., Malinowska, A., Klotz, C., Sperling, L., Dadlez, M., Koll, F., and Cohen, J. *Cildb : a knowledgebase for centrosomes and cilia*. Database : The Journal of Biological Databases and Curation, 2009 :bap022 (2009). doi :10.1093/database/bap022. (Cité page x.)

Arnaiz, O., Mathy, N., Baudry, C., Malinsky, S., Aury, J.-M., Denby Wilkes, C., Garnier, O., Labadie, K., Lauderdale, B. E., Le Mouël, A., Marmignon, A., Nowacki, M., Poulain, J., Prajer, M., Wincker, P., Meyer, E., Duharcourt, S., Duret, L., Bétermier, M., and Sperling, L. *The Paramecium germline genome provides a niche for intragenic parasitic DNA : evolutionary dynamics of internal eliminated sequences*. PLoS genetics, 8(10) :e1002984 (2012). doi :10.1371/journal.pgen.1002984. (Cité pages xi, xiii et xiv.)

Arnaiz, O., Meyer, E., and Sperling, L. *ParameciumDB 2019 : integrating genomic data across the genus for functional and evolutionary biology*. Nucleic Acids Research, (Database issue) (2019). (Cité page ix.)

Arnaiz, O. and Sperling, L. *ParameciumDB in 2011 : new tools and new data for functional and comparative genomics of the model ciliate Paramecium tetraurelia*. Nucleic Acids Research, 39(Database issue) :D632–636 (2011). doi :10.1093/nar/gkq918. (Cité page ix.)

Arnaiz, O., Van Dijk, E., Bétermier, M., Lhuillier-Akakpo, M., de Vanssay, A., Duharcourt, S., Sallet, E., Gouzy, J., and Sperling, L. *Improved methods and resources for paramecium genomics : transcription units, gene annotation and gene expression*. BMC genomics, 18(1) :483 (2017). doi :10.1186/s12864-017-3887-z. (Cité pages xi, xiii et xiv.)

Aury, J.-M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B. M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., Arnaiz, O., Billaut, A., Beisson, J., Blanc, I., Bouhouche, K., Câmara, F., Duharcourt, S., Guigo, R., Gogendeau, D., Katinka, M., Keller, A.-M., Kissmehl, R., Klotz, C., Koll, F., Le Mouël, A., Lepère, G., Malinsky, S., Nowacki, M., Nowak, J. K., Plattner, H., Poulain, J., Ruiz, F., Serrano, V., Zagulski, M., Dessen, P., Bétermier,

M., Weissenbach, J., Scarpelli, C., Schächter, V., Sperling, L., Meyer, E., Cohen, J., and Wincker, P. *Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia*. Nature, 444(7116) :171–178 (2006). doi :10.1038/nature05230. (Cité pages ix et xi.)

Bhullar, S., Denby Wilkes, C., Arnaiz, O., Nowacki, M., Sperling, L., and Meyer, E. *A mating-type mutagenesis screen identifies a zinc-finger protein required for specific DNA excision events in Paramecium*. Nucleic Acids Research, 46(18) :9550–9562 (2018). doi : 10.1093/nar/gky772.

Bischerour, J., Bhullar, S., Denby Wilkes, C., Régnier, V., Mathy, N., Dubois, E., Singh, A., Swart, E., Arnaiz, O., Sperling, L., Nowacki, M., and Bétermier, M. *Six domesticated PiggyBac transposases together carry out programmed DNA elimination in Paramecium*. eLife, 7 (2018). doi :10.7554/eLife.37927. (Cité page xii.)

Carradec, Q., Götz, U., Arnaiz, O., Pouch, J., Simon, M., Meyer, E., and Marker, S. *Primary and secondary siRNA synthesis triggered by RNAs from food bacteria in the ciliate Paramecium tetraurelia*. Nucleic Acids Research, 43(3) :1818–1833 (2015). doi :10.1093/nar/gku1331. (Cité page xi.)

Denby Wilkes, C., Arnaiz, O., and Sperling, L. *ParTIES : a toolbox for Paramecium interspersed DNA elimination studies*. Bioinformatics (Oxford, England), 32(4) :599–601 (2016). doi :10.1093/bioinformatics/btv691. (Cité pages xi, xiii et xiv.)

Duret, L., Cohen, J., Jubin, C., Dessen, P., Goût, J.-F., Mousset, S., Aury, J.-M., Jaillon, O., Noël, B., Arnaiz, O., Bétermier, M., Wincker, P., Meyer, E., and Sperling, L. *Analysis of sequence variability in the macronuclear DNA of Paramecium tetraurelia : a somatic view of the germline*. Genome Research, 18(4) :585–596 (2008). doi :10.1101/gr.074534.107. (Cité page xi.)

Frapporti, A., Miró Pina, C., Arnaiz, O., Holoch, D., Kawaguchi, T., Humbert, A., Eleftheriou, E., Lombard, B., Loew, D., Sperling, L., Guitot, K., Margueron, R., and Duharcourt, S. *The Polycomb protein Ezl1 mediates H3k9 and H3k27 methylation to repress transposable elements in Paramecium*. Nature Communications, 10(1) :2710 (2019). doi : 10.1038/s41467-019-10648-5. (Cité page xii.)

Gogendeau, D., Klotz, C., Arnaiz, O., Malinowska, A., Dadlez, M., de Loubresse, N. G., Ruiz, F., Koll, F., and Beisson, J. *Functional diversification of centrins and cell morphological complexity*. Journal of Cell Science, 121(Pt 1) :65–74 (2008). doi :10.1242/jcs.019414. (Cité page ix.)

Gogendeau, D., Lemullois, M., Aubusson-Fleury, A., Arnaiz, O., Cohen, J., Vesque, C., Schneider-Maunoury, S., Koll, F., and Tassin, A.-M. *MKS-NPHP module proteins regulate ciliary shedding in Paramecium*. bioRxiv, page 676395 (2019). doi :10.1101/676395. (Cité page ix.)

Gout, J.-F., Johri, P., Arnaiz, O., Doak, T. G., Bhullar, S., Couloux, A., Guérin, F., Malinsky, S., Sperling, L., Labadie, K., Meyer, E., Duharcourt, S., and Lynch, M. *Universal trends of post-duplication evolution revealed by the genomes of 13 Paramecium species sharing an ancestral whole-genome duplication*. bioRxiv, page 573576 (2019). doi :10.1101/573576. (Cité page xii.)

Gruchota, J., Denby Wilkes, C., Arnaiz, O., Sperling, L., and Nowak, J. K. *A meiosis-specific Spt5 homolog involved in non-coding transcription*. Nucleic Acids Research, 45(8) :4722–4732 (2017). doi :10.1093/nar/gkw1318. (Cité page xii.)

Guberman, J. M., Ai, J., Arnaiz, O., Baran, J., Blake, A., Baldock, R., Chelala, C., Croft, D., Cros, A., Cutts, R. J., Di Génova, A., Forbes, S., Fujisawa, T., Gadaleta, E., Goodstein, D. M., Gundem, G., Haggarty, B., Haider, S., Hall, M., Harris, T., Haw, R., Hu, S., Hubbard, S., Hsu, J., Iyer, V., Jones, P., Katayama, T., Kinsella, R., Kong, L., Lawson, D., Liang, Y., Lopez-Bigas, N., Luo, J., Lush, M., Mason, J., Moreews, F., Ndegwa, N., Oakley, D., Perez-Llamas, C., Primig, M., Rivkin, E., Rosanoff, S., Shepherd, R., Simon, R., Skarnes, B., Smedley, D., Sperling, L., Spooner, W., Stevenson, P., Stone, K., Teague, J., Wang, J., Wang, J., Whitty, B., Wong, D. T., Wong-Erasmus, M., Yao, L., Youens-Clark, K., Yung, C., Zhang, J., and Kasprzyk, A. *BioMart Central Portal : an open database network for the biological community*. Database : The Journal of Biological Databases and Curation, 2011 :bar041 (2011). doi :10.1093/database/bar041. (Cité page x.)

Guérin, F., Arnaiz, O., Boggetto, N., Denby Wilkes, C., Meyer, E., Sperling, L., and Duharcourt, S. *Flow cytometry sorting of nuclei enables the first global characterization of Paramecium germline DNA and transposable elements*. BMC genomics, 18(1) :327 (2017). doi : 10.1186/s12864-017-3713-7. (Cité pages xii, xiii et xiv.)

Maliszewska-Olejniczak, K., Gruchota, J., Gromadka, R., Denby Wilkes, C., Arnaiz, O., Mathy, N., Duharcourt, S., Bétermier, M., and Nowak, J. K. *TFIIS-Dependent Non-coding Transcription Regulates Developmental Genome Rearrangements*. PLoS genetics, 11(7) :e1005383 (2015). doi :10.1371/journal.pgen.1005383. (Cité page xii.)

Marker, S., Carradec, Q., Tanty, V., Arnaiz, O., and Meyer, E. *A forward genetic screen reveals essential and non-essential RNAi factors in Paramecium tetraurelia*. Nucleic Acids Research, 42(11) :7268–7280 (2014). doi :10.1093/nar/gku223. (Cité page xi.)

Marmignon, A., Bischerour, J., Silve, A., Fojcik, C., Dubois, E., Arnaiz, O., Kapusta, A., Malinsky, S., and Bétermier, M. *Ku-mediated coupling of DNA cleavage and repair during programmed genome rearrangements in the ciliate Paramecium tetraurelia*. PLoS genetics, 10(8) :e1004552 (2014). doi :10.1371/journal.pgen.1004552. (Cité page xii.)

Nekrasova, I., Nikitashina, V., Bhullar, S., Arnaiz, O., Singh, D. P., Meyer, E., and Potekhin, A. *Loss of a Fragile Chromosome Region leads to the Screwy Phenotype in Paramecium tetraurelia*. Genes, 10(7) (2019). doi :10.3390/genes10070513.

Nowacki, M., Higgins, B. P., Maquilan, G. M., Swart, E. C., Doak, T. G., and Landweber, L. F. *A functional role for transposases in a large eukaryotic genome*. Science, 324(5929) :935–8 (2009). doi :10.1126/science.1170023.

O'Connor, B. D., Day, A., Cain, S., Arnaiz, O., Sperling, L., and Stein, L. D. *GMODWeb : a web framework for the Generic Model Organism Database*. Genome Biology, 9(6) :R102 (2008). doi :10.1186/gb-2008-9-6-r102. (Cité page ix.)

Shi, L., Koll, F., Arnaiz, O., and Cohen, J. *The Ciliary Protein IFT57 in the Macronucleus of Paramecium*. The Journal of Eukaryotic Microbiology (2017). doi :10.1111/jeu.12423. (Cité page ix.)

Singh, D. P., Saudemont, B., Guglielmi, G., Arnaiz, O., Goût, J.-F., Prajer, M., Potekhin, A., Przybòs, E., Aubusson-Fleury, A., Bhullar, S., Bouhouche, K., Lhuillier-Akakpo, M., Tanty, V., Blugeon, C., Alberti, A., Labadie, K., Aury, J.-M., Sperling, L., Duharcourt, S., and Meyer, E. *Genome-defence small RNAs exapted for epigenetic mating-type inheritance*. Nature, 509(7501) :447–452 (2014). doi :10.1038/nature13318. (Cité pages xi et xii.)

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M. H., Baldock, R., Barbiera, G., Bardou, P., Beck, T., Blake, A., Bonierbale, M., Brookes, A. J., Bucci, G., Buetti, I., Burge, S., Cabau, C., Carlson, J. W., Chelala, C., Chrysostomou, C., Cittaro, D., Collin, O., Cordova, R., Cutts, R. J., Dassi, E., Di Genova, A., Djari, A., Esposito, A., Estrella, H., Eyras, E., Fernandez-Banet, J., Forbes, S., Free, R. C., Fujisawa, T., Gadaleta, E., Garcia-Manteiga, J. M., Goodstein, D., Gray, K., Guerra-Assunção, J. A., Haggarty, B., Han, D.-J., Han, B. W., Harris, T., Harshbarger, J., Hastings, R. K., Hayes, R. D., Hoede, C., Hu, S., Hu, Z.-L., Hutchins, L., Kan, Z., Kawaji, H., Keliet, A., Kerhornou, A., Kim, S., Kinsella, R., Klopp, C., Kong, L., Lawson, D., Lazarevic, D., Lee, J.-H., Letellier, T., Li, C.-Y., Lio, P., Liu, C.-J., Luo, J., Maass, A., Mariette, J., Maurel, T., Merella, S., Mohamed, A. M., Moreews, F., Nabihoudine, I., Ndegwa, N., Noirot, C., Perez-Llamas, C., Primig, M., Quattrone, A., Quesneville, H., Rambaldi, D., Reecy, J., Riba, M., Rosanoff, S., Saddiq, A. A., Salas, E., Sallou, O., Shepherd, R., Simon, R., Sperling, L., Spooner, W., Staines, D. M., Steinbach, D., Stone, K., Stupka, E., Teague, J. W., Dayem Ullah, A. Z., Wang, J., Ware, D., Wong-Erasmus, M., Youens-Clark, K., Zadissa, A., Zhang, S.-J., and Kasprzyk, A. *The BioMart community portal : an innovative alternative to large, centralized data repositories*. Nucleic Acids Research, 43(W1) :W589–598 (2015). doi :10.1093/nar/gkv350. (Cité page x.)

Table des matières

Remerciements **v**

Avant propos **vii**
.1 Contexte actuel de travail . . . vii
.2 Mon parcours professionnel et scientifique . . . viii
.2.0.1 Développement de systèmes d'information . . . ix
.2.0.2 Étude du génome de la paramécie . . . x
.2.0.3 Étude des réarrangements de génome . . . xii
.3 Pourquoi faire une thèse ? . . . xiii
.4 Ma thématique et l'organisation du manuscrit . . . xiii

Table des matières **xix**

Liste des figures **xxii**

Liste des tableaux **xxiv**

Liste des abréviations **xxv**

A Introduction **1**

I Les génomes eucaryotes 3
I.1 Caractéristiques des génomes eucaryotes . . . 3
I.1.1 Phylogénie . . . 3
I.1.2 Taille des génomes . . . 5
I.2 Organisation des génomes eucaryotes . . . 7
I.2.1 Structure des chromosomes . . . 7
I.2.2 Organisation génétique . . . 10
I.2.2.1 Les éléments répétés . . . 10
I.2.2.2 Les gènes . . . 12
I.3 Séquençage et assemblage de génomes . . . 16
I.3.1 Impact de l'évolution des technologies de séquençage . . . 16
I.3.2 Les technologies de séquençage . . . 19
I.3.2.1 Première génération . . . 20
I.3.2.2 Seconde génération . . . 20

I.3.2.3 Troisième génération 23
I.3.3 Les stratégies d'assemblage 26
I.3.4 Un assemblage de qualité 28

II Annotation des génomes **31**
II.1 Annotation de gènes 33
II.1.1 Masquer les répétitions 33
II.1.2 Les méthodes intrinsèques 34
II.1.3 Les méthodes extrinsèques 34
II.1.3.1 Évidences par homologie de séquence 35
II.1.3.2 Évidences d'expression 36
II.1.4 Méthodes intégratives 39
II.1.5 Curation humaine 40
II.2 Annotation des éléments transposables 42
II.2.1 Découverte et classification 42
II.2.1.1 Les éléments transposables à ARN 44
II.2.1.2 Les éléments transposables à ADN 44
II.2.2 Méthodes d'annotation 46
II.2.2.1 Approches basées sur la similarité de séquence 48
II.2.2.2 Approches *de novo* 48
II.2.2.3 Approches basées sur la structure des ET 49
II.2.2.4 Approches basées sur la génomique comparative 50

III La paramécie **51**
III.1 Place du modèle paramécie 51
III.1.1 Les eucaryotes 51
III.1.2 Les ciliés 53
III.1.3 Les paramécies 56
III.2 La biologie de la paramécie 58
III.2.1 Le cycle végétatif 58
III.2.2 Les processus sexuels 58
III.2.2.1 La conjugaison 58
III.2.2.2 L'autogamie 60
III.2.3 La génétique de la paramécie 61
III.2.3.1 Hérédité cytoplasmique 61
III.2.3.2 Outils moléculaires : transformation et extinction génique 61
III.3 Génomique de la paramécie 64
III.3.1 Les génomes micronucléaire et macronucléaire 64
III.3.2 Élimination des séquences micronucléaires 68
III.3.2.1 Les *Internal Eliminated Sequences* 68
III.3.2.2 Les séquences éliminées imprécisément 74
III.3.2.3 Le reconnaissance des séquences à éliminer 75

III.3.3 Les gènes codants 82
III.3.4 Duplication globale de génome 85

B Résultats **87**

IV Annotation du génome macronucléaire **89**

V Annotation du génome micronucléaire **105**
V.1 Annotation des IES 105
V.1.1 Identification des IES de *Paramecium tetraurelia* 105
V.1.2 ParTIES, *Paramecium Toolbox for Interspersed DNA Elimination Studies* 127
V.2 Annotation des éléments transposables 135

C Discussion et perspectives **155**

VI Les gènes et génomes MAC **159**
VI.1 La contrepartie d'une grande profondeur de séquençage 161
VI.2 Positionnement phylogénétique des WGD 165
VI.3 Les pseudogènes et gènes non-codants 167

VII Les génomes MIC **169**
VII.1 Les chromosomes MIC 169
VII.1.1 Assemblage des génomes MIC 169
VII.1.2 Quelle est la structure des chromosomes ? 170
VII.2 Que contiennent les génomes MIC ? 172
VII.2.1 Toutes les paramécies ont elles le même contenu en ET ? 172
VII.2.2 Les IES 174
VII.2.2.1 Toutes les paramécies ont elles des IES ? 174
VII.2.2.2 Avons nous catalogué toutes les IES ? 176
VII.2.2.3 Un rôle fonctionnel pour les IES ? 178
VII.2.3 Y a-t-il des gènes cachés dans le MIC ? 179

Bibliographie **181**

Liste des figures

1 Images de paramécies xi

I.1 Arbre phylogénétique des eucaryotes 4
I.2 Schéma représentant les gammes de taille de génome 5
I.3 Les niveaux de compaction de l'ADN dans une cellule eucaryote 8
I.4 Modèle du marquage épigénétique des chromosomes humains 9
I.5 Les types de séquences répétées et leurs mécanismes d'évolution 11
I.6 Structure d'un gène 14
I.7 Évolution du coût de sequençage d'une base d'ADN au cours du temps . . 17
I.8 Évolution du nombre de nucléotides assemblés et disponible au NCBI . . . 18
I.9 Principe de terminaison réversible cyclique de chaines 22
I.10 Plateformes de séquençage à lecture longue en temps réel 24
I.11 Principe d'assemblage de génome 27

II.1 Principes d'annotation pour les différents types de gènes 35
II.2 Répartition des séquences dans Uniprot par groupe taxonomique 37
II.3 Métriques de qualité pour l'annotation de génome 40
II.4 Interface WebApollo de ré-annotation de modèles de gène 41
II.5 La classification des transposons 43
II.6 Les grandes stratégies d'annotation des éléments transposables 47

III.1 Organisation cellulaire de la paramécie 52
III.2 Arbres phylogénétiques des eucaryotes 54
III.3 Phylogénie des alvéolés 55
III.4 Images de ciliés 55
III.5 Phylogénie des paramécies 57
III.6 Organisation nucléaire pendant le cycle sexuel de P. tetraurelia 59
III.7 Analyse génétique et type sexuel chez *P. tetraurelia* 62
III.8 Dimorphisme nucléaire et réarrangements de l'ADN chez *Paramecium* et *Tetrahymena* 65
III.9 Développement du génome macronucléaire chez *Oxytricha* 66
III.10 Propriétés de séquence des IES de *P. tetraurelia* 69
III.11 Modèle d'excision des IES chez *P. tetraurelia* 72
III.12 Réarrangements génomiques et organisation chromosomique 75

III.13Modèle de scanning chez *P. tetraurelia* 78
III.14Sensibilité des IES en fonction de leurs tailles 81
III.15Les introns chez *P. tetraurelia* 84
III.16Représentation des WGD successives du génome de *P. tetraurelia* 86

IV.1 Représentation circulaire de la synténie entre 2 chromosomes de *P. tetraurelia* 90

V.1 Protocole d'extraction d'ADN de cellules inactivées pour un gène 106
V.2 Densité en IES sur 8 grands chromosomes MAC 108
V.3 Les fonctionnalités de ParTIES 128
V.4 Scores de retention d'IES 130

VI.1 Représentation circulaire du scaffold 016 163
VI.2 Corrélation entre taille de génome et nombre de gènes ou d'IES 164
VI.3 Positionnement phylogénétique des WGD 166

VII.1 Complexité de génome occupée par les ET 173
VII.2 IES dans une région MIC spécifique 177

Liste des tableaux

I.1 Taille de génomes et nombre de gènes pour plusieurs organismes 6
I.2 Liste des génomes eucaryotes disponibles 19
I.3 Comparaison des technologies de séquençage 25

III.1 Statistiques sur les génomes haploïdes MIC et MAC de Tetrahymena, Paramecium et Oxytricha 64
III.2 Table récapitulative des IES sensibles 80
III.3 Caractéristiques des gènes 83

VI.1 Statistiques sur les génomes MIC et MAC, gènes et IES de paramécies . . . 160

Liste des abréviations

Dans ce document, par commodité j'ai choisi d'utiliser certains acronymes anglais.

ADN : Acide désoxyribonucléique

ADN-seq : Séquençage d'ADN par des technologies NGS

ARN : Acide ribonucléique

ARNm : ARN messager, intermédiaire pour la synthèse des protéines

ARN-seq : Séquençage d'ARN par des technologies NGS

BAC : *Bacterial Artificial Chromosome* ; Un chromosome bactérien artificiel

CDB : Cassure Double Brin de l'ADN

DIRS : *Dictyostelium Intermediate Repeat Sequence*

ET : Élément Transposable

EST : *Expressed Sequence Tag* ; un marqueur de séquence exprimée est une courte séquence d'un ADN complémentaire

FACS : *Fluorescence Activated Cell Sorting* ; tri cellulaire induit par fluorescence

GC : Contenu en dinucléotide G ou C

GMOD : Projet *Generic Model Organism Database*

HMM : *Hidden Markov Model* ; modèle de Markov caché

H3Kx : Histone H3 où x est le numéro de la lysine (K) modifiée

IES : *Internal Eliminated Sequence* ; séquence interstitielle éliminée lors des réarrangements de génome

indel : insertion ou délétion d'une ou plusieurs nucléotides dans une séquence, par rapport à une séquence de référence

Kb : Kilobase, 10^3 bases ou paires de bases

K-mer : sous-chaine de nucléotides de longueur K

LINE : *Long Interspersed Nuclear Element*

LTR : *Long Terminal Repeat* ; répétition terminale de grande taille

MAC : Macronoyau. Le noyau somatique ou génome somatique

MDS : *Macronucleus Destined Sequence* ; Sequences destinées au macronoyau

MIC : Micronoyau. Noyau germinal ou génome germinal

MICA : Méthode d'Identification par Comparaison d'Assemblages

MILORD : Méthode d'Identification ...

MIRAA : *Method of Identification by Read Alignment Anomalies* ; méthode d'identification par anomalies d'alignement de lectures

MITE : *Miniature Inverse-repeats Transposable Element*

Mb : Mégabase, 10^6 bases ou paires de bases

Ma : Million d'années

NGS : *Next Generation Sequencing* ; Nouvelle génération de technologie de séquençage

NHEJ : *Non-Homologous End Joining* ; ligature par jonction d'extrémités non-homologues

NMD : *None-sens Mediated Decay* ; système de reconnaissance et de dégradation des ARNm erronés

nt : nucléotide

ORF : *Open Reading Frame* ; phase ouverte de lecture

ONT : Technologie de séquençage *Oxford Nanopore Technology*

PacBio : Technologie de séquençage *Pacific Bioscience*

PCR : *Polymerase Chain Reaction* ; Réaction en chaîne par polymérase

pb : paires de bases

SNP : *Single Nucleotide Polymorphism* ; Polymorphisme nucléotidique au sein d'une même espèce

TA-indel : indel avec le dinucléotide TA à ses deux extrémités

TIR : *Terminal Inverted Repeat* ; répétition terminale inversée

UTR : *Untranslated Transcribed Region* ; Régions du transcrit non-traduites en protéine

WGD : *Whole Genome Duplication* ; duplication globale de génome

WT : *Wildtype* ; sauvage

Première partie

Introduction

Chapitre I

Les génomes eucaryotes

I.1 Caractéristiques des génomes eucaryotes

Mon travail de thèse porte sur l'annotation de génomes eucaryotes et plus particulièrement l'annotation du génome d'un cilié : la paramécie. Avant de définir où se placent phylogénétiquement les ciliés au sein des eucaryotes (voir **section III** (p.51)), il est important de définir ce qu'est un eucaryote. Les eucaryotes sont des organismes uni- ou pluri-cellulaires avec un compartiment nucléaire entouré par une membrane. Généralement, les cellules eucaryotes contiennent un organite mitochondrial, siège des voies respiratoires et énergétiques. Les plantes et les algues peuvent contenir des chloroplastes, lieu de la photosynthèse. Il y a environ 1.5 à 2 milliards d'années, des cyano- et eu-bactéries auraient été intégrées par endosymbiose aux cellules eucaryotes primitives (probablement des archées) pour donner les mitochondries et les chloroplastes (respectivement). Les mitochondries et les chloroplastes contiennent donc leurs propres séquences génomiques. Dans ce manuscrit, je vais me focaliser sur l'annotation des génomes nucléaires des organismes eucaryotes.

I.1.1 Phylogénie

En 1977, à l'aide d'études sur l'ARN ribosomique, Woese et al. (1990) ont décomposé le vivant en trois domaines : les bactéries, les archées et les eucaryotes. L'Empire eucaryote était divisé en cinq règnes : les protozoaires, les chromistes, les plantes, les champignons et les animaux. Aujourd'hui les eucaryotes sont classifiés en cinq super-groupes majeurs : SAR (avec les stramenopiles, les alvéolés, rhizaria, ...), Archaeplastida (plantes, algues,...), Excavata (flagellés), Amoebozoa (amibes, moisissures, ...) et les Opisthokonta (animaux, champignons, choanoflagellés, ...) (Adl et al. 2012). La **figure I.1** (p.4) montre les liens phylogénétiques entre ces groupes d'espèces. Les ciliés sont des alvéolés dans le super-groupe des SAR (**Figure III.2** p.54 et la section correspondante pour plus de détails). Reconstruire l'histoire évolutive de la vie est une tâche ardue. Elle passe par une classification des espèces les unes par rapport aux autres. Classiquement des mé-

thodes de génomique comparative sont utilisées. Cependant, nous n'avons accès qu'aux génomes d'une infime partie des organismes eucaryotes présents sur notre planète. De plus, ces organismes sont largement biaisés vers des organismes multicellulaires ayant divergé il y a seulement 550 millions d'années, alors que la vie serait apparue il y a ∼3 milliards d'années (Hugenholtz et al. 2016). Représentant la grande majorité de la biodiversité, les microorganismes sont parfois très difficiles à cultiver en laboratoire et donc à étudier. Aujourd'hui, les études métagénomiques sont aidées par une modernisation des techniques de biologie moléculaire et de séquençage (exemple le projet TARA Océan (Carradec et al. 2018)). Obtenir les génomes d'espèces représentatives de la diversité biologique permettra d'affiner les arbres phylogénétiques et de corriger les erreurs taxonomiques (Adl et al. 2019). Des efforts concertés de la part des experts scientifiques sont réalisés pour synthétiser et intégrer au mieux l'ensemble des données (Ruggiero et al. 2015). Par exemple, la communauté scientifique s'intéressant aux ciliés a établi un guide de recommandations des meilleures pratiques à adopter pour l'étude de la biodiversité, facilitant l'identification de nouvelles espèces (Warren et al. 2017).

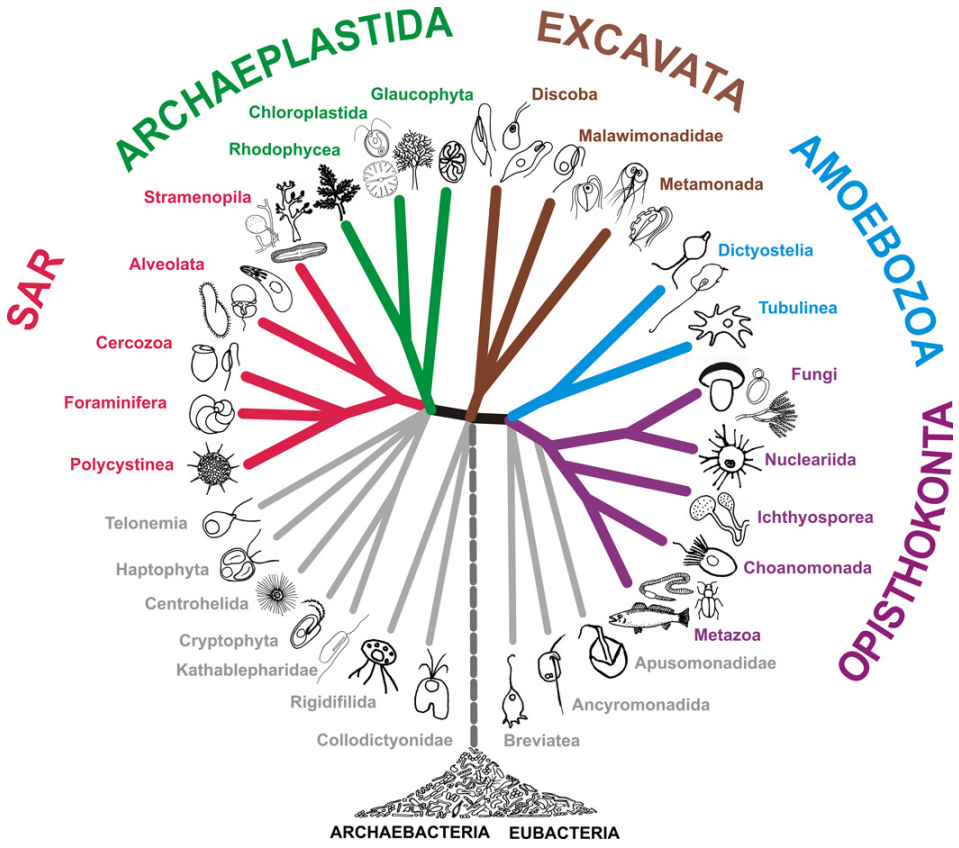


Figure I.1 – *Arbre phylogénétique des eucaryotes*

Classification des eucaryotes les divisant en cinq super-groupes principaux d'espèces. Les SAR, les Archaeplastida, les Excavata, les Amoebozoa et Opisthokonta. Figure tirée de Adl et al. (2012).

I.1.2 Taille des génomes

Les organismes ont des tailles de génome très diverses. La taille d'un génome est définie par le nombre de nucléotides (ou bases) qui composent son génome haploïde, ou par une valeur "C" en pg d'ADN (1 pg équivalent à 978 Mb). La "complexité" d'un organisme (terme vague je l'admets) n'est pas corrélée à la taille de son génome. Illustré par la **figure I.2** (p.5), certaines plantes ont des génomes beaucoup plus grands que le génome humain. Ce paradoxe de la valeur "C" peut s'expliquer par la présence de plus ou moins de séquences non-codantes ou répétées dans les génomes mais également par des événements de polyploïdisation.

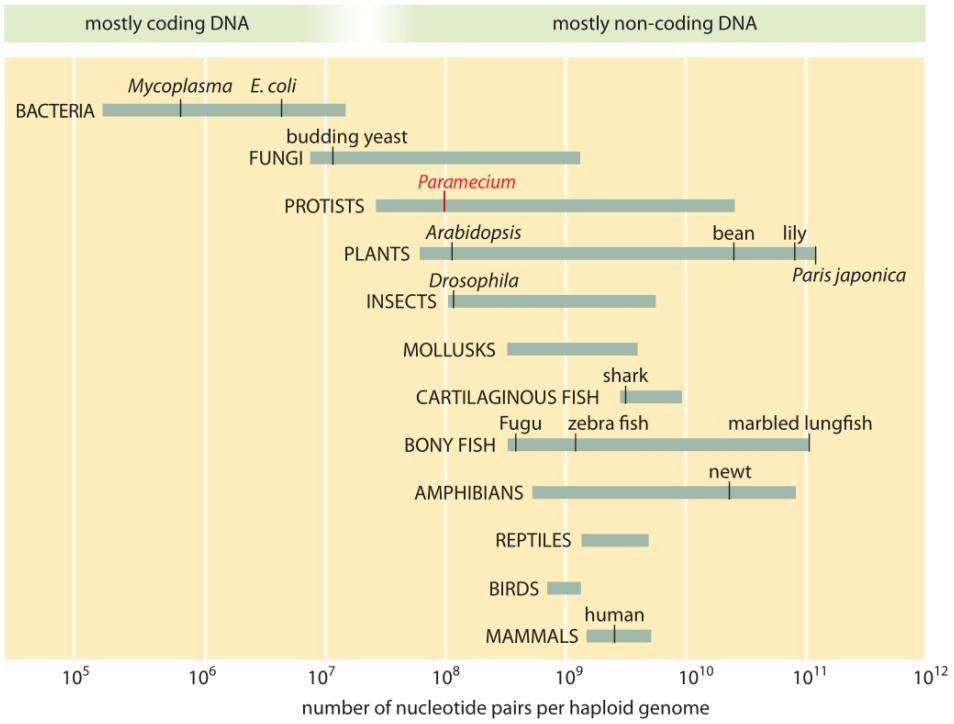


Figure I.2 – *Schéma représentant les gammes de taille de génome*

Dans un même groupe d'organismes les tailles de génomes peuvent varier de façon importante. La "complexité" de l'organisme n'est pas reliée à la taille de son génome. Figure tirée et modifiée du site http ://book.bionumbers.org/how-big-are-genomes/

A titre d'exemple, le génome hexaploïde du blé tendre (*Triticum aestivum*) est de 16 Gb, alors que l'homme n'a qu'un génome de 3.2 Gb. La plante *Paris japonica* a un génome de 150 Gb, un des génomes les plus grands connus (Pellicer et al. 2010). Et l'historiquement célèbre petit pois de Gregor Mendel a un génome diploïde 1.5 fois plus grand que le génome humain (4.5Gb), notamment dû à beaucoup de séquences hautement répétées (Kreplak et al. 2019). Le nombre de gènes codants n'est pas non plus relié à la taille du génome. Le nématode *Caenorhabditis elegans* compte environ 20 000 gènes pour un génome de 100 Mb, alors que l'homme a le même nombre de gènes mais un génome de 3 Gb (**Table I.1** p.6). Bien évidemment, il faut relativiser ces observations sur le nombre de gènes

compte tenu de l'épissage alternatif introduisant une diversité de protéines produites très importante, et notamment chez l'humain.

Groupe	**Espèce**	**Taille de génome**	**Nombre de gènes codants**	**Nombre de chromosomes haploïdes**
Bactérie	*E. coli*	4.6 Mb	4300	1
Champignon	*S cerevisiae*	12 Mb	6600	16
Protiste	*P. tetraurelia*	100 Mb	40000	?
Plante	*A. thaliana*	140 Mb	27000	5 (2n)
	P. japonica	150 Gb	?	20 (2n)
Insecte	*D. melanogaster*	140 Mb	14000	4 (2n)
Nématode	*C. elegans*	100 Mb	20000	6 (2n)
Poisson	*F. rubripes* (fugu)	400 Mb	19000	22
Mammifère	*H. sapiens*	3.2 Gb	21000	23 (2n)

TABLE I.1 – *Taille de génomes et nombre de gènes pour plusieurs organismes*

Le tableau rapporte la taille de génome, le nombre de gènes codants et le nombre de chromosomes pour une variété d'organismes. Informations tirées en partie du site http ://book.bionumbers.org/how-big-are-genomes/

I.2 Organisation des génomes eucaryotes

I.2.1 Structure des chromosomes

Le génome d'un organisme est l'ensemble des molécules d'ADN différentes contenues dans chacune de ses cellules. La double hélice d'ADN est compactée sous forme de chromosome(s). Alors que les procaryotes ont leur matériel génétique organisé dans un (ou plusieurs) chromosome généralement circulaire, les génomes eucaryotes se structurent sous la forme de chromosome(s) linéaire(s) (**Figure I.3** p. 8). Les télomères marquent les extrémités des chromosomes. Les télomères sont généralement des séquences non codantes, hautement répétées et ont un rôle protecteur des chromosomes. Sans télomères, la cellule perdrait de l'information génétique à chaque cycle de réplication car l'ADN polymérase ne peut pas copier les derniers nucléotides aux extrémités. Quand les télomères deviennent trop courts, la cellule entre en sénescence : c'est le vieillissement cellulaire. Aidées par une amplification particulièrement importante des séquences télomériques chez les ciliés, Greider and Blackburn (1985) ont découvert, chez *Tetrahymena*, l'existence de la télomérase impliquée dans la synthèse des télomères et donc la protection des chromosomes (E. Blackburn et C. Greider ont obtenu le prix Nobel de physiologie ou médecine en 2009 avec J. Szostak). Les chromosomes possèdent également des centromères composés de séquences répétées. Pendant la mitose, les centromères permettent l'assemblage du kinétochore pour séparer les chromosomes dans les cellules filles. Il existe deux types de centromères. Chez l'humain, le centromère monocentrique occupe une région spécifique du chromosome. Au contraire, les centromères holocentriques sont répartis sur plusieurs zones du chromosome (par exemple chez *C. elegans*) (Plohl et al. 2014).

L'ADN nucléaire des cellules eucaryotes est enroulé autour d'octamères d'histones (deux copies des histones H2A, H2B, H3 et H4) formant le nucléosome, unité de la chromatine (**Figure I.4** p.9). Un variant de l'histone H3 (CenH3) est utilisé pour enrouler l'ADN centromérique (Black and Bassett 2008). Par cytologie, on peut distinguer deux types de chromatine. L'euchromatine est peu compacte et l'ADN est assez accessible aux machineries cellulaires. En général, ces régions correspondent à des régions transcrites et donc des gènes exprimés. A l'inverse, l'hétérochromatine est très compacte et l'ADN y est peu accessible aux polymérases. Ces régions, transcriptionellement inactives, sont en général pauvres en gènes et riches en séquences répétées (**Figure I.4** p.9).

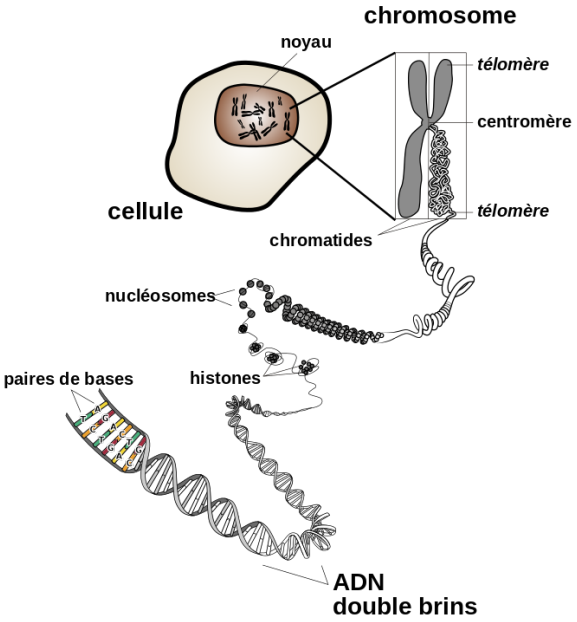


FIGURE I.3 – *Les niveaux de compaction de l'ADN dans une cellule eucaryote*

Tiré du site https ://fr.wikipedia.org/wiki/Chromosome

L'héterochromatine et l'euchromatine se distinguent par des contextes épigénétiques différents. En plus de la méthylation de l'ADN, les histones peuvent etre modifiées post-traductionnellement par l'ajout de groupements sur leurs queues N-terminales non structurées. Les principales marques d'histones sont la méthylation, l'acétylation, la phosphorylation et l'ubiquitination. Ces modifications peuvent se positionner sur des résidus différents de la queue d'histone et avec un nombre variable de groupements. Par exemple, une mono-méthylation, une di-méthylation ou une tri-méthylation sont possibles sur les lysines 9 et 27 de l'histone H3. Suivant les marques, la structure chromatinienne change, et donc l'accessibilité de l'ADN aux machineries cellulaires, définissant ainsi un véritable code des histones (STRAHL AND ALLIS 2000). L'hétérochromatine constitutive est marquée par la di- et tri-méthylation de l'histone H3 sur la lysine 9 (H3K9me2 et H3K9me3) alors que l'acétylation est caractéristique de l'euchromatine dite "ouverte". Au sein de l'euchromatine, les gènes inactifs sont marqués par une tri-méthylation de la lysine 27 (H3K27me3) (BARSKI ET AL. 2007). La méthylation de la lysine 4 de l'histone H3 (H3K4me) est plutôt localisée au début des gènes et marque les sites de début de transcription (SANTOS-ROSA ET AL. 2002). Une haute densité en H3K36me3 indique la fin des gènes actifs. Enfin, la monométhylation de H3K4, H3K9, H3K27, H4K20 et H2BK5 pointe les régions transcrites activement (SCHONES AND ZHAO 2008).

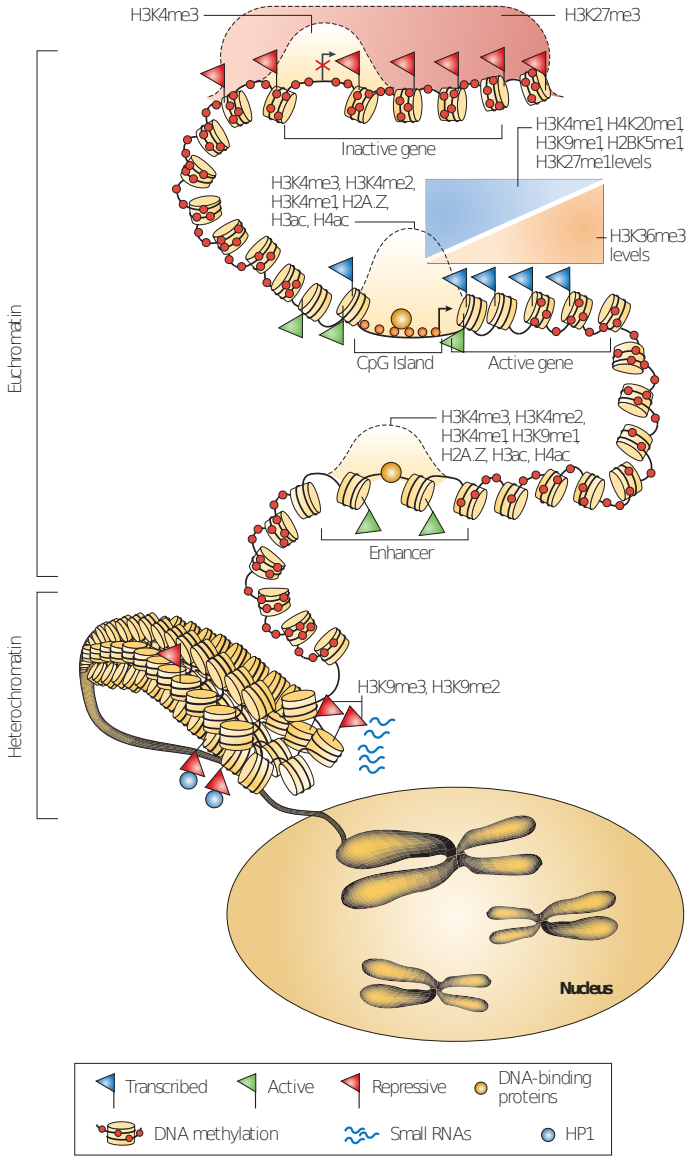


FIGURE I.4 – *Modèle du marquage épigénétique des chromosomes humains*

Les chromosomes sont divisés en régions accessibles (euchromatine), et faiblement accessibles (hétérochromatine). L'hétérochromatine est marquée par la di- et triméthylation de l'histone H3 sur la lysine 9 (H3K9me2 et H3K9me3). La méthylation de l'ADN est présente à travers tout le génome sauf sur la plupart des îlots CpG, les promoteurs et éventuellement les séquences régulatrices. La marque H3K27me3 est présente sur de larges domaines comprenant des gènes inactifs. Les modification H3K4me3, H3K4me2, H3K4me1 ainsi que l'acétylation du variant H2A.Z marquent les sites de début de transcription. Enfin la monométhylation de H3K4, H3K9, H3K27, H4K20 et H2BK5 marque les régions transcrites activement, avec un pic à l'extrémité 5′ des gènes, alors que la triméthylation de H3K36 marque leur extrémité 3′. Figure tirée de SCHONES AND ZHAO (2008).

I.2.2 Organisation génétique

Le génome d'un organisme est l'ensemble de son ADN pouvant être décrit sous la forme de séquence(s). Avoir la séquence d'un génome n'est que la première étape. Décrypter l'information qu'il code est essentiel. Autrement dit, il faut l'annoter. L'annotation de génome étant au cœur de mon travail et de ce manuscrit, il est important de décrire les éléments à annoter. Dans cette section, je vais découper les éléments composant les génomes en deux parties : les séquences répétées telles que les satellites ou éléments transposables et les gènes, qu'ils soient codants ou non.

I.2.2.1 Les éléments répétés

Les génomes procaryotes sont constitués essentiellement de gènes codants (87% codant pour *E coli*). A l'inverse les génomes eucaryotes contiennent, en général, beaucoup moins de régions codantes (70% *S cerevisiae*, 28% *A thaliana* et 1.3% codant pour *H sapiens*). Les séquences non codantes regroupent notamment les séquences répétées, les régions intergéniques et les séquences interrompant les régions codantes et devant être épissées lors de la maturation des transcrits (voir **section I.2.2.2** p.12).

Les régions répétées peuvent occuper une grande proportion des génomes. En effet, plus de 50% du génome humain est composé de séquences répétées (Jurka et al. 2007). Pour Richard et al. (2008), les séquences répétées sont classées en deux familles : les répétitions en tandem et les répétitions dispersées dans les génomes (**Figure I.5** p. 11). Les répétitions en tandem intègrent les gènes paralogues en tandem, les satellites (minisatellites et microsatellites) (Walker 1971) et les répétitions de l'ADN ribosomal. Les répétitions dispersées regroupent les éléments transposables, les tRNA, des gènes paralogues et certains pseudogènes. De plus, des événements dramatiques comme les duplications globales de génome engendrent une forme de répétition (la **section III.3.4** p.85 est consacrée aux duplications globales de génome (WGD pour *Whole Genome Duplication*), caractéristiques des génomes de paramécies) .

Les satellites sont des séquences répétées en tandem pouvant atteindre plusieurs millions de paires de bases. Les centromères et télomères, et plus généralement l'hétérochromatine, sont composés de satellites. Suivant la taille de l'unité de répétition, elles sont cataloguées en micro- ou mini-satellites. Les microsatellites sont des répétitions contiguës d'un motif de 1 à 6 nucléotides (ou 1 à 12 nucléotides selon les sources). Il en existe plusieurs milliers d'occurrences dans les génomes (1 818 chez *S. cerevisiae*, ∼38 000 pour *A. thaliana* et ∼253 000 microsatellites chez *H. sapiens*) (Richard et al. 2008). Ils sont très utilisés comme marqueurs génétiques chez l'humain. Les minisatellites sont des répétitions de 10 à 500 paires de bases. On les retrouve à plusieurs centaines ou milliers d'occurrences dans les génomes (113 chez *S. cerevisiae*, ∼720 pour *A. thaliana* et ∼6 000 minisatellites chez *H. sapiens*) (Richard et al. 2008). Majoritairement non fonctionnels,

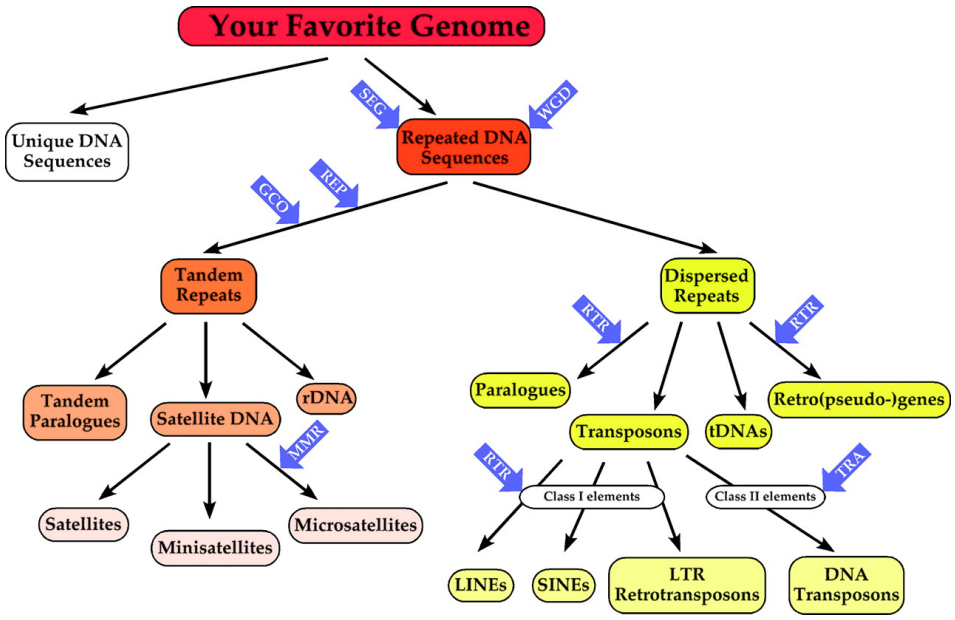


FIGURE I.5 – *Les types de séquences répétées et leurs mécanismes d'évolution*

Les deux principales catégories d'éléments répétés (répétitions tandem et répétées dispersées) sont présentées, ainsi que les sous-catégories, comme décrit dans le texte. Les flèches bleues indiquent les mécanismes moléculaires impliqués dans la propagation et l'évolution de séquences répétées. REP, glissement de réplication ; GCO, conversion de gène ; WGD, duplication du génome entier ; SEG, duplications segmentaires ; RTR, transcription inverse ; TRA, transposition Figure tirée de RICHARD ET AL. (2008).

certains minisatellites ont pourtant un rôle dans la régulation transcriptionnelle des gènes (Kennedy et al. 1995).

Les éléments transposables (ET), aussi appelés transposons, sont des éléments génomiques pouvant migrer d'un *locus* à un autre de manière plus ou moins autonome. Récompensée par un prix Nobel en 1983, B. McClintock a découvert les éléments mobiles lors de ses études cytogénétiques sur le maïs dans les années 1930-1940, bien avant l'avènement des approches de biologie moléculaire (McClintock 1950). Leurs séquences peuvent contenir un gène codant pour une transposase ou une intégrase, permettant leur déplacement dans les génomes. Les ET, aussi qualifiés de parasites moléculaires, peuvent occuper une proportion importante des génomes. Par exemple, 45% du génome humain est constitué d'ET (Jurka et al. 2007). Bien que délétères en cas d'insertion dans un gène essentiel, ces éléments sont également considérés comme un des moteurs de l'évolution et de la biodiversité des espèces (Platt et al. 2018). Les organismes ont trouvé des moyens de contrôler la transposition des ET, notamment via une hétérochromatinisation des séquences (voir **section I.2.1** p.8). N'étant pas sous pression sélective les copies d'ET accumulent énormément de mutations, les rendant compliquées à détecter. Les ET sont catégorisés en deux classes suivant le type de molécule qu'ils utilisent pour se mobiliser (Finnegan 1989). Les ET de classe I, aussi appelés retroélements ou rétrotransposons, utilisent un intermédiaire ARN pour transposer selon le principe du "copier-coller". Après transcription et retro-transcription, ces éléments s'intègrent dans un nouveau *locus*. Les ET de classe II utilisent un intermédiaire ADN pour transposer. Les ET de classe II les plus étudiés transposent selon le principe du "couper-coller". En effet, ces éléments s'excisent de leur *locus* d'origine, en laissant plus ou moins de cicatrices, témoin de leur passage, et colonisent d'autres *loci*. Une classification unifiée des ET a été proposée par Wicker et al. (2007) basée sur la structure des éléments et leurs mécanismes de transposition. Les différentes classes et familles d'ET seront décrites avec plus de détails dans la **section II.2** (p.42).

I.2.2.2 Les gènes

Précédemment, nous avons vu que les séquences répétées, globalement non codantes, pouvaient occuper une grande proportion des génomes. Pourtant, la partie codante, n'occupant qu'1% du génome humain, est d'un intérêt tout particulier car l'expression des gènes codants détermine le phénotype de l'organisme. D'après Gerstein et al. (2007), un gène est une union de séquences génomiques codant pour un ensemble cohérent de produit(s) fonctionnel(s). Il existe une variété de types de gènes. Dans les paragraphes suivants, je vais différencier les gènes non-codants, les gènes codants et les pseudogènes. Même si ces descriptions resteront brèves et un peu scolaires, il me paraissait important de définir les éléments que j'annote.

Les gènes non codants sont transcrits mais ne codent pas pour des protéines. La très grande majorité de ces transcrits ont un rôle fonctionnel (Eddy 2001). Les ARN de transfert (ARNt) traduisent l'information des triplets de codon de l'ARNm en acides aminés. Les ARN ribosomiques (ARNr), représentant 80% de l'ARN total de la cellule, sont au cœur de l'ossature du ribosome, un complexe ribonucléoprotéique. Les snoARN (*Small Nucleolar RNA*) ont une fonction dans les nucléoles et les snRNA (*Small Nuclear RNA*) dans le spliceosome pour épisser les introns de l'ARN primaire. Les microARN (*miARN*) d'une vingtaine de nucléotides ont des rôles dans la régulation des ARN (Ambros 2001).

Les gènes codants La **figure I.6** (p. 14) schématise la structure d'un gène codant eucaryote (Shafee 2017). Un brin d'ADN est transcrit dans le sens 5' vers 3' par l'ARN polymérase II. Un gène codant est composé d'un cadre ouvert de lecture (ORF *Open Reading Frame*) et de séquences régulatrices. Sans rentrer dans des détails sémantiques, aujourd'hui, et notamment en bioinformatique, une ORF est définit préférentiellement comme une séquence séparée par deux codons terminateurs (discuté par Sieber et al. (2018)), alors que la séquence codante (ou CDS pour *CoDing Sequence*) démarre par un codon initiateur et se finit par un codon terminateur.

Les séquences régulatrices, localisées aux extrémités du gène, sont composées d'une région promotrice et de séquences activatrices ou répressives. Ces dernières peuvent être éloignées du gène et modulent l'activité des promoteurs. Le promoteur, situé à l'extrémité 5' de la CDS, est constitué d'une partie principale et d'une partie proximale. Le promoteur principal fixe l'ARN polymérase et définit le site de départ de la transcription (TSS *Transcription Start Site*). Il peut contenir des motifs de type TATA ou CCAAT à environ 20-30 paires de base en amont du TSS. La région proximale du promoteur se lie à des facteurs de transcription modifiant l'affinité du promoteur principal pour l'ARN polymérase (Juven-Gershon et al. 2008, Haberle and Stark 2018). Les facteurs de transcription régulent la transcription des gènes en fonction du stade cellulaire ou du type cellulaire (Andersson et al. 2015). La partie transcrite du gène est composée d'exons et d'introns pour générer une molécule de pré-ARN messager. Après épissage des introns, seuls les exons seront retenus dans l'ARN messager mature (Matera and Wang 2014). Une coiffe est ajoutée en 5' et une queue de poly-adénosine en 3', stabilisant la molécule (Wu and Brewer 2012).

Après transport dans le cytoplasme, la traduction de l'ARNm débutera au codon initiateur (ou codon *start*) (en général une méthionine) et se finira au codon terminateur (ou codon *stop*). Grâce aux ARNt, chaque codon (triplet de nucléotides) va correspondre à un acide aminé. En revanche, un acide aminé se rapporte à plusieurs codons (le code génétique). Tous les codons, codant pour le même acide aminé, ne sont pas retrouvés aux mêmes fréquences dans les séquences. Ce biais d'utilisation de codons varie d'une espèce à une autre et peut également varier entre gènes d'un même organisme. En effet, il est connu que l'usage des codons a des implications fonctionnelles pour le contrôle de la traduction et l'expression des gènes (Quax et al. 2015). Il faut aussi noter que tous

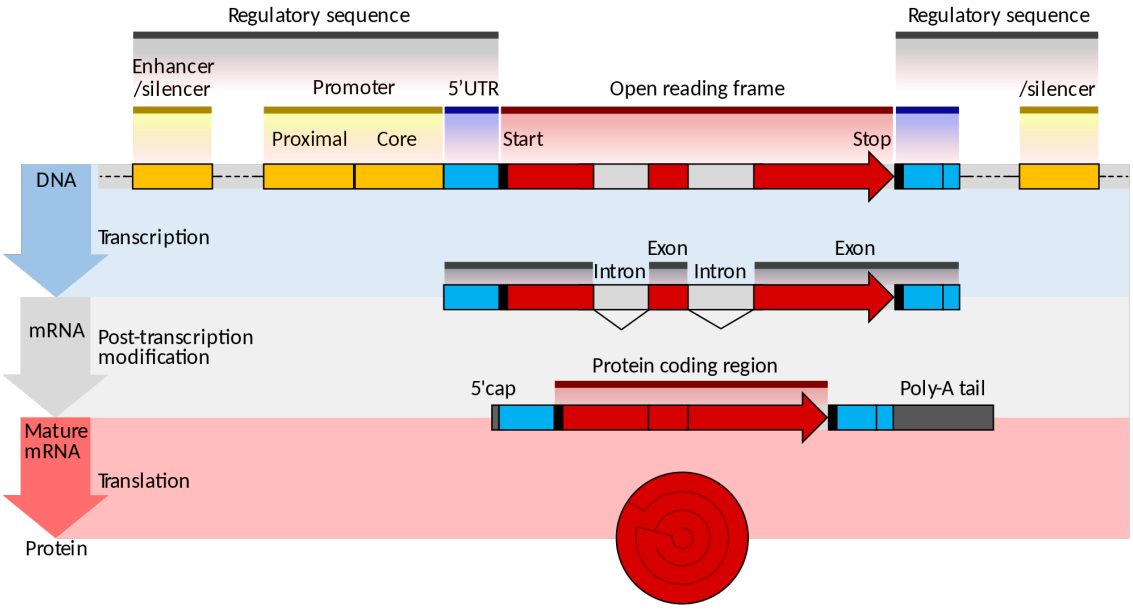


FIGURE I.6 – *Structure d'un gène*

La structure d'un gène eucaryote codant pour une protéine. La séquence régulatrice contrôle quand et où l'expression se produit pour la région codant pour la protéine (rouge). Les régions promotrices et activatrices (jaune) régulent la transcription du gène en un pré-ARNm modifié pour éliminer les introns (gris clair) et ajoutent une coiffe 5 ′et une queue poly-A (gris foncé). Les régions non traduites 5 ′et 3′ de l'ARNm (bleu) régulent la traduction dans le produit protéique final. Figure tirée de SHAFEE (2017)

les organismes n'utilisent pas le même code génétique (Osawa et al. 1992, Jukes and Osawa 1993). Une douzaine de codes alternatifs sont répertoriés. Par exemple, le génome mitochondrial des vertébrés utilise des codons terminateurs différents du génome nucléaire (Barrell et al. 1979). Le génome nucléaire de la paramécie utilise le code "*The Ciliate, Dasycladacean and Hexamita Nuclear Code*" (numéro 6) et sa mitochondrie le code "*The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code*" (numéro 4) (Caron and Meyer 1985, Preer et al. 1985, Pritchard et al. 1990). Dans le code génétique utilisé par les ciliés, il n'existe qu'un seul codon initiateur (ATG pour la Méthionine) et un seul codon terminateur (TGA).

Les exons non traduits ou UTR (*UnTranslated Region*) contiennent des éléments régulateurs (Wu and Brewer 2012). La séquence 3'UTR participe à la signalisation de la fin de la transcription à l'ARN polymérase. La séquence 5'UTR lie le ribosome pour la traduction. De plus, les protéines peuvent être maturées par des modifications post-traductionnelles (acétylation, biotinylation, méthylation, ...) altérant ou modifiant leurs fonctions. Le code des histones de la **section I.2.1** (p.8) en est un exemple. Il est vrai que la **figure I.6** (p. I.6) est une vision simplifiée du gène. En effet deux gènes peuvent se chevaucher, notamment au niveau des séquences régulatrices. De plus, l'épissage alternatif (Kornblihtt et al. 2013) et le trans-épissage (Lasda and Blumenthal 2011, Matera and Wang 2014) participent à la diversité des molécules produites par un gène.

Les pseudogènes ne produisent pas de protéines, en raison de la présence de codons stop interrompant leurs phases ouvertes de lecture ou à un changement de phase de lecture suite à des *InDels* (Ohno 1972). Les séquences des pseudogènes, n'étant plus sous pression sélective, s'éloignent progressivement de la composition de l'ADN codant et se rapprochent d'une séquence aléatoire (Echols et al. 2002). Comme les élements transposables, plus le pseudogène accumule des mutations, plus son identification sera difficile. Il existe deux classes de pseudogènes. Les pseudogènes processés (Vanin 1985) sont issus de la retrotranscription d'un ARNm par une reverse transcriptase de transposon (le plus souvent d'un rétrotransposon de type LINE). L'ADN codant est ensuite intégré dans le génome. Par définition, ces gènes sont dépourvus d'introns et de séquences régulatrices. Ces rétropseudogènes sont considérés comme des séquences répétées (voir **figure I.5** dans la **section I.2.2.1** p.10). Les pseudogènes non processés sont issus d'une duplication de gène. La plupart des pseudogènes n'ont pas de fonction mais plusieurs études ont démontré un possible rôle des pseudogènes dans la régulation de l'expression et la fonction des gènes (Balakirev and Ayala 2003, Pavlicek et al. 2006, Kovalenko and Patrushev 2018).

I.3 Séquençage et assemblage de génomes

J'ai commencé à faire de la génomique à partir de 2006. En 2007, j'ai vécu l'arrivée des nouvelles technologies de séquençage (NGS pour *Next Generation Sequencing*). Cette révolution a profondément changé mon activité. C'est la raison pour laquelle je trouvais intéressant de consacrer une partie de ce manuscrit à l'évolution des technologies de séquençage et à l'impact qu'elle a eu sur les études génomiques. Un petit zoom sera fait sur l'épopée du séquençage du génome humain.

I.3.1 Impact de l'évolution des technologies de séquençage

Le séquençage détermine l'ordre des nucléotides dans une molécule d'ADN. En 1977, la méthode de séquençage révolutionnaire *Sanger* a été utilisée pour séquencer l'ADN simple brin du premier génome d'un bactériophage (Sanger et al. 1977). Aujourd'hui, les technologies de séquençage nous permettent d'accéder à de petits génomes complets très simplement. Pourtant, et même si les techniques se sont largement améliorées, les génomes plus imposants restent encore compliqués à obtenir. Pour séquencer ces grands génomes, l'ADN est fragmenté par des méthodes chimiques ou physiques. Deux stratégies peuvent être utilisées avant le séquençage : l'ordonnancement hiérarchique ou l'approche globale. L'approche par ordonnancement hiérarchique trie les fragments clonés dans des vecteurs par hybridation spécifique ou par profil de restriction. Très utile pour les grands génomes, cette approche permet aux grands consortiums de se répartir le travail. L'approche de séquençage globale (WGS pour *Whole Genome Shotgun*) n'implique pas de trier les fragments avant le séquençage (Staden 1979). Plusieurs banques de fragments aléatoires d'ADN sont générées. Des logiciels, appelés assembleurs, réalisent un traitement bioinformatique classant et ordonnançant les fragments aléatoires produits et reconstituent la séquence et la structure du génome d'intérêt (voir **section I.3.3** p.26). Popularisée par le projet de séquençage du génome humain, l'approche WGS est probablement la plus employée aujourd'hui.

Le projet du génome humain est lancé dans les années 1990 par un consortium international de laboratoires. Les scientifiques se répartissent les tâches et utilisent des cartes génétiques d'ordonnancement des 23 chromosomes humains. En 1998, Craig Venter fonde une entreprise privée (*Celera genomics*) et acquiert de nombreux séquenceurs. Avec la méthode WGS, moins chère et plus rapide, il clame qu'il décryptera le génome en quelques années. La course commence entre le consortium international et *Celera Genomics*. La même année, les deux versions des séquences brutes sont publiées dans *Nature* (Lander et al. 2001) pour le consortium public et dans *Science* (Venter et al. 2001) pour *Celera genomics*. Je présume que la course a été remportée par le consortium international car *Celera Genomics* a avoué s'être appuyée sur des données mises à disposition par le consortium public. Toujours est-il, l'approche de Venter et ses collègues sera promise à un grand

avenir. En 2003, avec une version finalisée de l'assemblage, la fin du projet de séquençage du génome humain est annoncée (International Human Genome Sequencing Consortium 2004). Et des projets comme ENCODE (Consortium 2004) (*Encyclopedia of DNA Elements*) voient le jour pour identifier les éléments fonctionnels du génome humain.

La méthode de séquençage *Sanger* génère des milliers de fragments (ou lectures) de plusieurs centaines de bases de longueur (Sanger et al. 1977). En 2007, les technologies de séquençage de seconde génération apparaissent. Ces technologies NGS sont basées sur une parallélisation massive des procédures, faisant entrer la biologie dans l'ère des *Big Data*. En comparaison à la méthode *Sanger*, ces nouvelles méthodes génèrent un nombre de lectures courtes beaucoup plus important, et à moindre coût. Par exemple, le projet de séquençage du génome humain, utilisant un séquençage de type *Sanger*, a couté environ 3 milliards de dollars sur dix ans. Aujourd'hui en trois jours et pour 1000 dollars, on peut re-séquencer le génome humain. La **figure I.7** (p. 17) illustre l'évolution du coût de séquençage par mégabase d'ADN. La droite blanche sur la **figure I.7** présente la loi de *Moore*, qui décrit l'évolution de la puissance informatique en fonction du prix, caractérisée par un doublement tous les deux ans. En 2007, la chute drastique du coût de la base séquencée correspond au remplacement des séquenceurs *Sanger* par des séquenceurs de seconde génération. A partir de cette date, les technologies de séquençage évoluent plus rapidement que la puissance informatique.

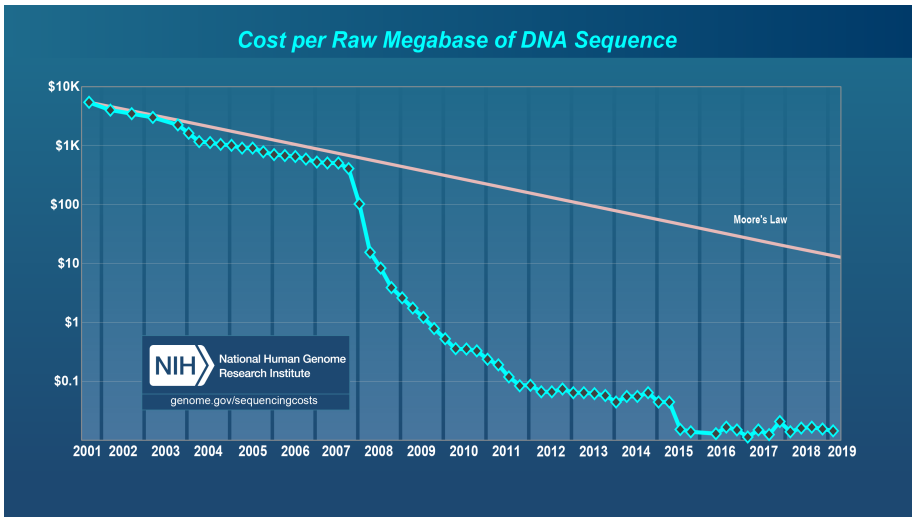


Figure I.7 – *Évolution du coût de sequençage d'une base d'ADN au cours du temps*

Données issues du *National Human Genome Research Institute* (NHGRI)

Plus récemment, des technologies de séquençage de troisième génération ont vu le jour (Schadt et al. 2010). Elles génèrent des séquences beaucoup plus longues et avec un débit aussi important que les technologies de seconde génération. La baisse du coût de séquençage de 2015 sur la **figure I.7** (p. 17) cadre avec l'essor de ces nouvelles méthodes. Cependant, ces techniques souffrent encore d'un inconvénient de taille : le taux d'erreur

de séquençage. En effet, le taux d'erreurs varie entre 5 et 15% suivant les technologies, contre <1% pour les NGS de deuxième génération (voir **Table I.2** p.19). En constante évolution, ce problème va probablement être amélioré dans un futur proche (Heather and Chain 2016).

La chute du coût de séquençage a eu un impact important sur le nombre de projets de séquençage de génomes. La **figure I.8** (p. 18) montre l'évolution exponentielle du nombre de bases assemblées et disponibles dans les bases de données publiques. La **Table I.2** (p.19) référence les 8832 génomes eucaryotes accessibles, avec 26% d'animaux, 54% de champignons, 10% de plantes et 8% de protistes. Même si les champignons sont particulièrement bien représentés, ils n'occupent que 4% de la complexité assemblée, alors que les mammifères en occupent 45%. Avoir accès au génome d'une espèce devient si aisé que les scientifiques peuvent maintenant s'intéresser à la variabilité entre individus au sein d'une même espèce. En effet, un assemblage de génome n'est qu'une séquence consensus de plusieurs individus d'une espèce ou la séquence d'un individu d'une espèce. En étudiant la variabilité entre individus, les chercheurs peuvent relier la constitution génétique à la susceptibilité d'avoir une maladie. A titre d'exemple, le projet international des *1000 Genomes* a séquencé les génomes de 2500 humains afin de cataloguer les variations génétiques entre populations (1000 Genomes Project Consortium et al. 2010, Sudmant et al. 2015).

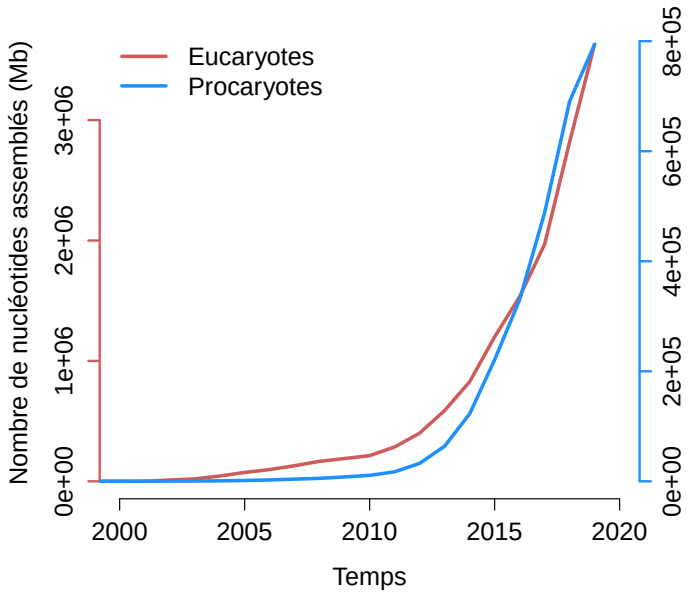


Figure I.8 – *Évolution du nombre de nucléotides assemblés et disponible au NCBI*

Les courbes rouge pour les eucaryotes et bleue pour les procaryotes montrent l'évolution exponentielle du nombre de nucléotides assemblés contenus dans les bases de données. Données issues du *NCBI* en juillet 2019

Groupe	Sous groupe	Nombre de génomes	Taille cumulée de génome (Mb)
Animaux	Mammifères	722	1 656 250
	Insectes	626	233 814
	Poissons	317	266 617
	Oiseaux	183	206 636
	Vers ronds	175	25 276
	Vers plats	50	26 519
	Reptiles	47	88 887
	Amphibiens	9	54 648
	Autres	225	167 064
Champignons	Ascomycètes	3 792	10 7087
	Basidiomycètes	801	32 712
	Autres	213	12 565
Plantes	plantes terrestres	832	726 202
	algues vertes	88	7 116
	Autres	3	2 001
Protistes	Apicomplexes	265	7 994
	Cinétoplastes	117	3 858
	Autres	367	25 443
Total		**8832**	**3 650 689**

TABLE I.2 – *Liste des génomes eucaryotes disponibles*

Nombre et tailles cumulées de génomes disponibles au NCBI en juillet 2019

I.3.2 Les technologies de séquençage

DANS cette section, je vais m'appliquer à décrire le principe des différentes techniques de séquençage par ordre d'apparition, sans vouloir être exhaustif. En effet, je ne vais me focaliser que sur les techniques auxquelles j'ai été confronté ou qui ont un fort potentiel (**Table I.3** p.25). Tout d'abord, avec le WGS, deux approches existent pour la préparation des matrices (les molécules à séquencer) : les matrices amplifiées et les matrices à molécule unique.

L'amplification PCR des matrices permet d'accéder à des molécules très peu abondantes. Cependant, l'étape de PCR n'est pas dénuée de conséquences. Au-delà des erreurs de la polymérase, l'amplification est biaisée selon la composition nucléotidique des molécules. Des séquences plus riches en G et C ont plus de chance d'être amplifiées que des séquences riches en A et T. Pourtant, les progrès bio-technologiques sur l'efficacité des polymérases tendent à réduire ces biais. Une autre conséquence de l'amplification PCR, pouvant être problématique selon les projets, est que toutes les modifications épigénétiques sur l'ADN d'origine sont perdues.

A l'inverse, les matrices à molécule unique ne sont pas amplifiées. Les molécules d'ADN contenues dans la cellule sont directement séquencées. Malgré une sensibilité plus faible, les matrices à molécule unique ne souffrent d'aucun biais et sont particulièrement désignées pour des analyses quantitatives (expression de gènes par ARN-seq (WANG ET AL. 2009)). Il faut néanmoins être conscient que la plupart des technologies de séquen-

çage nécessitent une quantité minimale de matériel à engager (de l'ordre de quelques ng pour les moins exigeantes). Sans amplification, ce pré-requis peut s'avérer problématique selon les modèles d'étude ou les échantillons.

Les séquenceurs *Sanger* à terminaison de chaine inaugurent la première génération de séquenceurs. A partir de 2005, les technologies de séquençage à haut débit font leur apparition. En s'affranchissant des étapes de construction de banques génomiques, elles génèrent des millions de séquences à moindre coût. Il existe deux stratégies principales pour les séquenceurs de deuxième génération : par addition de nucléotides uniques (pyroséquençage *454 Roche*) et par terminaison cyclique réversible (*Illumina*). Enfin, le séquençage de molécules uniques (*Pacific Biosciences* ou *Oxford NanoPore Technology*) annonce la troisième génération de séquenceurs. Commune à toutes les technologies NGS, les matrices sont spatialement distinguables et traitées indépendamment et de façon simultanée. Les matrices peuvent être fixées ou immobilisées sur un support solide, ou isolées dans des cavités.

I.3.2.1 Première génération

Le **séquençage *Sanger***, aussi connu sous le nom de séquençage par terminaison de chaine, a été développé dans les années 1970 (Sanger et al. 1977). La même année une méthode radicalement différente a également été développée par Maxam and Gilbert (1977), s'appuyant sur une dégradation chimique sélective de l'ADN. En 1980, Gilbert et Sanger ont été récompensés par le prix Nobel de chimie. Néanmoins, l'approche *Sanger* s'est rapidement popularisée et a contribué à l'émergence de la génomique.

L'ADN à séquencer est cloné dans des banques génomiques (de plasmides, fosmides, cosmides ou de BACS). Puis, à l'aide d'une amorce, l'ADN matrice est polymérisé en présence de nucléotides permettant l'élongation et une faible proportion de nucléotides imposant l'arrêt de l'ADN polymérase. Ainsi, par des arrêts aléatoires de l'ADN polymérase, des molécules de taille variable sont générées. Après migration sur un gel de polyacrylamide, la séquence est déduite à partir de la taille des molécules et l'identification du nucléotide terminateur de chaine. La méthode *Sanger* a largement évolué depuis son invention. Les marquages fluorescents ont remplacé les marquages radioactifs. Les polymérases ont été améliorées en qualité et en efficacité. Grâce à l'automatisation, en quelques heures, un biologiste peut avoir la séquence de la molécule de plusieurs centaines de bases contenue dans son tube *Eppendorf*.

I.3.2.2 Seconde génération

Le **séquençage 454** de *Roche*®est aussi appelé pyroséquençage. Personnellement, je n'ai pas été confronté directement à cette technologie mais elle a indéniablement fait entrer le séquençage dans l'ère du haut débit (Ronaghi et al. 1996; 1998).

L'ADN est fragmenté puis deux adaptateurs différents sont ligaturés à chacune des

extrémités des fragments. Ces adaptateurs permettront l'amplification PCR, nécessaire à l'amplification du signal de détection. Les matrices sont fixées sur des billes. Avec un dosage adéquat, chaque bille ne portera qu'un seul fragment. Enfin, les billes sont placées dans une émulsion pour l'étape de PCR. Chaque goutte d'émulsion ne contiendra qu'une bille. A la fin des cycles d'amplification, les billes seront recouvertes de molécules d'ADN simple brin identiques. Les billes sont déposées sur des plaques PTP (*Pico Titer Plate*) composées de millions de trous ne permettant d'accueillir qu'une seule bille. La plaque est mise en contact successivement avec des solutions ne contenant qu'un seul des 4 nucléotides. En cas de polymérisation, un signal lumineux est détecté. L'ordre et l'intensité des pics de lumière permettent de déduire la séquence d'ADN fixée sur chacune des billes. Au cours des années, les plaques PTP et les billes ont été améliorés afin d'augmenter la taille et la qualité des lectures séquencées (Margulies et al. 2005). Cependant cette technologie souffre d'un haut taux d'erreur, notamment au niveau des homopolymères (**Table I.3 p.25**). Remplacée par des séquenceurs de troisième génération, la technologie 454 a presque totalement disparu.

Le séquençage par synthèse d'*Illumina* (*Solexa*®) est de loin la technologie NGS la plus utilisée depuis une dizaine d'années. Comme le séquençage *Sanger*, *Illumina* utilise des nucléotides terminateurs de chaine. Après fragmentation de l'ADN, des adaptateurs sont ligaturés aux extrémités des molécules. Les molécules (simple brin) sont attachées de manière covalente sur une plaque de verre (*flowcell*) (Fedurco et al. 2006). Chaque molécule, répartie sur la plaque, est amplifiée localement formant ainsi des groupes (*clusters*) denses de molécules identiques attachées sur la plaque. Ces groupes de molécules permettront d'amplifier le signal. Plusieurs millions de groupes ou amplicons spatialement distinguables se forment sur la *flowcell*. Le principe de détermination de la séquence par terminaison réversible de chaine est décrit sur la **figure I.9** (p.22) (Bentley et al. 2008). Cette méthode implique notamment l'utilisation de nucléotides couplés à un fluorophore. Les quatre types de nucléotides se distinguent par des fluorophores différents. Ces nucléotides sont également modifiés pour empêcher l'incorporation du nucléotide suivant sur la chaine (terminateur). Contrairement à la technologie *Sanger*, cette propriété terminatrice est réversible par clivage. La procédure est composée de plusieurs cycles de trois étapes : (i) incorporation d'un nucléotide par complémentarité de séquence (ii) lecture d'un signal lumineux correspondant à la base incorporée (iii) clivage de la partie terminatrice et fluorescente du nucléotide incorporé, permettant l'élongation de la chaine au prochain cycle. A chaque cycle, un nucléotide est donc déterminé parallèlement pour chacun des millions de clusters positionnés sur la plaque. Comparé à la technologie *454*, le séquençage *Illumina* génère des lectures beaucoup plus courtes. Cependant le taux d'erreur est très bas et correspond le plus souvent à des substitutions. Aujourd'hui, en une semaine, une expérience peut générer jusqu'à 6 milliards de séquences, d'une longueur de 50 à 300 nucléotides (**Table I.3** p.25).

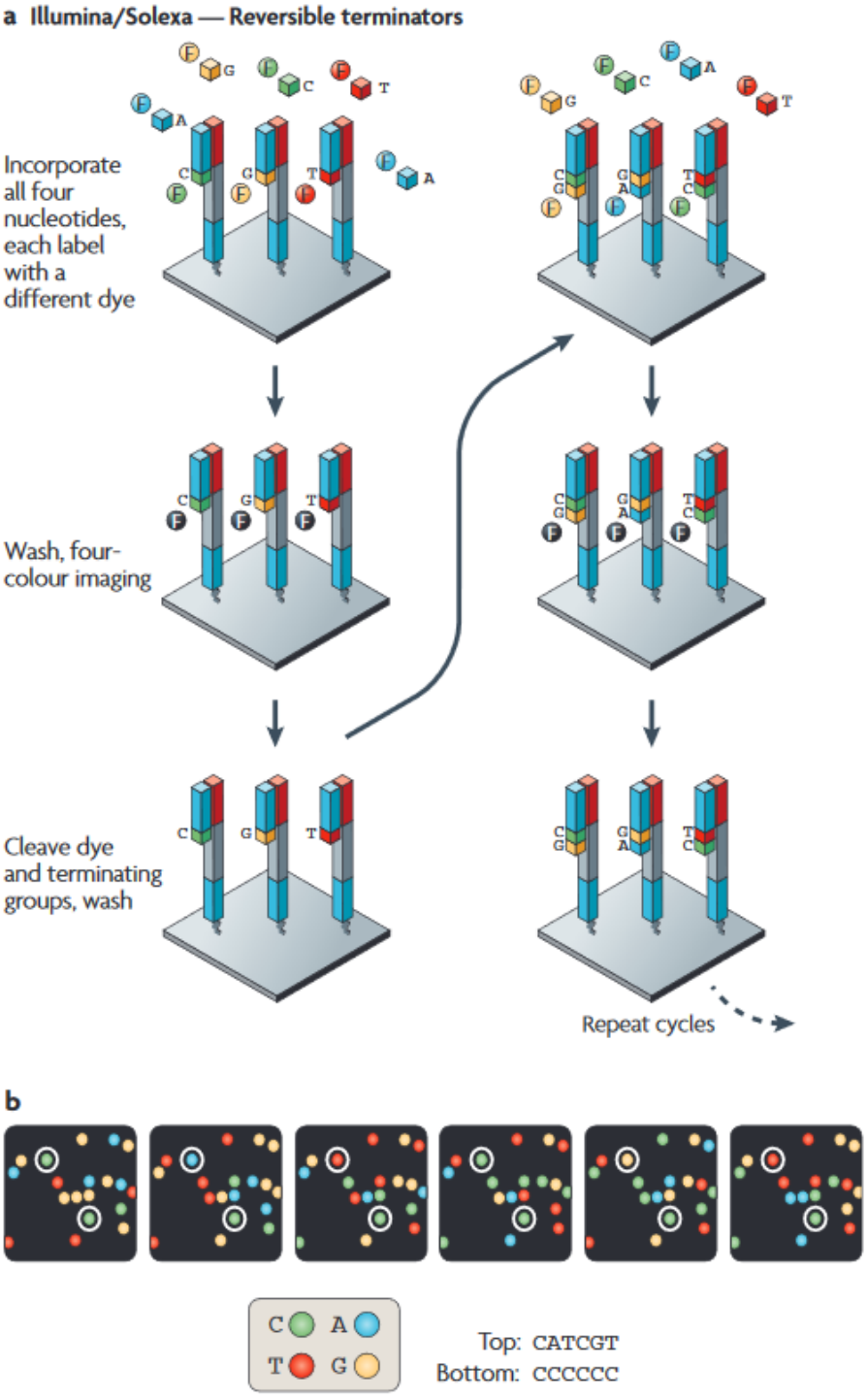


FIGURE I.9 – *Principe de terminaison réversible cyclique de chaines*

(a) Les trois étapes de la méthode de terminaison réversible cyclique utilisée par les séquenceurs *Illumina*. Sur chacun des 3 *clusters* schématisés les nucléotides modifiés sont incorporés par complémentarité de séquence. Après lavage, une photographie de la plaque permettra de déterminer le nucléotide incorporé en fonction de son fluorophore (b) sur chacun des *clusters*. Le clivage élimine les colorants fluorescents et autorise l'extension de la chaine. Figure tirée de METZKER (2010).

I.3.2.3 Troisième génération

Les technologies de seconde génération génèrent des lectures de haute qualité mais relativement courtes. Dans les génomes, l'architecture des éléments répétés peut être assez complexe et d'une taille importante. Les lectures courtes, même appariées (voir **section I.3.3** p.26), ont des difficultés à résoudre ces structures. Malgré un taux d'erreur important, les séquenceurs de troisième génération produisent des lectures longues (jusqu'à plusieurs dizaines de Kb) pouvant aider à la résolution de ces structures complexes. Par ailleurs, les lectures longues sont utiles pour l'étude des transcrits, et en particulier la définition des transcrits alternatifs d'un gène (Hardwick et al. 2019).

Le séquençage en temps réel de molécules uniques (SMRT pour *Single Molecule Real Time*) de **Pacific Bioscience®** (Eid et al. 2009) est une technologie de troisième génération générant des lectures longues (**Figure I.10** p.24 et **Table I.3** p.25). Contrairement à la technologie *Illumina*, la lecture des nucléotides incorporés se fait en temps réel et de manière continue. En effet, la polymérase ne fait pas d'arrêt et elle est fixée dans de petites chambres de détection (ZMW pour *Zero-Mode Waveguide* Levene et al. (2003)) de quelques picolitres. Dans chacun des puits, et à chaque incorporation de nucléotides fluorescents, un système de laser et de caméra mesure le temps d'émission et le type de lumière pour en déduire la séquence. Cette plateforme tire profit d'un lieu de polymérisation, et donc d'émission de lumière, fixe. Une circularisation des fragments permet de lire plusieurs fois la même molécule et ainsi générer une lecture consensus minimisant le taux d'erreur. De plus, avec un apprentissage adéquat, la technologie SMRT est capable d'identifier des nucléotides épigénétiquement modifiés, rallongeant le temps de passage de la polymérase.

Oxford®Nanopore Technologies (ONT) est également une plateforme de troisième génération. Contrairement aux technologies décrites précédemment, ONT ne lit pas l'incorporation de nucléotides par complémentarité sur une séquence matrice. ONT détecte directement la composition d'un ADN simple brin passant à travers un pore protéique ou synthétique (Clarke et al. 2009). Pour déduire la séquence, des variations de tension ionique sont mesurées et interprétées pour correspondre à un nucléotide spécifique où plus précisément à une succession de plusieurs nucléotides de longueur K (K-mer) (Jain et al. 2015). Les brins sens et antisens sont lus afin d'obtenir une séquence consensus de meilleure qualité. La technique "$1D^2$" lit un brin puis l'autre. Alors que, la technique "2D" utilise un ADN circularisé grâce à des adaptateurs en épingle à cheveux et les deux brins sont lus l'un à la suite de l'autre (**Figure I.10** p.24 et **Table I.3** p. 25). Ne nécessitant aucun appareillage optique, le séquenceur ONT peut être extrêmement portatif. Certains modèles ont une dimension comparable à une clé USB. Dans le cadre de projets itinérants comme sur le bateau TARA, ce genre de système est tout à fait approprié.

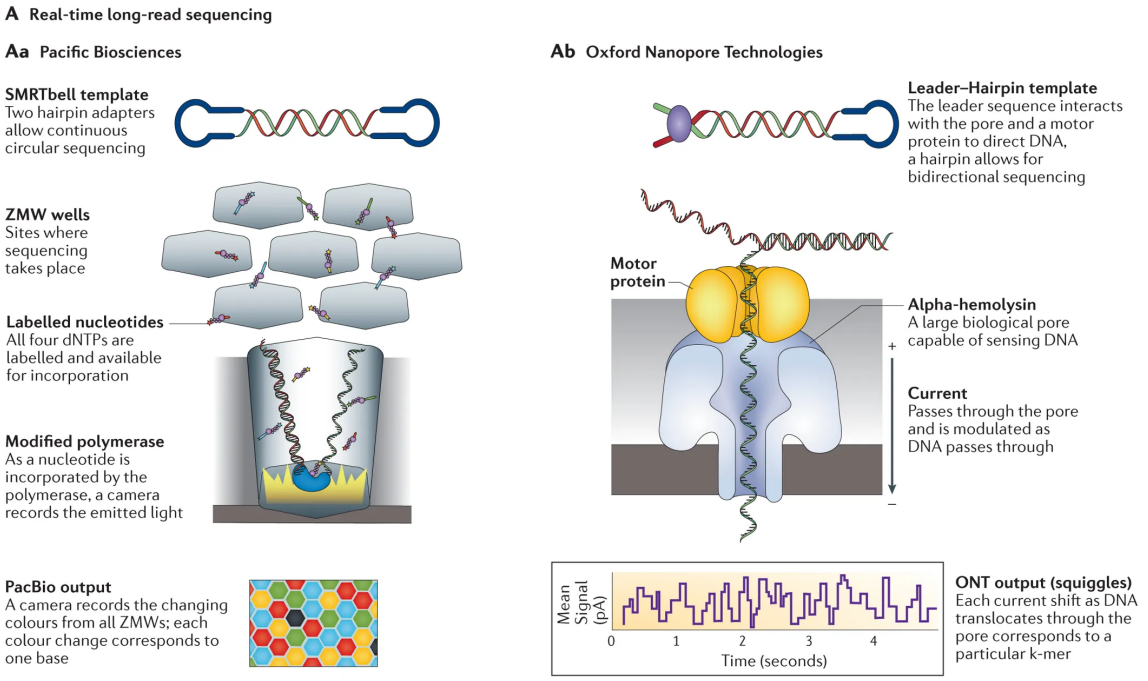


FIGURE I.10 – *Plateformes de séquençage à lecture longue en temps réel*

Aa technologie *Pacific Biosciences* (PacBio). Les fragments matrices sont ligaturés à des adaptateurs en épingle à cheveux aux deux extrémités, circularisant la molécule. Grâce à des amorces, une ADN polymérase est liée à ces molécules d'ADN circularisées puis elles sont déposées au fond de puits (ZMW). La polymérase incorpore en continu des nucléotides marqués par des fluorophores. Une caméra enregistre la couleur émise et notamment les changements de couleurs traduisant l'incorporation d'une nouvelle base différente. **Ab** *Oxford Nanopore Technologies* (ONT). Un adaptateur "guide" (*leader*) et un adaptateur en épingle à cheveux sont fixés à chacune des extrémités des fragments. L'adaptateur guide dirige l'ADN vers un pore. La translocation de l'ADN simple brin au travers de ce pore entraine un décalage de tension. L'ampleur et la durée du décalage sont enregistrées et peuvent être interprétées comme une séquence spécifique. L'adaptateur en épingle à cheveux permettra de lire le brin complémentaire afin de générer une séquence consensus de meilleure qualité ("2D"). Figure tirée de GOODWIN ET AL. (2016).

Génération	Plateforme	Amplification ou molécule unique	Méthode	Longueur des lectures	Taux d'erreur en %	Type d'erreur majoritaire	Nombre de lectures générées	Coût relatif par Mb (USD)
1	Sanger - AB1	PCR	Terminaison de chaine *via* des didésoxyribonucléotides	600-1000	0.001	InDel et substitutions	100	500
2	454 Roche - GS FLX+	PCR	pyroséquençage	700	1	InDel	1e6	∼10
2	Illumina Solexa - HiSeq	PCR	Terminaison réversible de chaine	50-300	0.1	Substitution	1e9	∼0.1
3	Pacific Biosciences - RSII	SMRT	Lecture contenue d'émission de lumière	8-20k	13% sans circularisation	InDel	1e5	∼0.5
3	Oxford Nanopore Technology (ONT) - MinION	SMRT	Interprétation de variations de tension ionique après passage au travers d'un pore	10 à 100 Kb	12	InDel et substitutions	1e4	∼10

TABLE I.3 – *Comparaison des technologies de séquençage*

Tableau adapté de DEROCLES ET AL. (2018)

I.3.3 Les stratégies d'assemblage

Le séquençage donne accès à la séquence de fragments d'ADN plus ou moins longs appartenant au génome d'intérêt. L'assemblage est la procédure bioinformatique permettant de reconstituer la séquence des chromosomes (ou ce qui s'en rapproche le plus) à partir des lectures issues d'un séquençage. La fragmentation des molécules d'ADN se faisant aléatoirement, une région de génome est représentée par plusieurs molécules différentes. Le processus d'assemblage s'assimile à la reconstruction d'un livre à partir de plusieurs exemplaires du livre déchiquetés aléatoirement en petits morceaux (Green 2001, Nagarajan and Pop 2013). L'idée de base derrière tous les assembleurs est que tous les fragments d'ADN identiques proviennent de la même région du génome. Autrement dit, tous les morceaux de "phrase" identiques proviennent de la même page du livre si on continue la métaphore du livre. On peut d'ores et déjà imaginer que ce postulat est faux car il existe des segments d'ADN parfaitement identiques (ou presque) dispersés dans un génome. De plus, avec des lectures de séquençage courtes, le risque d'avoir deux segments identiques augmente. Les assembleurs analysent les lectures chevauchantes et forment une séquence unique consensus contiguë plus grande (*contig*) (**Figure I.11** p. 27). A cette étape, les contigs ne représentent que des morceaux de chromosome. En effet, le manque de lectures ou la présence de séquences répétées empêchent les assembleurs de construire, sans ambiguïté, des contigs plus longs. Des informations de liaisons de longues distances sont utilisées pour ordonner les contigs entre eux. Des cartographies optiques, des cartes de recombinaison, des lectures pairées de longue distance ou des lectures très longues peuvent être utilisées pour joindre plusieurs contigs et former des *scaffolds*. Certains assembleurs, comme SOAPdenovo (Li et al. 2010), intègrent directement cette étape mais des logiciels, comme SSPACE (Boetzer et al. 2011), sont dédiés à cette tâche. Dans le meilleur des cas, les scaffolds sont plus ou moins équivalents aux chromosomes. Il faut garder en mémoire que plus le génome sera composé d'éléments répétés, plus la tâche de l'assembleur sera difficile.

L'évolution des technologies de séquençage et plus spécifiquement les caractéristiques des lectures obtenues (longueur des lectures, nombre de lectures et taux d'erreur) ont largement guidé l'évolution des méthodes d'assemblage. Les stratégies d'assemblage sont réparties en trois catégories. La méthode *Glouton* (*Greedy*) fait des choix avec le meilleur bénéfice immédiat sans pouvoir revenir en arrière. Cet algorithme a une vision locale de l'assemblage et ne tient pas compte des relations globales pouvant exister entre fragments. Des assembleurs comme TIGR (Sutton et al. 1995) ou VCAKE (Jeck et al. 2007) utilisent des heuristiques pour éviter le mauvais assemblage des séquences répétées. La méthode d'assemblage par chevauchement (*Overlap-Layout-Consensus OLC*) est beaucoup plus utilisée que la méthode précédente. Cette méthode identifie les lectures montrant un certain chevauchement et les organise en graphes. Les graphes intègrent les relations provenant de toutes les lectures. Après simplification des graphes, une séquence consensus est déduite. Des logiciels comme Celera (Myers et al. 2000) ou Arachne (Batzoglou

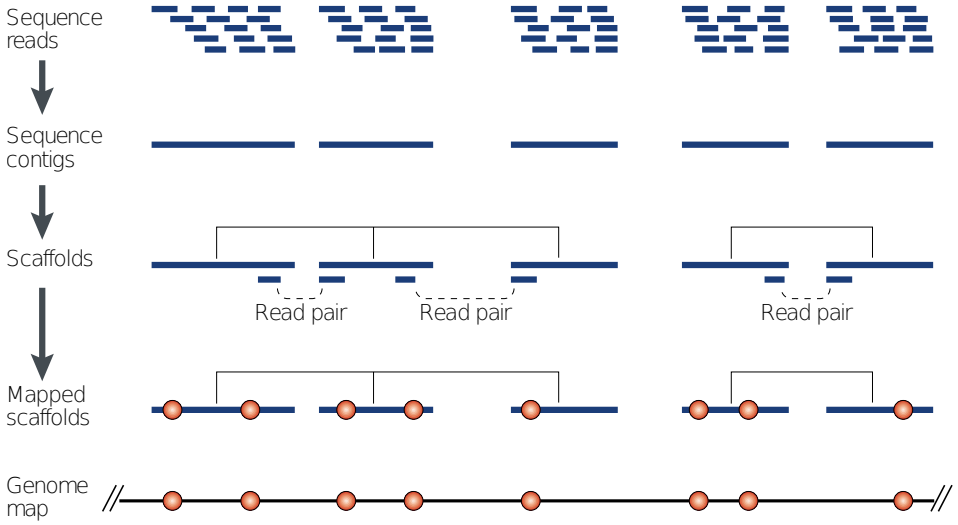


FIGURE I.11 – *Principe d'assemblage de génome*

Les lectures individuelles chevauchantes, issues du séquençage, sont assemblées en contigs. Ces contigs sont ensuite groupés en scaffolds grâce à des informations longue distance. Dans cet exemple, chaque lecture de la paire s'aligne sur un seul contig permettant l'établissement d'un lien non ambigu. La dernière étape consiste à identifier dans les scaffolds des marqueurs génétiques connus, ce qui permet d'associer les scaffolds à des positions chromosomiques. Figure reproduite de GREEN (2001).

ET AL. 2002) demandent des ressources computationnelles très importantes. L'émergence des NGS avec leurs millions de séquences a rendu ces classes d'algorithmes assez difficiles à mettre en œuvre, même si des approches optimisées comme SGA (SIMPSON AND DURBIN 2012) voient le jour. Le troisième type de méthode est basé sur les graphes de De Bruijn. Ces graphes sont construits en utilisant des sous chaines de nucléotides de longueur K (K-mer) extraites des lectures. La séquence consensus est déduite de l'interprétation du graphe. Cette méthode a des points communs avec la méthode OLC mais l'utilisation de mots de taille réduite accélère largement les calculs. Étant basé sur une identité parfaite entre les K-mer, ces méthodes sont particulièrement sensibles aux erreurs de séquençage. Des étapes de filtration et/ou correction des lectures sont parfois nécessaires afin d'obtenir un assemblage de qualité. Cette approche a été popularisée avec l'assembleur Euler (PEVZNER ET AL. 2001), puis largement utilisée avec des logiciels comme Velvet (ZERBINO AND BIRNEY 2008), SOAPdenovo (LI ET AL. 2010) et ALLPATHS (BUTLER ET AL. 2008). Des logiciels comme SPAdes (BANKEVICH ET AL. 2012), miniasm (LI 2016), Canu (KOREN ET AL. 2017) (basé sur Celera) ou plus récemment SMARTdenovo, Flye (KOLMOGOROV ET AL. 2019) et Wtdbg2 (RUAN AND LI 2020) sont accoutumés à un taux d'erreur important des lectures longues de troisième génération. Pourtant de nombreux algorithmes de correction hybride existent pour améliorer la qualité des lectures ONT ou PacBio en utilisant les lectures de 2ème génération (WANG AND AU 2020).

I.3.4 Un assemblage de qualité

La qualité d'un assemblage s'évalue selon plusieurs paramètres. Plus un assemblage va se rapprocher de la réalité du génome d'intérêt meilleur il sera. Une bonne connaissance de l'organisme et de son génome est donc inestimable. Des informations indépendantes sur le nombre de chromosomes, la taille du génome, sa ploïdie, sa composition nucléotidique moyenne (taux de G+C) permettent de juger de la qualité d'un assemblage. Avoir accès au génome d'une espèce proche est également un atout. Un logiciel comme QUAST (Gurevich et al. 2013) évalue la qualité des assemblages en donnant de nombreuses statistiques et permet de comparer les génomes.

Les *contigs* sont les séquences consensus contiguës calculées à partir des lectures séquencées. Plusieurs *contigs* ordonnés, orientés et liés entre eux par des informations de liaison longue distance forment les *scaffolds*. La taille du plus petit et du plus grand *scaffold* sont des valeurs facilement appréhendables. La somme des longueurs des *scaffolds*, aussi appelée complexité de l'assemblage, doit se rapprocher de la taille estimée du génome. Compte tenu de la difficulté d'assembler les séquences répétées, un assemblage avec une couverture de génome de 90-95% est considéré comme bon. Évidemment cette valeur dépend de la quantité, de la taille, et de la ressemblance des séquences répétées. Les séquences répétées trop ressemblantes ne sont pas différentiables par l'assembleur et sont souvent retrouvées en une seule occurrence dans l'assemblage. Cette occurrence, parfois isolée sur un *contig*, risque d'être une version consensus chimérique des répétitions. On parle de *collapse* de séquences. Souvent composés de séquences répétées, les centromères et télomères sont rarement bien assemblés.

Le N50 est une des métriques les plus utilisées pour juger de la qualité d'un assemblage. Cette mesure est définie comme la taille pour laquelle la longueur combinée de tous les *scaffolds* (ou *contigs*) plus grands que cette valeur représente au moins 50% de la complexité de l'assemblage (Narzisi and Mishra 2011). C'est une valeur représentative de la fragmentation de l'assemblage. Plus le N50 est élevé, meilleur est l'assemblage. Il faut néanmoins prendre garde à ne pas artificiellement surévaluer le N50. Par exemple, si l'étape de scaffolding est réalisée avec trop de permissivité, la moindre information de liaison longue distance va avoir tendance à lier des *contigs* artificiellement et ainsi augmenter le N50. Dans une optique d'annotation de gènes, un N50 minimum correspondant à la taille moyenne des gènes est nécessaire. Dans le cas contraire il est recommandé de générer plus de données ou des données d'un type différent. D'autre part, il est fréquent que la séquence entre deux contigs liés soit indéterminée. Des "N" représentent cette incertitude de nucléotide. Le nombre de *N* et le nombre de régions avec des *N*s contigus (*Gap*) sont des paramètres pour estimer les lacunes dans l'assemblage.

Il est rare que le premier assemblage d'un génome d'une espèce soit parfait. On parle d'un assemblage "brouillon" (ou *draft*). De nombreuses études indépendantes sont nécessaires pour corriger un assemblage afin qu'il atteigne un stade mature. Seuls quelques génomes de grands organismes modèles, sans compter celui de l'homme, peuvent se pré-

valoir d'avoir atteint le stade d'un assemblage finalisé. Plusieurs types d'erreurs existent dans un assemblage avec plus ou moins de répercussions sur les analyses futures et notamment l'annotation. Des liaisons chimériques ou manquantes entre *contigs* vont entrainer une incohérence entre le nombre de chromosomes et le nombre de *scaffolds*. Des erreurs plus locales, comme de petites inversions, des substitutions, des manques de nucléotides ou des nucléotides surnuméraires sont très problématiques pour l'annotation. En effet, une substitution peut révéler ou effacer un codon terminateur ou initiateur de traduction des gènes codants. Une insertion ou une délétion (*InDel*) de nucléotides décalent les phases ouvertes de lecture ne pouvant conduire qu'à une annotation erronée. Nous verrons dans la **section IV** et **V.1.1** (p.89 et p.105) des Résultats qu'une amélioration de l'assemblage de *P. tetraurelia* a permis une meilleure annotation des gènes et de séquences liées à des transposons (IES, voir **section III.3.2.1** p.68). La technologie de séquençage employée et la couverture du génome en lectures influencent le type et le nombre d'erreurs. Par exemple, des lectures courtes de type *Illumina* vont avoir tendance à fragmenter l'assemblage. En revanche, des lectures longues provenant de technologies de troisième génération seront plus promptes à faire des erreurs de type *InDel*.

En 2007, l'arrivée des NGS et ses lectures courtes, a entrainé une augmentation du nombre de génomes disponibles (voir **section I.3** p.16) (Goodwin et al. 2016). Cependant, nul ne peut nier que les meilleurs assemblages (aussi les plus coûteux) ont été réalisés avec des lectures *Sanger*. La troisième génération de séquenceur permet d'obtenir des tailles de lectures très importantes (van Dijk et al. 2018) et donc des assemblages intéressants. Cependant le taux d'erreur reste un problème important (Henson et al. 2012, Sohn and Nam 2018). Il est devenu courant d'allier les deux types de technologies pour profiter des avantages de l'un et de l'autre. Par exemple, avec des logiciels comme Pilon (Walker et al. 2014), des lectures de séquençage *Illumina* sont utilisées pour corriger l'assemblage réalisé à partir de lectures longues. Nous avons également vu dans le paragraphe I.3.3 (p.26) précédent qu'il était possible de corriger les lectures ONT ou PacBio pour obtenir un meilleur assemblage (Wang and Au 2020). Il est possible que ces étapes ne soient bientôt plus nécessaires avec la constante amélioration de la qualité de séquençage des lectures longues.

Chapitre II

Annotation des génomes

Dans le chapitre précédent, j'ai introduit la notion de génome et présenté les moyens d'accéder à sa séquence. En étant provocateur, je dirais qu'aujourd'hui, séquencer et assembler un génome devient presque routinier. Cependant, avoir la séquence génomique de son organisme préféré n'a que peu d'intérêt si l'on ne la relie pas aux fonctions biologiques en décryptant l'information codée par celle-ci. Autrement dit l'annoter. Dans ce chapitre, je vais me focaliser sur la description des méthodes d'annotation des gènes et des éléments transposables.

Avant tout, une bonne annotation ne peut être espérée qu'avec un génome de qualité. Il faut savoir qu'une amélioration significative de l'assemblage bénéficiera à la qualité de l'annotation (Elsik et al. 2014). Certaines espèces jouissent d'un génome de qualité et d'une très bonne annotation car ils profitent de plusieurs décennies d'étude dédiée.

Dans le chapitre précédent, nous avons vu que les technologies NGS ont fait chuter le coût de séquençage, provoquant la parution de nombreux génomes (voir **section I.3** p.16). Le nombre important de génomes séquencés et annotés apporte son lot d'avantages et d'inconvénients. Personne ne peut nier que les meilleurs assemblages, comme celui de la drosophile (Adams et al. 2000, Celniker et al. 2002) ou de l'homme (Venter et al. 2001, International Human Genome Sequencing Consortium 2004), ont été réalisés à partir de lectures *Sanger*. Des stratégies d'annotation utilisent les connaissances acquises sur des espèces proches. Donc avoir accès aux annotations d'une variété large d'espèces facilite l'annotation de génomes taxonomiquement isolés. En revanche, face à ce déluge de génomes, la plupart des annotations sont réalisées de façon complètement automatique et sans curation humaine. Des erreurs inévitables vont polluer les bases de données et malheureusement ces erreurs contamineront de futures annotations (Salzberg 2019).

Dans le cadre d'un projet d'annotation de génome, les scientifiques se fixent des objectifs en fonction des données et des moyens disponibles. La taille de la communauté et les moyens financiers/humains mis en jeu sont prépondérants dans ce choix, et souvent liés à l'histoire et à l'impact sociétal à plus ou moins court terme. Malgré un enjeu économique majeur, l'annotation du génome de certaines espèces peut se révéler un travail de titan. Je pense notamment aux plantes et plus spécifiquement au blé. L'annotation des

107 000 gènes du génome hexaploïde de *Triticum aestivum* (blé tendre), de 15Gb et contenant 85% de séquences répétées, a pris près de 13 ans (International Wheat Genome Sequencing Consortium (IWGSC) et al. 2018).

Nous verrons dans ce chapitre que l'annotation se décline en deux catégories : l'annotation structurale définit la position et structure des éléments sur le génome et l'annotation fonctionnelle prédit, ou mieux définit, la fonction des éléments. Même s'il est souvent requis de commencer par masquer les séquences répétées, l'annotation structurale des gènes codants est classiquement le premier objectif. Les gènes codants sont, en principe, les plus reconnaissables car des contraintes d'évolution s'appliquent sur leurs séquences. L'identification des introns est un point critique de l'annotation des gènes codants. Le séquençage de l'ARN messager est couramment utilisé pour caractériser les introns. Le ARN-seq (séquençage d'ARN par des NGS) est capable de mettre en évidence des gènes très peu exprimés, indétectables par des EST (*Expressed Sequence Tag*). L'annotation structurale se poursuivra par les gènes non codants, les pseudogènes, les éléments transposables et autres séquences répétées. Encore plus ambitieux, l'annotation des séquences régulatrices ou la description de l'état épigénétique des séquences requièrent des données spécifiques (projets ENCODE (Consortium 2004) et modENCODE (Celniker et al. 2002)). L'annotation fonctionnelle s'appuie sur les éléments caractérisés par l'annotation structurale. Donc la qualité de l'annotation structurale impactera la qualité de l'annotation fonctionnelle. Plus généralement, il faut garder en tête que les analyses futures se baseront sur ces annotations. Il est donc critique d'avoir la meilleure annotation possible, tant au niveau qualité qu'au niveau de l'exhaustivité. Tout le jeu de l'annotation sera de trouver un juste milieu entre sensibilité (éviter les faux négatifs) et spécificité (éviter les faux positifs) de l'annotation.

II.1 Annotation de gènes

Plusieurs méthodes existent pour réaliser une annotation structurale de gènes (Brent 2005; 2007; 2008, Mudge and Harrow 2016). Les méthodes intrinsèques ou *ab initio* utilisent seulement la séquence génomique et des algorithmes mathématiques pour prédire la présence et la structure d'un gène. Les méthodes extrinsèques utilisent des évidences déduites de données indépendantes de la séquence génomique. Ces données proviennent, par exemple, de bases de données ou de résultats de séquençage. Ces informations sont transférées sur le génome d'intérêt pour pronostiquer la présence d'un gène et sa structure. Enfin, les méthodes combinatoires intègrent les prédictions des méthodes intrinsèques et des méthodes extrinsèques afin de donner la meilleure annotation possible en accord avec l'ensemble des informations disponibles. Ces dernières profitent des avantages des deux précédentes méthodes.

II.1.1 Masquer les répétitions

Nous avons vu dans le chapitre précédent que les séquences répétées sont problématiques pour le processus d'assemblage. Les répétitions perturbent également les logiciels automatiques d'annotation. C'est pourquoi, la première étape pour l'annotation génique d'un génome est souvent le masquage de ses répétitions. Dispersées dans le génome, les répétitions sont des séquences de faible complexité et/ou répétées en tandem (microsatellites ou minisatellites) ou des éléments transposables (voir **sections I.2.2.1** p.10 et **II.2** p.42). Les génomes eucaryotes peuvent être très riches en séquences répétées. Par exemple le génome humain est composé d'au moins 47% de répétitions (Lander et al. 2001, Jurka et al. 2007) et grimpe jusqu'à 83% pour le petit pois (Kreplak et al. 2019). On comprend intuitivement que les répétitions perturbent, ou tout du moins compliquent, le processus d'annotation des gènes, d'où la nécessité de les masquer. Cette étape est particulièrement critique et demande un ajustement adéquat. En effet certains gènes codants pour des protéines sont composés de répétitions (tétratricopeptide ou WD40). A *contrario*, certains éléments transposables ont des phases ouvertes de lecture et peuvent être interprétés par les prédicteurs de gènes. Le masquage doit donc être réalisé avec la plus grande attention. L'annotation de répétitions simples est relativement efficace avec des outils comme TRF (Benson 1999) ou TAREAN (Novák et al. 2017). En revanche, l'annotation précise des séquences répétées complexes nécessite des outils spécifiques et doit être opérée indépendamment de l'annotation des autres éléments qui compose le génome. Il existe deux grands types d'outils bioinformatiques pour les caractériser (Bergman and Quesneville 2007, Treangen and Salzberg 2011). Les outils *de novo* (Bao and Eddy 2002, Price et al. 2005) utilisant uniquement la séquence génomique (McClure et al. 2005, Buisine et al. 2008, Han and Wessler 2010) et les outils basés sur l'homologie utilisant une banque d'éléments déjà annotés (RepBase Bao et al. (2015)). L'approche par homologie est probablement la plus utilisée et notamment via l'outil RepeatMasker

(Smit 1996) utilisant un moteur BLAST (Altschul et al. 1997). Néanmoins, un problème se pose si les éléments contenus dans la librairie de séquences sont trop divergents des éléments présents dans le génome de l'espèce étudiée. Dans ce cas, nous verrons dans la **section II.2** (p.42) que la création d'une nouvelle banque d'éléments s'impose.

II.1.2 Les méthodes intrinsèques

Les méthodes intrinsèques (ou *ab initio*) utilisent des procédures mathématiques sur la séquence génomique pour détecter les éléments à annoter. Autrement dit, elles recherchent dans le texte génomique brut des signaux ou motifs liés à la présence d'éléments structurant le gène : les exons et introns. L'idée sous-jacente est de repérer les phases ouvertes de lecture les plus longues et donc les plus susceptibles de correspondre à des exons codants. Ces méthodes ont l'avantage (et l'inconvénient) d'être uniquement basées sur la séquence génomique. Elles étaient très populaires quand les évidences extrinsèques étaient rares et surtout coûteuses à produire. Particulièrement efficace chez les bactéries (avec l'exemple de Glimmer (Majoros et al. 2004)), elles sont plus complexes pour les eucaryotes. En effet, la présence d'introns complique la prédiction. Plus longs et nombreux seront les introns d'un gène, plus difficile sera la tâche du prédicteur de gènes. Les méthodes intrinsèques utilisent en général des modèles probabilistes, tels que les modèles de Markov cachés (HMM *Hidden Markov Model*). Il est souvent nécessaire de fournir un jeu avéré de gènes afin d'entrainer le programme à reconnaitre les caractéristiques des gènes de l'espèce d'intérêt. En effet, la fréquence des codons, la taille des introns et exons, la composition des séquences géniques et intergéniques, les motifs particuliers de séquences régulatrices sont autant de caractéristiques propres à chaque organisme. Les gènes codants sont détectables grâce à des signaux forts tels que les codons initiateurs de traduction, les codons terminateurs et les sites donneurs/accepteurs pour l'épissage des introns. En effet, les introns commencent généralement par 5'GT et finissent par AG 3'. On peut également rechercher des variations dans la composition nucléotidique. Les exons codants auront tendance à être plus riches en nucléotides G ou C que les introns ou les régions intergéniques. La combinaison de ces informations permet de prédire la structure des gènes. Des programmes comme GeneID (Parra et al. 2000), Fgenesh (Solovyev et al. 1995), GenScan (Burge and Karlin 1997) ou SNAP (Korf 2004) obtiennent de bons résultats. Pour l'apprentissage, avoir un jeu de gènes conséquent et représentatif est crucial, en particulier si l'espèce étudiée est phylogénétiquement distante des espèces disponibles. Malgré le perfectionnement de ces méthodes, de nombreuses erreurs de prédictions persistent. Il est souvent nécessaire d'avoir recours à d'autres types de méthodes.

II.1.3 Les méthodes extrinsèques

Beaucoup de logiciels d'annotation de génome *ab initio*, comme Augustus (Stanke and Waack 2003, Stanke et al. 2006), Gaze (Howe et al. 2002) ou EuGene (Foissac

and Schiex 2005) peuvent utiliser des évidences externes pour affiner leurs prédictions. Les évidences externes sont des indices de la présence d'un gène à un *locus* donné sur le génome cible indépendamment de la séquence intrinsèque. Les évidences externes sont décomposées en deux catégories : les évidences liées à l'homologie de séquence et les évidences liées à la transcription des gènes.

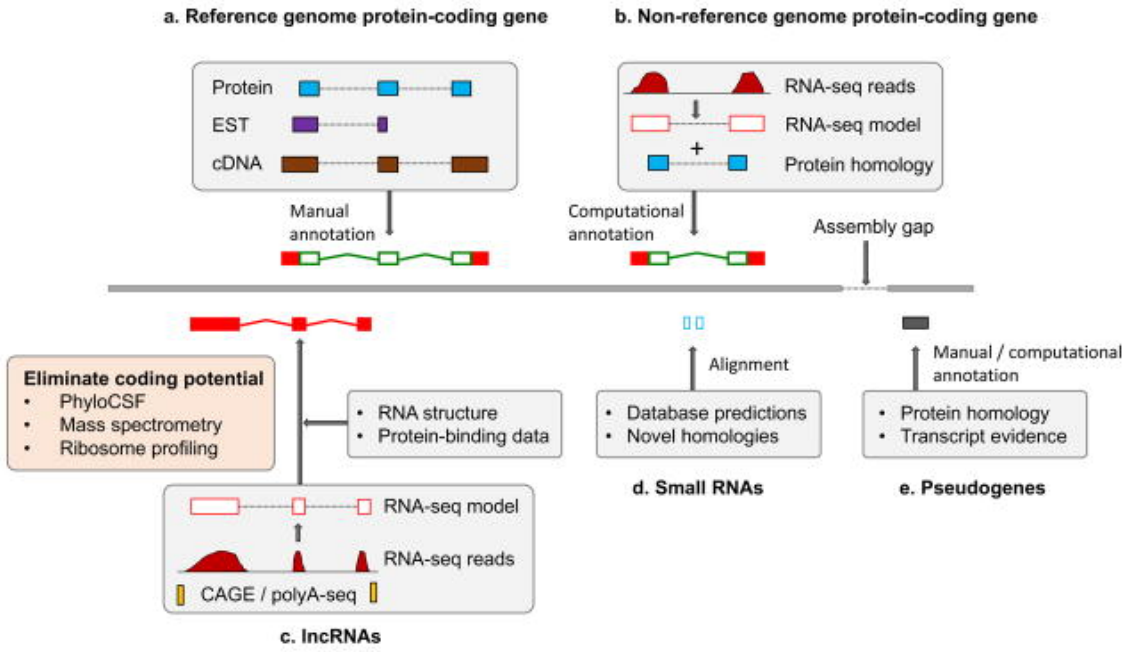


Figure II.1 – *Principes d'annotation pour les différents types de gènes*

La figure schématise les approches d'annotation possibles en fonction du type de gène. **a.** Les gènes codant pour des protéines sont annotés sur la base d'alignements génomiques de séquences protéiques ou de séquences issues de séquencage d'EST ou de cDNA. Avec des outils comme WebApollo (Dunn et al. 2019), une curation manuelle du modèle de gène est possible (voir **section II.1.5** p.40). **b.** Les gènes codants sont également annotés à l'aide de données ARN-seq (voir **section II.1.3.2** p.36) et d'informations d'homologie protéique. **c.** Comme les gènes codants, l'annotation de longs ARN non-codants utilise aussi des données ARN-seq (polyA) et CAGE-seq (voir **section II.1.3.2** p.36). De plus, des méthodes de génomique comparative (PhyloCSF Lin et al. (2011)), de spectrométrie de masse, de profilage des ribosomes, d'interaction protéique ou de prédiction de structure d'ARN apportent d'autres types d'évidences. **d.** Les petits ARN sont prédits à l'aide de bases de données comme Rfam (Nawrocki et al. 2015) ou miRBase (Kozomara and Griffiths-Jones 2014). De nouveaux *loci* sont révélés grâce au séquençage de petits ARN (Bryant et al. 2019). **e.** L'annotation de pseudogènes est basée essentiellement sur une analyse d'homologie avec des paralogues ou orthologues. Une annotation manuelle est le plus souvent requise. Figure tirée de Mudge and Harrow (2016)

II.1.3.1 Évidences par homologie de séquence

Les évidences par homologie sont basées sur l'hypothèse qu'il existe, entre les espèces, une certaine conservation des éléments à annoter. L'idée est de se servir de la connaissance déjà acquise. Cette approche reste assez limitée, si le génome d'intérêt est éloigné

évolutivement des génomes déjà annotés. De ce fait, plus les organismes annotés, contenus dans les bases de données, seront variés et représentatifs de la diversité du vivant, plus l'annotation d'une nouvelle espèce sera facilitée. Il est donc important de ne pas négliger l'annotation de génomes exotiques au profit des grands organismes modèles ou d'intérêts sociétaux. Nous avons vu dans la **section I.3** (p.16) que le nombre de génomes séquencés et annotés ne cesse d'augmenter. Cette flambée est liée à la facilité à obtenir la séquence complète d'un génome. Aujourd'hui, la base de données Uniprot-TrEMBL compte plus de 170 millions de séquences protéiques (Apweiler et al. 2004). Cependant, dans la plupart des cas, ces séquences proviennent d'une annotation automatique. Des erreurs, inhérentes à ce genre d'approche, risquent de polluer les bases de données (Salzberg 2019). En revanche, une banque de données comme Uniprot-SwissProt est un exemple d'une source de données fiable (Boutet et al. 2016). En effet, SwissProt regroupe environ 500 000 séquences protéiques manuellement certifiées par des curateurs. Seule une étude fonctionnelle des protéines autorise l'entrée dans SwissProt, conduisant à ce faible nombre de séquences accessibles. De plus, la proportion de séquences de mammifères est largement sur-représentée dans UniProt-SwissProt par rapport à l'ensemble des séquences contenues dans UniProt (**figure II.2** p.37), ce biais pouvant s'expliquer par la proportion d'équipes scientifiques travaillant sur ces modèles biologiques. Malgré ses limites, SwissProt reste une ressource inestimable pour les annotateurs.

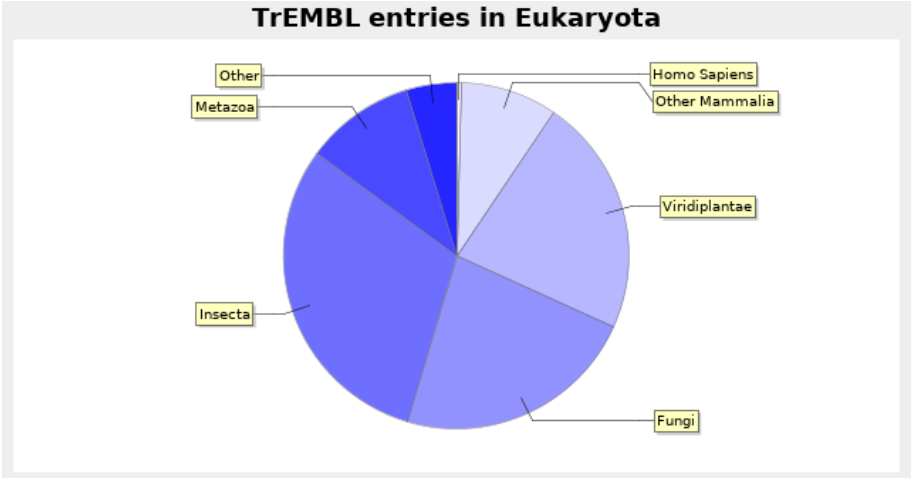
Classiquement, les séquences protéiques sont alignées sur le génome d'intérêt (**figure II.1** p.35), en utilisant les six phases de lecture de traduction, avec des logiciels comme BLASTx (Altschul et al. 1997), GeneWise (Birney et al. 2004) ou Exonerate (Slater and Birney 2005). Le bioinformaticien dégage l'information de ces alignements, et décrit les évidences pertinentes sous un format compréhensible par le prédicteur de gènes. Les gènes non-codants (tRNA, snoRNA, miRNA) sont détectés par des logiciels adaptés comme tRNAscan-SE (Lowe and Eddy 1997) ou Infernal (Nawrocki and Eddy 2013). Ces programmes interrogent des banques de données dédiées comme Rfam (Nawrocki et al. 2015) ou miRBase (Kozomara and Griffiths-Jones 2014) (**figure II.1** p.35).

II.1.3.2 Évidences d'expression

Dans cette partie, je ne vais parler que de l'expression des gènes mesurée par le niveau de leurs transcrits (ARN). Pourtant, des approches très puissantes par spectrométrie de masse révèlent l'existence de protéines et donc de gènes transcrits (**figure II.1** p.35).

Dans une condition donnée, les transcrits sont extraits, purifiés et séquencés. Selon la population d'ARN visée, l'étape de purification est essentielle. Il est important de savoir que 80% des ARN cellulaires sont des ARN ribosomiques synthétisés par l'ARN polymérase I (Russell and Zomerdijk 2006). Si l'étude porte sur les ARN messagers, une purification des molécules polyadénylées (polyA-seq Derti et al. (2012)), est particulièrement cruciale pour éviter de séquencer un grand nombre de lectures non désirées. Si le génome à annoter est éloigné du point de vue évolutif des génomes bien représentés dans

A



B

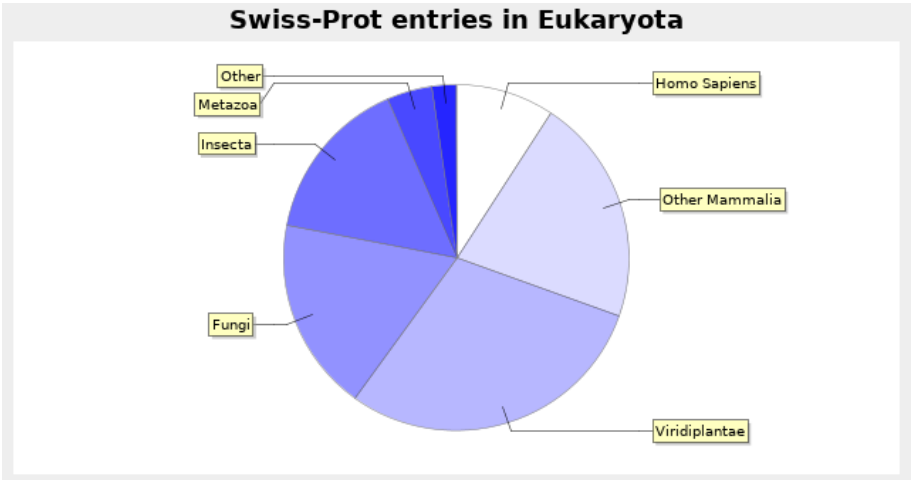


FIGURE II.2 – *Répartition des séquences dans Uniprot par groupe taxonomique*

A Proportion de séquences par groupe taxonomique dans UniProt-TrEMBL **B** Proportion de séquences par groupe taxonomique dans UniProt-SwissProt. Les métazoaires sont divisés entre trois catégories : l'homme (Homo sapiens), les mammifères hormis l'homme ("Other Mammalia") et le reste des métazoaires ("Metazoa"). Figures tirées de `https://www.uniprot.org/statistics/`

les bases de données, la production de données ARNm-seq est probablement inévitable. Générer ces données d'expression a un coût. Il y a quelques années, des lectures *Sanger* d'EST (*Expressed Sequence Tag*) étaient exploitées. La longueur des séquences compensait leur nombre relativement faible. Aujourd'hui, le ARN-seq *Illumina* a largement supplanté les EST. Un séquençage de lectures de 50 à 100 paires de bases pairées, avec conservation de l'information du brin transcrit, est certainement le plus pratiqué en ce moment. Les techniques de seconde génération sont plus sensibles et donnent accès aux gènes peu transcrits, et ainsi permettent d'avoir une vision plus exhaustive du transcriptome (Conesa et al. 2016). Cependant, les lectures courtes, comparées aux lectures *Sanger*, discriminent moins bien les formes alternatives de transcription des gènes. En effet, là ou une EST était capable de couvrir plusieurs exons, les lectures *Illumina* en sont incapables. Les technologies de troisième génération, combinant longueur et débit, devraient régler ces problèmes.

Les lectures ARN-seq sont alignées sur le génome en utilisant des logiciels comme TopHat2 (Kim et al. 2013) (ou son héritier HiSat (Kim et al. 2015)), autorisant des sauts (*gap*) dans les alignements dus aux introns. Une analyse des alignements permet de déduire la structure des unités de transcription (voir **figure II.1** p.35) par assemblage guidé avec Cufflinks (Trapnell et al. 2010) ou par analyse de couverture en lecture de séquençage avec GMORSE (Denoeud et al. 2008) ou TrUC (Arnaiz et al. 2017). Le CAGE-seq (*Cap Analysis of Gene Expression*) , ou Cap-seq, séquence préférentiellement les extrémités 5' des ARN protégées par une coiffe (Takahashi et al. 2012). Cette technique permet de mettre en évidence les sites d'initiation de la transcription (TSS pour *Transcription Start Site*) (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. 2014). Le séquençage des petits ARN et leurs alignements sur le génome aident à détecter les sites de production tels que les clusters de piRNA (Brennecke et al. 2007).

Si le génome de référence n'est pas disponible, il est toujours possible d'assembler les données ARN-seq avec Trinity (Grabherr et al. 2011) ou Velvet-Oases (Schulz et al. 2012). Les cadres ouverts de lecture sont recherchés sur ces séquences dépourvues d'introns. L'assemblage de transcrits est une approche intéressante, même si le génome est disponible. Avoir accès à ces transcrits peut compenser les lacunes ou erreurs de l'assemblage (Orgeur et al. 2018). A l'instar des séquences répétées pour l'assemblage de génomes, l'épissage alternatif complique grandement les procédures d'assemblage de transcrits.

Malgré une profondeur de séquençage importante, seuls les gènes exprimés dans la condition d'extraction des ARN seront détectables. Afin d'avoir un transcriptome plus exhaustif, il est requis de multiplier les extractions d'ARN dans différentes situations cellulaires et environnementales ou différents tissus. Au delà d'une évidence de transcription, les données ARN-seq sont utilisées pour quantifier le niveau des transcrits, approximant le niveau d'expression d'un gène dans un échantillon. En effet le niveau d'ARNm est déterminé par la synthèse et la dégradation de la molécule. Après alignement des lectures ARN-seq sur le génome de référence, des programmes, comme DESeq2 (Love et al. 2014) ou edgeR (Robinson et al. 2010), comparent le niveau d'expression des gènes entre deux

conditions. Ces analyses d'expression différentielle de gènes nécessitent des réplicats biologiques pour déterminer le niveau de bruit dans les mesures. Par une comparaison du nombre de K-mers observés dans des données de séquençage, et donc sans avoir recours à la séquence génomique, des outils comme DE-kupl (Audoux et al. 2017) sont également capables de réaliser des analyses d'expression différentielle (Pinskaya et al. 2019). Les données ARN-seq sont une source d'information très importante et contrairement aux puces à ADN, les données ARN-seq ne dépendent pas d'une annotation. En effet, si le génome et/ou l'annotation changent, les lectures pourront être réalignées et ré-analysées.

II.1.4 Méthodes intégratives

Que les méthodes soient intrinsèques ou extrinsèques, l'objectif est d'obtenir la meilleure annotation possible. Les méthodes dites "intégratives" combinent des résultats hétérogènes, générés par des méthodes intrinsèques ou extrinsèques, et les synthétisent pour fournir la version de l'annotation la plus consensuelle possible. Ces programmes sont appelés *combiner* (ou *chooser*). Des programmes comme JIGSAW (Allen and Salzberg 2005), EvidenceModeler (Haas et al. 2008) ou Evigan (Liu et al. 2008) choisissent la meilleure combinaison d'exons et proposent une prédiction. Des logiciels comme Gaze (Howe et al. 2002), Augustus (Stanke et al. 2006) ou EuGene (Foissac and Schiex 2005) fonctionnent également comme des *combiners*. Ils intègrent différents types d'évidences (intrinsèques fournies par des prédicteurs *ab initio* ou extrinsèques) pour guider et affiner leurs propres prédictions.

Si elle n'est pas automatique (comme pour Evigan (Liu et al. 2008)), la configuration de ces programmes est cruciale. Il faut renseigner les paramètres en fonction des caractéristiques des gènes à annoter et ajuster le poids de chaque type d'évidence. Autrement dit, l'impact qu'a l'évidence sur le choix fait par le *combiner*. Pour régler ces paramètres, un jeu de gènes avérés (n'ayant pas été utilisés pour l'entrainement des prédicteurs *ab initio*), est employé pour calculer des métriques de qualité de l'annotation produite. Même si plusieurs méthodes de calcul sont retrouvées dans la littérature, la sensibilité et la spécificité sont les deux mesures les plus populaires (**Figure II.3** p.40). La sensibilité est la fraction du jeu de gènes avérés correctement trouvée par la prédiction, alors que la spécificité est la fraction de l'annotation correctement prédite. La sensibilité, mesurant le taux de faux négatifs, et la spécificité, le taux de faux positifs, sont combinés en une métrique appelée "précision" (Burset and Guigó 1996). Les trois métriques peuvent être calculées pour n'importe quelle partie structurale du modèle de gène telles que la position de début et fin des transcrits, des exons ou des introns.

Annoter un génome est chronophage. Le bioinformaticien génère plusieurs jeux de prédictions *ab initio* et plusieurs jeux d'évidences externes d'homologie ou d'expression. Il teste différentes combinaisons de logiciels et passe un temps certain à ajuster les paramètres. A mon sens, cette démarche est certainement la plus qualitative mais représente un temps considérable. Quand est ce qu'une annotation est finie ? Probablement jamais.

$$SN = \frac{VP}{VP + FN} \quad SP = \frac{VP}{VP + FP} \quad AC = \frac{SN + SP}{2}$$

FIGURE II.3 – *Métriques de qualité pour l'annotation de génome*

La sensibilité (SN), la spécificité (SP) et la précision (AC pour *accuracy*) sont trois mesures couramment utilisées pour évaluer la performance du prédicteur de gènes et la qualité d'une annotation. La sensibilité est la fraction de l'annotation de référence correctement trouvée par la prédiction. La spécificité est la fraction de l'annotation correctement prédite. La sensibilité et la spécificité sont combinées en une seule mesure appelée précision (BURSET AND GUIGÓ 1996). VP pour Vrai Positif, FN Faux Négatif et FP Faux Positif. Informations tirées de YANDELL AND ENCE (2012)

Seul quelques organismes ont la "prétention" d'avoir une annotation de génome complète ou tout au moins stable.

Des procédures intégrées comme MAKER2 (HOLT AND YANDELL 2011) ou PASA (HAAS ET AL. 2011) sont fournies pour gagner du temps. De grands portails comme NCBI (THIBAUD-NISSEN ET AL. 2013) ou EnsEMBL (CURWEN ET AL. 2004) proposent des chaines de procédures complètes d'annotation.

II.1.5 Curation humaine

Les séquences génomiques et protéiques sont classiquement disponibles sous le format FASTA. Les annotations de gènes nécessitent un format plus élaboré. Ils doivent notamment renseigner les positions génomiques des exons et introns ainsi que leurs affiliations à un certain transcrit. Les formats GFF3 (et sa variante le format GTF), BED, GenBank et EMBL sont des fichiers textes largement répandus. L'utilisation de vocabulaires contrôlés tels que SO (*Sequence Ontology*) ou GO (*Gene Ontology*), rendent les annotations interopérables entre outils et donc compatibles avec bon nombre de logiciels. Les interfaces IGV (*Integrated Genome Viewer*) (ROBINSON ET AL. 2011), JBrowse (WESTESSON ET AL. 2013, BUELS ET AL. 2016) (héritier de GBrowse (STEIN ET AL. 2002, STEIN 2013)) ou UCSC (CASPER ET AL. 2018) permettent de visualiser des annotations mais également des résultats d'alignements de lectures NGS (formats SAM/BAM exploitables avec samtools (LI ET AL. 2009) ou Wig/BigWig). Le projet GMOD (*Generic Model Organism Database*) propose une série d'outils pour stocker, visualiser, analyser et distribuer des résultats génomiques. Au-delà des bien connus GBrowse et JBrowse, le projet GMOD distribue le schéma relationnel libre de base de données *chado* (MUNGALL ET AL. 2007) pouvant stocker des annotations de gènes, mais également capable d'intégrer une variété de données biologiques (phénotypes, génotypes, souches, phylogénie, publications, etc ...). Adepte fervent des outils GMOD, je les ai utilisés pour développer les bases de données ParameciumDB (ARNAIZ ET AL. 2007, ARNAIZ AND SPERLING 2011, ARNAIZ ET AL. 2019) et Cildb (ARNAIZ ET AL. 2009; 2014).

L'annotation automatique n'étant pas parfaite, des logiciels comme WebApollo (DUNN ET AL. 2019) (voir **Figure II.4** p.41) ou Artemis (RUTHERFORD ET AL. 2000), laissent les bio-

logistes ou curateurs corriger les modèles de gènes erronés. Pour de petits génomes avec "peu" de gènes, cette tâche est chronophage mais envisageable. Il est, cependant inconcevable, pour une personne ou un laboratoire, d'imaginer ce genre de procédure pour corriger de grands génomes avec beaucoup de gènes. Pourtant de grandes campagnes de séances de travail communautaire sont organisées (Munoz-Torres et al. 2011, Wang et al. 2012). Pendant plusieurs jours consécutifs, un travail conjoint et intensif de chercheurs ou de curateurs, s'effectue sur un site dédié (ou à distance avec WebApollo) pour corriger un maximum de modèles de gène. Au delà de cet objectif de curation, ces rendez-vous ont également un rôle pédagogique pour les étudiants.

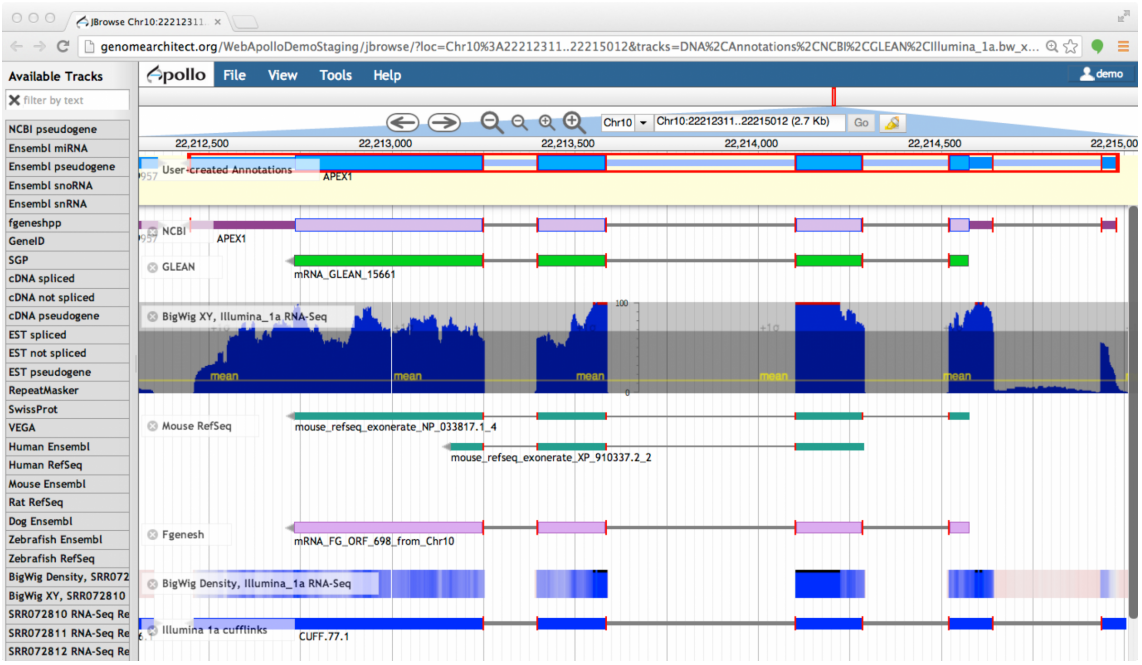


Figure II.4 – *Interface WebApollo de ré-annotation de modèles de gène*

Capture d'écran de l'interface WebApollo `http://genomearchitect.github.io/users-guide/`

II.2 Annotation des éléments transposables

II.2.1 Découverte et classification

Barbara McClintock décrit pour la première fois l'existence de gènes mobiles et le principe de transposition (McClintock 1950). Les éléments transposables (ET) (ou transposons) peuvent être considérés comme des séquences parasitaires égoïstes. En augmentant leurs nombres de copies, les transposons envahissent littéralement les génomes. Par exemple, les ET occupent 25% du génome du riz (Yu et al. 2002), 45% du génome humain (Lander et al. 2001) ou 85% du génome du maïs (Schnable et al. 2009). Une intégration d'un ET dans un gène essentiel est délétère pour l'organisme hôte. C'est pourquoi, l'hôte a élaboré des systèmes pour contrôler les ET en passant notamment par une extinction génétique via un marquage épigénétique et une hétérochromatisation (voir **section I.2.1** p.8). Les ET sont très étudiés pour leurs rôles dans l'évolution des espèces, leurs implications dans le façonnement de la structure des génomes, les réarrangements de génomes ou leurs contributions aux fonctions géniques de l'hôte. Nous avons déjà évoqué que la nature répétée des ET complique à la fois les assemblages de génomes, et l'annotation des gènes. Contrairement aux gènes codants, les copies d'ET ne sont, la plupart du temps, pas sous pression de sélection. Chaque copie observable aujourd'hui a eu sa propre histoire évolutive (Miller and Capy 2006). Les multi-copies d'un ET sont donc plus ou moins dégénérées, et retracer son origine évolutive est parfois compliqué. Néanmoins, l'idée sous-jacente est qu'un consensus de séquences de différentes copies d'un élément correspondrait à la forme originelle de l'élément.

Il existe deux grandes classes d'ET, définies par le type de molécule intermédiaire qu'ils utilisent pour transposer (Finnegan 1989) (Piégu et al. (2015) en proposeraient plutôt 8) . Les transposons de classe I utilisent un intermédiaire ARN, alors que les transposons de classe II utilisent un intermédiaire à ADN. Wicker et al. (2007) ont proposé une classification unifiée des ET (**Figure II.5** p.43). Les classes I et II d'ET contiennent plusieurs ordres et chaque ordre contient plusieurs super-familles, et une super-famille rassemble plusieurs familles. Nous verrons dans la suite de cette section que la classification d'un élément dans telle ou telle famille est assez subtile et dépend principalement d'une certaine similarité de séquence dans ses gènes ou ses séquences répétées terminales. Des éléments autonomes (contenant les gènes nécessaires à une transposition) ou non-autonomes peuvent appartenir à une même famille. Dans le cas d'une similarité faible (<80%), une nouvelle famille est créée. Des outils comme PASTEC permettent de classifier automatiquement les séquences consensus des éléments (Hoede et al. 2014).

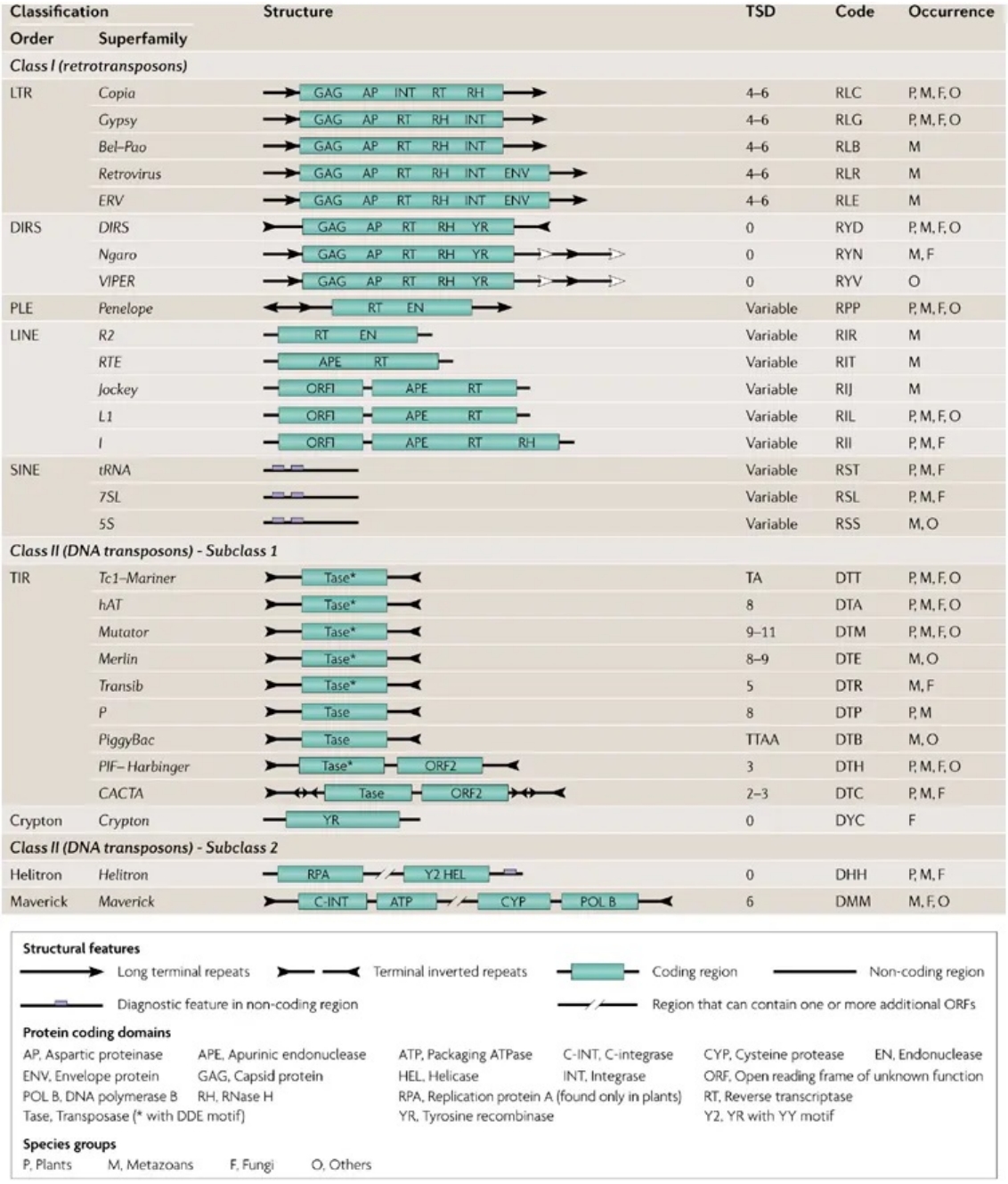


FIGURE II.5 – *La classification des transposons*

La classification est hiérarchisée et divise les éléments transposables en deux grandes classes sur la base de la présence ou l'absence d'un intermédiaire de transposition à ARN. Chaque classe est subdivisée en sous-classes, ordres, et super-familles. La taille du site cible (TSD : *target site duplication*) est caractéristique de la plupart des super-familles et peut être utilisé comme signature. Afin de faciliter l'identification, la classification propose un code à trois lettres qui décrit tous les groupes majeurs en plus de la famille de chaque élément transposable. DIRS, *Dictyostelium intermediate repeat sequence* ; LINE, *long interspersed nuclear element* ; LTR, *long terminal repeat* ; PLE, *Penelope-like elements* ; SINE, short interspersed nuclear element ; TIR, terminal inverted repeat. Figure tirée de WICKER ET AL. (2007).

II.2.1.1 Les éléments transposables à ARN

La transposition des ET de classe I utilise un intermédiaire ARN. Une copie génomique du TE est transcrite en ARN puis rétro-transcrite en ADN, afin de s'intégrer à un autre *locus* génomique de l'hôte. Ces transposons à ARN sont également appelés rétrotransposons. Les rétrotransposons sont catégorisés en cinq ordres selon leurs compositions géniques, la structure de leurs séquences terminales et leurs caractéristiques mécanistiques (**Figure II.5** p.43).

Les transposons LTR (*Long Terminal Repeats*), proche des rétrovirus, montrent des répétitions à leurs extrémités. Le mécanisme d'intégration des ET LTR conduit à l'ajout de quelques bases (4 à 6 nucléotides) de chaque côté du site d'insertion. Généralement, les ET à LTR arborent les gènes GAG (protéine de capside), AP (protéinase aspartique), RT (transcriptase reverse), RH (RNase H), INT (intégrase DDE aussi appelée transposase) et parfois ENV (Envelope) . En fonction de la composition génique et la structure, les ET LTR sont divisés en plusieurs super-familles (*Copia*, *Gypsy*, *BelPao*, *Retrovirus* et *ERV*).

Contrairement aux ET de type LTR, les ET appartenant à l'ordre des LINE (*Long Interspersed Nuclear Element*) n'ont pas de répétitions à leurs extrémités et sont partagés en cinq grandes super-familles (*R2*, *RTE*, *Jockey*, *I* et *L1*). Les LINE contiennent notamment une RT. Chez les animaux, les LINE, et notamment les L1, prédominent ($\sim$25% de LINE dont 22% de L1 chez l'humain) alors qu'ils sont relativement rares chez les plantes (Pevsner 2015).

Les éléments autonomes DIRS (*Dictyostelium Intermediate Repeat Sequence*) et PLE (*Penelope*) contiennent une RT. Les SINE (*Short Interspersed Nuclear Element*) sont non-autonomes et portent des séquences promotrices de l'ARN polymérase III. Ils utilisent la machinerie des autres retroélements pour transposer. Les LARD (*Large Retrotransposons Derivatives*) et les TRIM (*Terminal Repeat retrotransposon In Miniature*) sont également des éléments non-autonomes dérivant d'éléments LTR (Kalendar et al. 2004, Witte et al. 2001).

II.2.1.2 Les éléments transposables à ADN

Les éléments transposables de classe II n'utilisent pas d'intermédiaires ARN pour transposer. Ils sont présents chez la plupart des eucaryotes mais également chez les bactéries sous la forme d'IS (*Insertion Sequence*) (Mahillon and Chandler 1998). En général, les ET de classe II occupent une proportion du génome moins importante que les éléments de classe I. A titre d'exemple, seul 3% du génome humain correspond à des ET à ADN contre 21% de LINE (Treangen and Salzberg 2011). Le mécanisme de coupure des brins d'ADN définit deux sous-classes d'ET à ADN (**Figure II.5** p.43).

Les éléments à ADN de sous-classe 1 ont un mécanisme de transposition dit de "couper-coller". En effet, l'ADN est cassé sur les deux brins afin que l'élément s'excise du *locus* d'origine. L'élément excisé peut alors se ré-insérer à un autre *locus*. La transposition elle-même n'entraine pas une augmentation du nombre de copies de l'élément. Seul

une transposition d'un ET d'un *locus* déjà répliqué à un *locus* pas encore répliqué conduit à l'ajout d'une copie de l'élément. La sous-classe 1 comporte les éléments de l'ordre TIR (*Terminal Inverted Repeat*) et de l'ordre des *Cryptons*. Les éléments TIR sont caractérisés par des répétitions inversées à leurs bornes. Ils contiennent un gène codant pour une transposase avec, en général, un motif protéique (DDE ou DDD) caractéristique d'une activité catalytique. D'après la classification de Wicker et al. (2007), les TIR sont divisés en neuf super-familles : *Tc1-Mariner*, *hAT*, *Mutator*, *Merlin*, *Transib* (en lien avec la protéine Rag1 de la recombinaison V(D)J), *P*, *PiggyBac*, *PIF-Harbinger* et *CACTA*. Le nombre, la composition des nucléotides ajoutés lors de l'événement d'insertion, les cicatrices laissées lors de son départ ainsi que le motif préférentiel d'insertion séparent ces super-familles. Par exemple, les ET *Tc1-Mariner* s'insèrent au niveau d'un site TA. Leur intégration entraine une duplication du site TA, et leur excision laisse quelques nucléotides, trace (*footprint*) de son passage. En revanche, les ET de la famille *PiggyBac* préfèrent s'insérer dans un site TTAA et n'entrainent aucune cicatrice après excision (Skipper et al. 2013). Ces deux familles d'éléments sont particulièrement importantes pour comprendre les réarrangements programmés de génome de la paramécie (voir **section III.3.2** p.68). Les MITE (*Miniatures Inverted Repeats Transposable Elements*) grossissent la liste des éléments de la sous-classe 1 d'ET à ADN. Ces éléments non-autonomes sont bornés par des séquences répétées inversées et utilisent la machinerie des ET autonomes de classe II pour transposer.

Contrairement à la sous-classe 1, la sous-classe 2 d'ET ne coupe qu'un seul des deux brins d'ADN de la copie donneuse lors de leur transposition. La réplication des brins permet de générer une nouvelle copie pouvant s'insérer à un autre *locus*. Comme les ET de classe I, ces transposons sont dits "copier-coller". Il existe les familles *Helitron* (utilisant un système de réplication en "cercle roulant" ou *rolling circle*) et *Maverick*. Découverts et particulièrement abondants chez les plantes, mais également chez les ciliés (voir **section VII.2** p.172), les *Helitrons* ont des mécanismes de transposition encore mal compris (Kapitonov and Jurka 2007). Comme tous les ET, ils peuvent influer sur l'expression des gènes et semblent avoir un rôle majeur dans l'évolution des génomes hôtes (Thomas and Pritham 2015) .

II.2.2 Méthodes d'annotation

À la vue de l'occupation des ET dans les génomes, de nombreuses méthodes computationnelles ont été développées pour les annoter. Évidemment la nature répétée des ET complique grandement la tâche. Et pour compliquer le tout, les ET ont une fâcheuse tendance à s'insérer préférentiellement dans d'autres copies d'ET. Dans les génomes aujourd'hui, on observe les copies d'un ET pouvant être plus ou moins altérées et donc parfois éloignées de la forme active de l'ET d'origine. A partir de ces copies, on réalise une séquence consensus représentant la meilleure approximation disponible de l'élément originel actif (Bergman and Quesneville 2007). Les séquences consensus sont utilisées comme "sonde" pour retrouver les copies, car la distance évolutive entre une copie et le consensus est souvent plus faible que la distance entre deux copies.

L'annotation des ET consiste en deux étapes : la découverte des ET, autrement dit la recherche des familles de séquences répétées, puis la détection des ET avec une identification exhaustive des copies des familles précédemment identifiées. Des programmes et méthodes sont spécialisés pour chacune de ces deux étapes. Ces méthodes se regroupent en quatre grandes catégories (i) les méthodes basées sur l'homologie de séquence avec des éléments connus (ii) les méthodes dites *de novo* qui tentent de trouver les ET en se basant sur leur nature répétée (iii) les méthodes recherchant des caractéristiques structurelles des ET (iv) les méthodes utilisant la comparaison de génomes (**Figure II.6** p.47). Comme souvent, une approche combinant plusieurs méthodes donne souvent les meilleurs résultats (Quesneville et al. 2005). La suite logicielle REPET (Flutre et al. 2011) intègre et combine un certain nombre des approches décrites ci-dessus. Les revues Bergman and Quesneville (2007), Lerat (2010), Goerner-Potvin and Bourque (2018) font le tour des différentes approches que je ne vais détailler que sommairement.

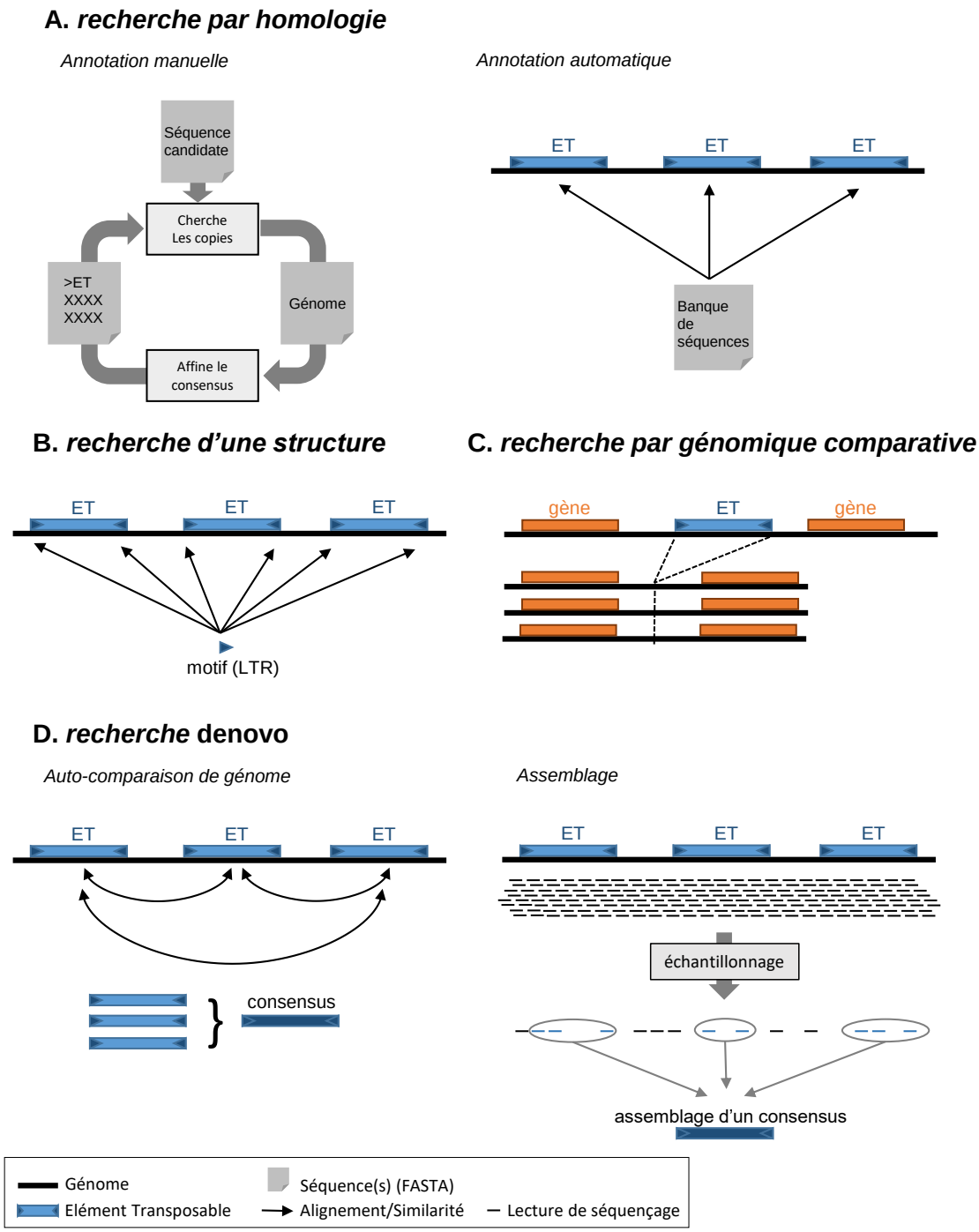


FIGURE II.6 – *Les grandes stratégies d'annotation des éléments transposables*

A. La recherche par homologie peut se faire de manière manuelle ou automatique. L'approche manuelle va partir d'une séquence cible. La découverte de copies va améliorer la séquence consensus de l'élément permettant l'identification de nouvelles copies. La recherche automatique va chercher l'ensemble des copies sur un génome à partir d'une banque de séquences consensus **B.** La recherche de structures va détecter des caractéristiques propres à chaque type d'élément. La recherche de motifs LTR en est un exemple. **C.** Des études de génomique comparative vont détecter des insertions récentes d'éléments dans une espèce par rapport à des espèces proches. **D.** Approche *de novo* La méthode d'auto-comparaison de séquences similaires au sein d'un génome. L'alignement multiple des occurrences permet de déduire une séquence consensus. L'approche *de novo* par assemblage utilise des données de séquençage. Par échantillonnage des lectures, seules les régions répétées auront une chance d'être assemblées.

II.2.2.1 Approches basées sur la similarité de séquence

Les occurrences d'ET sont recherchées dans un génome à partir de séquences consensus d'ET. Ces séquences peuvent être des séquences nucléotidiques (des copies ou des séquences consensus correspondant à l'élément d'origine) ou des séquences protéiques (traduction des phases ouvertes de lecture des séquences consensus). Les approches par similarité sont impliquées dans les deux étapes de l'annotation des ET : la découverte et la détection. Ces méthodes sont assez sensibles et détectent des éléments en faible nombre de copies mais privilégient l'identification de familles déjà connues ou ayant suffisamment peu divergé. Par essence, une recherche avec des séquences protéiques se limite aux super-familles d'ET montrant une partie codante. Des traces de séquences codantes sont scrutées sur le génome à partir de banques de séquences protéiques, et à l'aide d'outils comme BLAST (Altschul et al. 1990) ou à partir de profils HMM Pfam (El-Gebali et al. 2019) avec des outils comme HMMER3 (Eddy 2009). La **figure II.6** (p.47) schématise deux approches par similarité de séquence. L'approche dite "manuelle" utilise une séquence amorce et par des recherches récursives de copies identifie et raffine de nouveaux éléments. Le travail d'annotation des ET du génome de la paramécie a été réalisé en suivant cette méthodologie (Arnaiz et al. 2012, Guérin et al. 2017). L'approche automatique utilise une banque de séquences pour retrouver l'ensemble des copies des éléments. La base de données Dfam rassemble des profils HMM d'une collection d'alignements de séquences consensus ou de copies d'ET (Hubley et al. 2016). De son côté la bibliothèque de séquences nucléotidiques RepBase regroupe des séquences d'éléments mobiles pour une variété d'eucaryotes (Jurka et al. 2005, Bao et al. 2015). RepBase reste une ressource incontournable dans le domaine. Des programmes comme RepeatMasker (Smit 1996), Censor (Jurka et al. 1996), BLASTER (Quesneville et al. 2003) ou simplement BLAST (Altschul et al. 1990) peuvent l'exploiter. Les données de RepBase peuvent être complétées avec des séquences plus proches de l'espèce cible, provenant d'annotations manuelles ou d'autres types d'approches (par exemple *de novo*). Contrairement à l'approche *de novo*, l'approche basée sur l'homologie ne détecte pas des éléments très divergents ou complètement nouveaux.

II.2.2.2 Approches *de novo*

Les approches *de novo* profitent de la nature répétée des ET. Sans aucune information préalable, ces méthodes visent à découvrir tous les éléments répétés d'un génome et potentiellement de nouveaux éléments. Plus l'élément sera répété dans le génome plus il sera aisé de le détecter. A l'inverse, les éléments avec un faible nombre de copies, ou avec des copies très dégénérées d'un même élément, seront difficilement détectables. Le nombre croissant de génomes disponibles rend ces approches particulièrement attractives. Malheureusement, les approches *de novo* génèrent de nombreux faux positifs. Trois classes

de méthodes se distinguent : (i) par analyse de K-mer (ii) par auto-comparaison de génome (iii) par assemblage.

Analyse de K-mer Des programmes comme REPuter (Kurtz and Schleiermacher 1999), REAS (Li et al. 2005), Tallymer (Kurtz et al. 2008), Jellyfish (Marçais and Kingsford 2011) ou RepeatExplorer (Novák et al. 2013) repèrent des K-mers sur-représentés dans la séquence génomique ou dans les lectures de séquençage. Ces "mots" correspondent, potentiellement, à des portions d'éléments répétés. Au delà de l'annotation des ET, ces méthodes sont également utilisées pour masquer les régions répétées. Les analyses de K-mer détectent aussi bien des familles multi-géniques, des duplications segmentales que de véritables portions d'ET (Lerat 2010).

Auto-comparaison de génome Comme leur nom l'indique, les méthodes d'auto-comparaison de génome alignent l'ensemble du génome sur lui-même. Les séquences similaires sont regroupées et des alignements multiples permettent de déduire des séquences consensus correspondant à des familles d'éléments. Les logiciels comme RECON (Bao and Eddy 2002), PILER (Edgar and Myers 2005) ou BLASTER (Quesneville et al. 2005) diffèrent par le programme utilisé pour l'alignement du génome mais surtout par l'algorithme de regroupement des séquences similaires. Cette procédure s'effectuant à partir d'un assemblage de génome, la qualité de celui-ci (le bon assemblage des régions répétées, sans trop de *collapse* voir **section I.3.4** p.28) est critique.

Assemblage L'approche *de novo* utilise des données de séquençage pour effectuer l'assemblage des ET. Nous avons vu dans le chapitre précédent que les assembleurs génomiques classiques éprouvent quelques difficultés avec les séquences répétées. Pourtant, Tedna est un exemple d'assembleur de données *Illumina* dédié aux ET (Zytnicki et al. 2014). Par ailleurs, l'outil dnaPipeTE (Goubert et al. 2015) est basé sur le postulat que des lectures correspondant à des régions répétées seront sur-représentées dans un séquençage, par rapport à des régions non répétées. Par échantillonnage extrême (<1X) des lectures, puis une procédure d'assemblage, seules les régions répétées auront une chance d'être assemblées. Évidemment plus les copies des ET seront dégénérées, plus il sera difficile de les reconstituer.

II.2.2.3 Approches basées sur la structure des ET

Ces approches détectent les caractéristiques structurelles de certains types d'ET. Elles recherchent des répétitions terminales (LTR,TIR) (**Figure II.6** p.47) ou des motifs particuliers comme les sites dupliqués lors de l'insertion d'un élément. Par définition, les familles d'ET non-structurés ne sont pas détectables. De plus, ces méthodes se basent exclusivement sur notre connaissance des ET et de leurs caractéristiques communes. De nouveaux éléments peuvent être trouvés mais pas de nouvelles classes. Beaucoup de pro-

grammes sont spécifiques d'une certaine classe d'élément ou d'un certain type de structure. Sans être exhaustif, en voici quelques exemples : LTRHarvest (Ellinghaus et al. 2008), RTanalyzer (Lucier et al. 2007), SINEDR (Tu et al. 2004), FindMITE (Tu 2001) ou HelitronFinder (Du et al. 2008).

II.2.2.4 Approches basées sur la génomique comparative

Les outils basés sur la génomique comparative utilisent, comme son nom l'indique, les génomes de plusieurs espèces (ou souches). L'analyse des ruptures d'alignements multiples de ces génomes fait ressortir des régions où, potentiellement, un ET s'est inséré (**Figure II.6** p.47) (Caspi and Pachter 2006, Quadrana et al. 2016). Afin de mettre en œuvre ce genre d'approche, il faut avoir accès à des génomes de bonne qualité et de plusieurs espèces adéquatement distantes les unes des autres. En effet, si les génomes sont trop proches, évolutivement parlant, alors aucune insertion ne pourra être détectée. Au contraire, si les génomes sont trop distants alors les génomes auront du mal à s'aligner et la plupart des différences seront dues à des réarrangements chromosomiques et non pas des insertions d'ET.

Chapitre III

La paramécie

Mon travail de thèse porte sur l'annotation des génomes de paramécies. Dans ce chapitre, je décrirai la place phylogénétique des paramécies, et je présenterai des notions sur la biologie de ces organismes, ainsi que leurs caractéristiques génomiques.

Les paramécies sont des eucaryotes unicellulaires appartenant au groupe des ciliés. Les ciliés sont des organismes phylogénétiquement éloignés des autres organismes modèles. Ils ont la particularité de présenter deux types de noyaux dans leur cytoplasme : un noyau germinal et un noyau somatique (voir **Figure III.1** p.52). Ces noyaux contiennent des matériels génétiques différents devant être annotés spécifiquement.

III.1 Place du modèle paramécie

III.1.1 Les eucaryotes

La **figure III.2** (p.54) montre la place des ciliés dans un arbre phylogénétique des eucaryotes. Appartenant au groupe des alvéolés, *Paramecium* et *Tetrahymena* sont les ciliés les plus étudiés dans le monde. Ces deux espèces sont, néanmoins, très éloignées phylogénétiquement. Leur divergence est au moins comparable à la séparation des mammifères et des arthropodes (300 à 500 Ma) (Baroin-Tourancheau et al. 1992, Xiong et al. 2019). Les ciliés sont dans une radiation très profonde de l'arbre des eucaryotes et regroupent des organismes morphologiquement différents (voir **Figure III.4** p.55). La précision du positionnement phylogénétique des espèces dans les arbres dépend essentiellement des séquences fournies pour les construire (Adl et al. 2012) (voir **section I.1.1** p.3). Plus un groupe d'espèces est étudié, et donc plus nous avons accès à une variété de séquences, plus le positionnement phylogénétique sera stable et robuste. Les organismes de laboratoire ne sont qu'un échantillonnage non-représentatif du vivant. Il est vrai que les espèces avec un enjeu économique ou un lien avec la santé humaine sont préférées. Par exemple, les groupes des animaux, champignons et plantes représentent 85% des génomes séquencés alors qu'ils ne représentent que 23% de la biodiversité (Burki 2014) (voir **section I.3** p.16). Pourtant, à l'image de TARA Océan, des projets métagénomiques à très grande

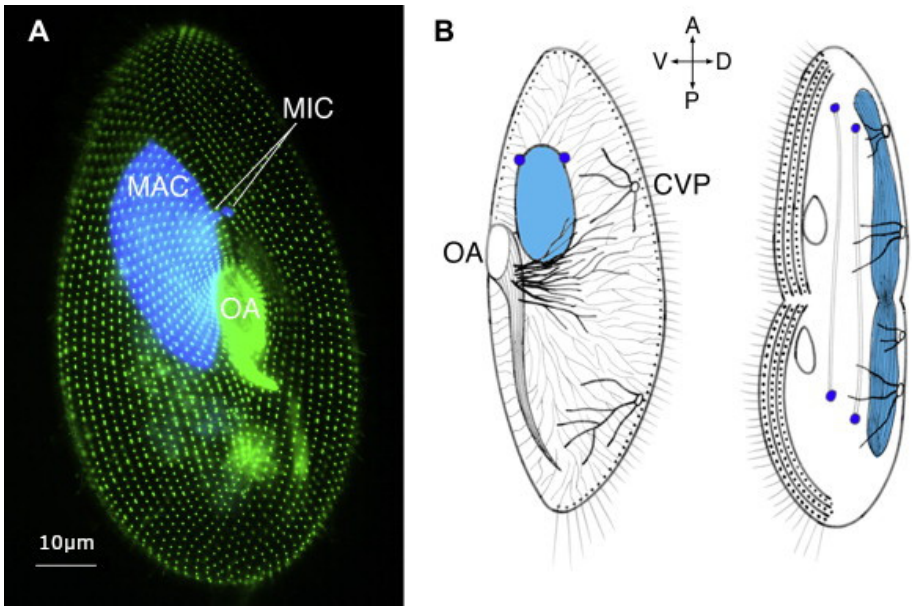


FIGURE III.1 – *Organisation cellulaire de la paramécie*

A. Paramécie en immunoflorescence. Les corps basaux couvrant le cortex cellulaire sont marqués en vert. Les corps basaux se prolongent par un organite ciliaire. Les 2 petits noyaux germinaux (MIC pour *micronucleus*) et le gros noyau somatique (MAC pour *macronucleus*) apparaissent en bleu par un marquage DAPI. La bouche (OA pour *Oral Apparatus*) permet l'ingestion de particules alimentaires. L'OA est très brillant en raison d'une densité en corps basaux très importante. **B.** La cellule de $\sim$120 μm présente une polarité antéro-postérieure et dorso-ventrale. Les vacuoles pulsatiles (CVP pour *Contractile Vacuole Pore*) ont un rôle dans l'équilibre hydrique de la cellule, et sont rigidifiées par un réseau microtubulaire. Le dessin de droite schématise une paramécie en division végétative avec une ségrégation des noyaux germinaux dans les deux cellules filles et une fission a-mitotique du noyau somatique. Figure tirée de SPERLING (2011)

échelle voient le jour, facilités par la modernisation des technologies de séquençage (Carradec et al. 2018). L'émergence de ces programmes de recherche permet de combler les lacunes de notre vision du monde vivant

III.1.2 Les ciliés

Le groupe monophylétique des alvéolés regroupe un grand nombre d'espèces présentant une large diversité morphologique. Les **figures III.2** (p.54) et **III.3** (p.55) montrent la division des *Alveolata* en trois sous clades. Les ciliés (*Oxytricha*, *Euplotes*, *Paramecium*, ou *Tetrahymena*), les dinoflagellés (*Symbiodinium*) et les apicomplexes (*Toxoplasma* ou *Plasmodium*) parasites des métazoaires. Les *Ciliophora* sont des protozoaires unicellulaires caractérisés par la présence de cils à leurs surfaces. Ces cils motiles servent essentiellement à la locomotion et à la nutrition. Ces organismes vivent en milieu aqueux (eaux douces, marines ou saumâtres) et existent sous forme libre, parasitaire ou symbiotique. Ils se nourrissent de particules organiques ou bactéries, et ne sont pas contre le cannibalisme.

Les génomes les plus étudiés sont probablement ceux de *Paramecium*, *Tetrahymena* et *Oxytricha*. Bien que petite, la communauté des laboratoires travaillant sur les ciliés est répartie dans le monde. Les ciliés sont des organismes intéressants pour plusieurs aspects. En écotoxicologie, ils sont de bons bioindicateurs de la pureté des sols et des eaux (lac, rivière ou eaux usées) (Lara and Acosta-Mercado 2012). La structure et le fonctionnement des cils sont très conservés au cours de l'évolution. Une large communauté s'intéresse au fonctionnement et la biogenèse des cils et flagelles, en lien avec des maladies humaines (les ciliopathies) (Bachmann-Gagescu 2014, Mitchison and Valente 2017). Pour intégrer des données extraites d'études ciliaires à haut débit, en connexion avec des informations génétiques sur des maladies humaines, j'ai construit, en collaboration avec l'équipe de J. Cohen et AM Tassin, un système d'information : Cildb (Arnaiz et al. 2009; 2014).

Plus en lien avec le cœur de mon sujet de recherche, les ciliés possèdent une autre particularité fascinante pour des unicellulaires. Ils présentent deux types de noyaux : un micronoyau (MIC pour *micronucleus*) germinal subissant la méiose et transmettant l'information génétique à la génération suivante et un macronoyau (MAC pour *macronucleus*) somatique assurant la transcription génique pendant la vie végétative de la cellule. Nous verrons qu'à chaque cycle sexuel, le MAC polyploïde est perdu et un nouveau MAC est généré à partir d'un MIC. Des processus de développement impliquant des réarrangements de génomes avec élimination programmée de matériel génétique ont lieu chez la plupart des grands groupes de métazoaires (Wang and Davis 2014), mais les ciliés sont les seuls unicellulaires à le faire (voir **section III.3.1** p.64). A leur manière, les ciliés ont différencié leur lignée germinale et somatique. L'organisation, la taille, la ploïdie et le nombre de ces deux types de noyaux varient selon les espèces. Du point vue de l'annotation, et même s'ils dérivent l'un de l'autre, ces deux noyaux contiennent des génomes différents.

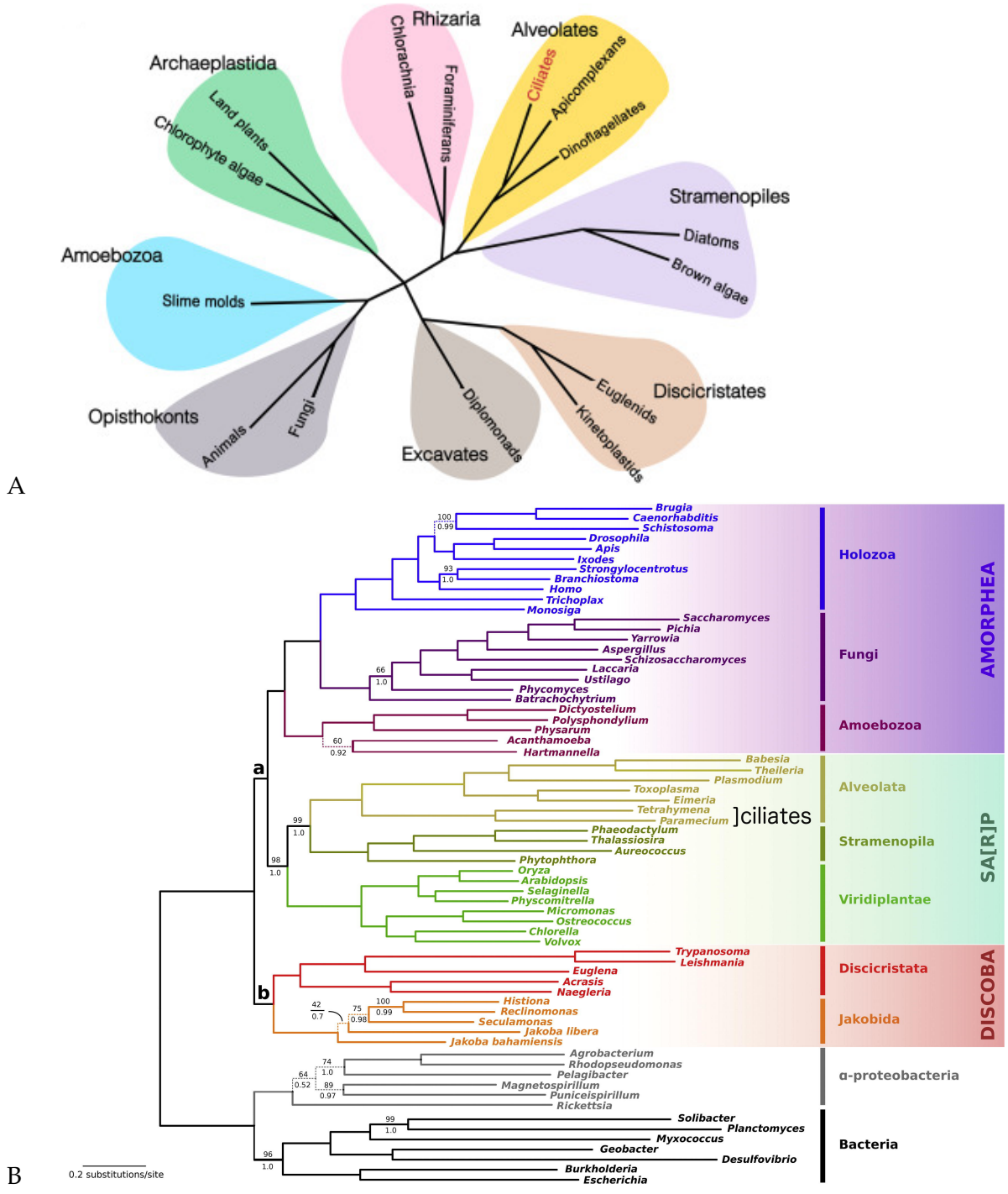


FIGURE III.2 – *Arbres phylogénétiques des eucaryotes*

A. Arbre sans racines (basé sur un consensus de données moléculaires et ultrastructurales) donnant une image large de la diversité des eucaryotes. *Paramecium* appartient au phylum des ciliés (*ciliates*), une partie du clade alvéolée. Figure tirée de (BALDAUF 2003) et modifiée par DUHARCOURT AND SPERLING (2018). **B.** Phylogénie enracinée plus fine des eucaryotes basée sur un jeu de 37 protéines eucaryotes avec une certaine homologie chez la bactérie. Les groupes d'espèces sont indiqués en couleur. Figure tirée de HE ET AL. (2014).

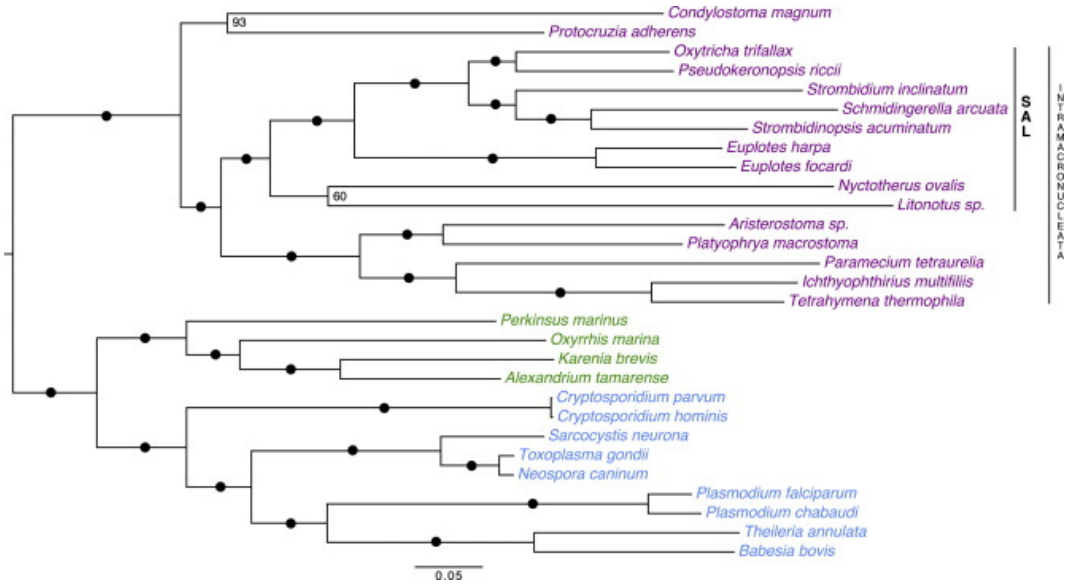


FIGURE III.3 – *Phylogénie des alvéolés*

Le groupe des *alveolata* comporte les ciliés (violet), les dinoflagellés (vert) et les parasites apicomplexes (bleu). La phylogénie est basée sur un jeu de 158 gènes. Les valeurs de bootstrap sont indiquées sur les nœuds, les ronds noirs indiquent une valeur de 100%. Figure tirée de GENTEKAKI ET AL. (2014).

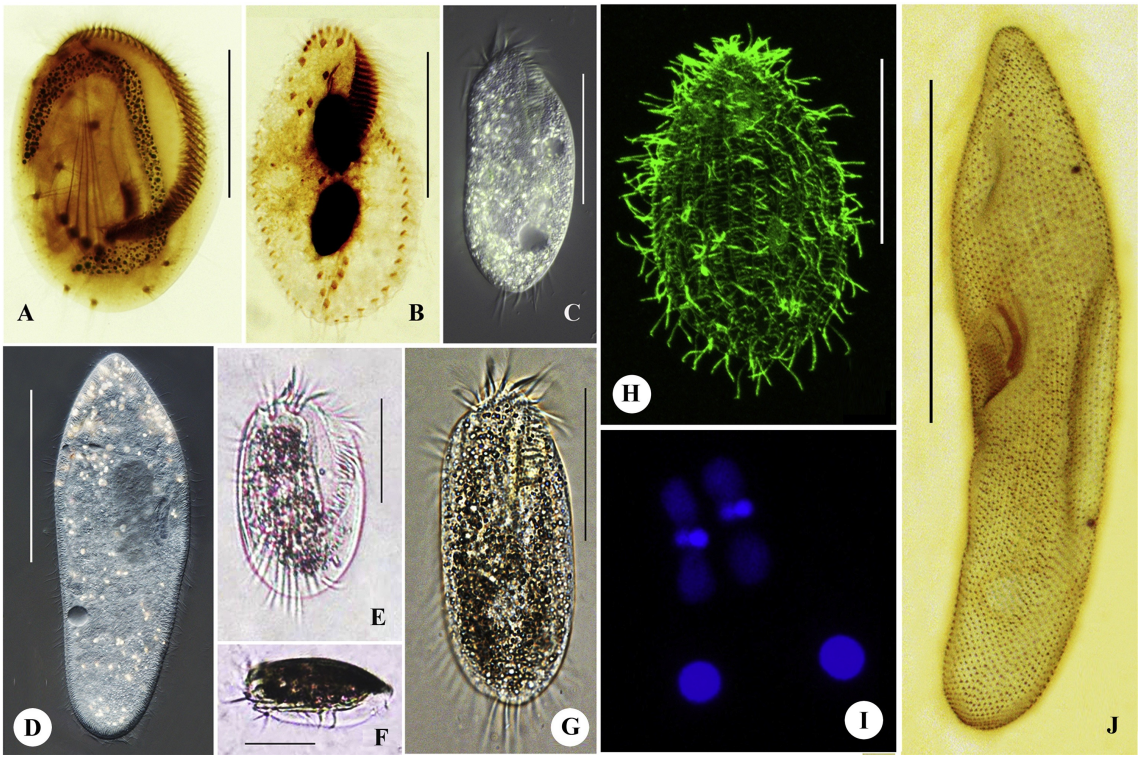


FIGURE III.4 – *Images de ciliés*

Images de différents ciliés : *Euplotes harpa* (A, E et F), l'hypotriche *Oxytricha triffalax* (B, C et G), *Paramecium caudatum* (D et J) et *Tetrahymena thermophila* (H, I). Barres d'échelles : 60 μm (A, E, F), 50 μm (B, C, G), 20 μm (H), 100 μm (D, J). Figure tirée de WANG ET AL. (2017).

III.1.3 Les paramécies

Aux États-Unis, dans les années 1930, T.M. Sonneborn a établi l'organisme *Paramecium* comme modèle d'étude. Il a été séduit par cette grande cellule ($\sim$120-150μm) facilement observable à la loupe binoculaire et se prêtant bien à diverses études génétiques. Des tests de croisements ont dévoilé l'existence d'un groupe *aurelia* regroupant au moins 15 espèces morphologiquement indiscernables mais sexuellement incompatibles (voir **Figure III.5** p.57) (Sonneborn 1975, Coleman 2005). Après un séjour dans le laboratoire de T.M. Sonneborn, J. Beisson a exporté le modèle paramécie en France et notamment l'espèce *Paramecium tetraurelia*. Facilement cultivable, cette espèce est particulièrement bien adaptée aux conditions de laboratoire. De plus, des croisements, et donc des études génétiques, sont possibles.

Outre une hérédité mendélienne, la paramécie révèle une transmission héréditaire cytoplasmique d'un certain nombre de caractères. L'héritage du schéma cortical des corps basaux (structure analogue aux centrioles et socle d'assemblage de l'organite ciliaire) dans la cellule en est un exemple (**Figure III.1** p.52) (Beisson 2008). Nous verrons dans la **section III.3.2** (p.68) que le développement du MAC met également en jeu une hérédité macronucléaire non-mendélienne impliquant des mécanismes épigénétiques.

Le premier génome MAC de paramécie (*Paramecium tetraurelia*) a été publié en 2006 (Aury et al. 2006), puis les génomes MAC de *Paramecium caudatum*, *Paramecium biaurelia* et *Paramecium sexaurelia* en 2014 (McGrath et al. 2014b;a). De récentes études affinent le positionnement phylogénétique des paramécies par l'alignement de séquences protéiques mitochondriales (Yi et al. 2014, Johri et al. 2019, Arnaiz et al. 2019). Sur la **figure III.5** (p.57), le groupe *aurelia* se divise en trois sous-clades : une sous-clade contenant *P. sexaurelia*, *P. jenningsi* et *P. sonneborni*, une sous-clade avec *P. decaurelia*, *P. dodecaurelia*, *P. octaurelia* et *P. tetraurelia*, et enfin une sous-clade avec *P. biaurelia*, *P. novaurelia*, *P. primaurelia*, *P. pentaurelia*, *P. quadecaurelia* et *P. tredecaurelia*. Les espèces *P. caudatum* et *P. multimicronucleatum* restent assez distantes du groupe *aurelia*.

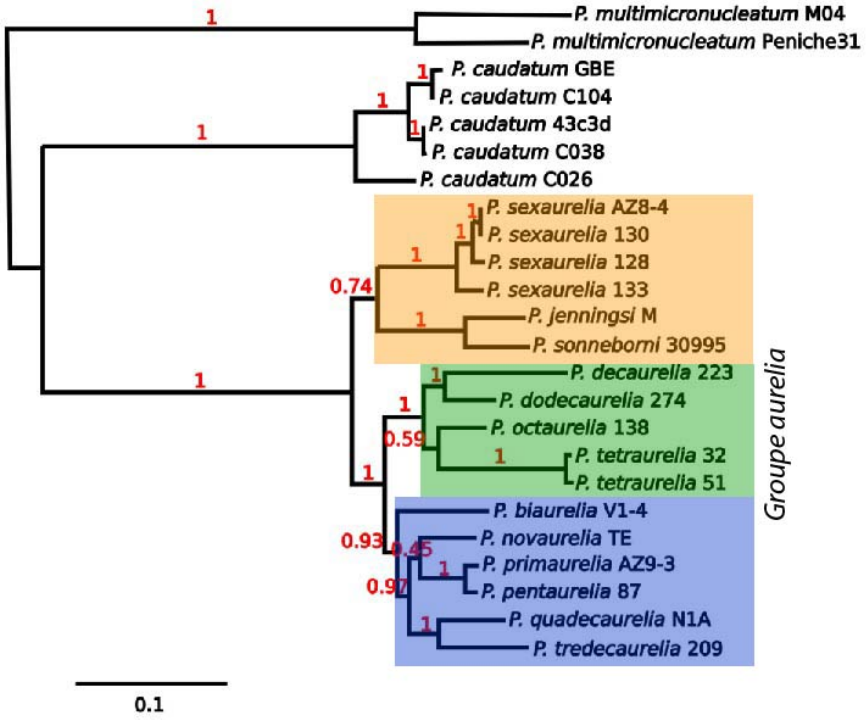


FIGURE III.5 – *Phylogénie des paramécies*

Arbre phylogénétique d'espèces de paramécies basé sur 46 protéines mitochondriales. Le groupe d'espèces *aurelia* trois sous-clades : une sous-clade avec *P. sexaurelia*, *P. jenningsi* et *P. sonneborni* (en orange), une sous-clade avec *P. dec-*, *P. dodec-*, *P. oct-* et *P. tetr-aurelia* (en vert), et enfin une sous-clade avec *P. bi-*, *P. nov-*, *P. prim-*, *P. pent-*, *P. quadec-* et *P. tredec-aurelia* (en bleu). Figure modifiée de ARNAIZ ET AL. (2019).

III.2 La biologie de la paramécie

III.2.1 Le cycle végétatif

Durant sa vie végétative, la paramécie se divise par fission binaire. Le noyau MAC se sépare en deux compartiments sans mitose, et de façon non équitable (Tucker et al. 1980). Après mitose fermée, les noyaux MIC se répartissent dans les deux cellules filles (**Figure III.1B** p.52). En milieu riche en bactéries et dans des conditions de laboratoire, *Paramecium tetraurelia* réalise environ 4 à 5 divisions par jour. Après une centaine de divisions végétatives sans processus sexuel, la paramécie meurt par sénescence (Gilley and Blackburn 1994). En effet, contrairement à l'homme où les chromosomes se raccourcissent par les télomères au cours du vieillissement, les chromosomes MAC se fragmentent au cours de la vie clonale de la paramécie, accumulant des dommages à l'ADN.

III.2.2 Les processus sexuels

En plus des divisions végétatives, les paramécies se reproduisent sexuellement. Contrairement à l'espèce *P. caudatum* (voir **Figure III.5** p.57), les paramécies du groupe *aurelia* sont capables d'opérer deux types de processus sexuels : une conjugaison entre deux cellules de types sexuels compatibles ou une auto-fertilisation appelée autogamie. En laboratoire, la reproduction sexuée est induite par une carence alimentaire.

III.2.2.1 La conjugaison

La conjugaison est une reproduction sexuée entre deux cellules de types sexuels compatibles. L'agglutination de cellules réactives (cellules carencées n'ayant pas encore réalisé trop (∼20) de divisions végétatives) permet la formation de couples (**Figure 1B** p.xi). La **figure III.6** (p.59) schématise l'organisation nucléaire pendant le cycle sexuel de la paramécie. Dans chacun des deux partenaires, un noyau parmi huit produits méiotiques MIC n'est pas dégradé et subit une mitose. La fécondation s'accomplit par échange réciproque des pronoyaux. La fusion des noyaux gamétiques maternel et paternel donne naissance au noyau zygotique (la caryogamie), puis le noyau zygotique subit deux mitoses successives. Deux produits vont devenir les MIC et les deux autres se développeront en nouveaux MAC. Ces ébauches de MAC sont appelées *anlagen*. A la division caryonidale, les noyaux MIC et MAC se répartissent dans les deux cellules filles. Durant toutes ces étapes l'ancien MAC va se fragmenter (en ∼30 compartiments) et se dégrader progressivement. Après quelques divisions végétatives, les fragments de l'ancien MAC auront complètement disparu (Berger 1967).

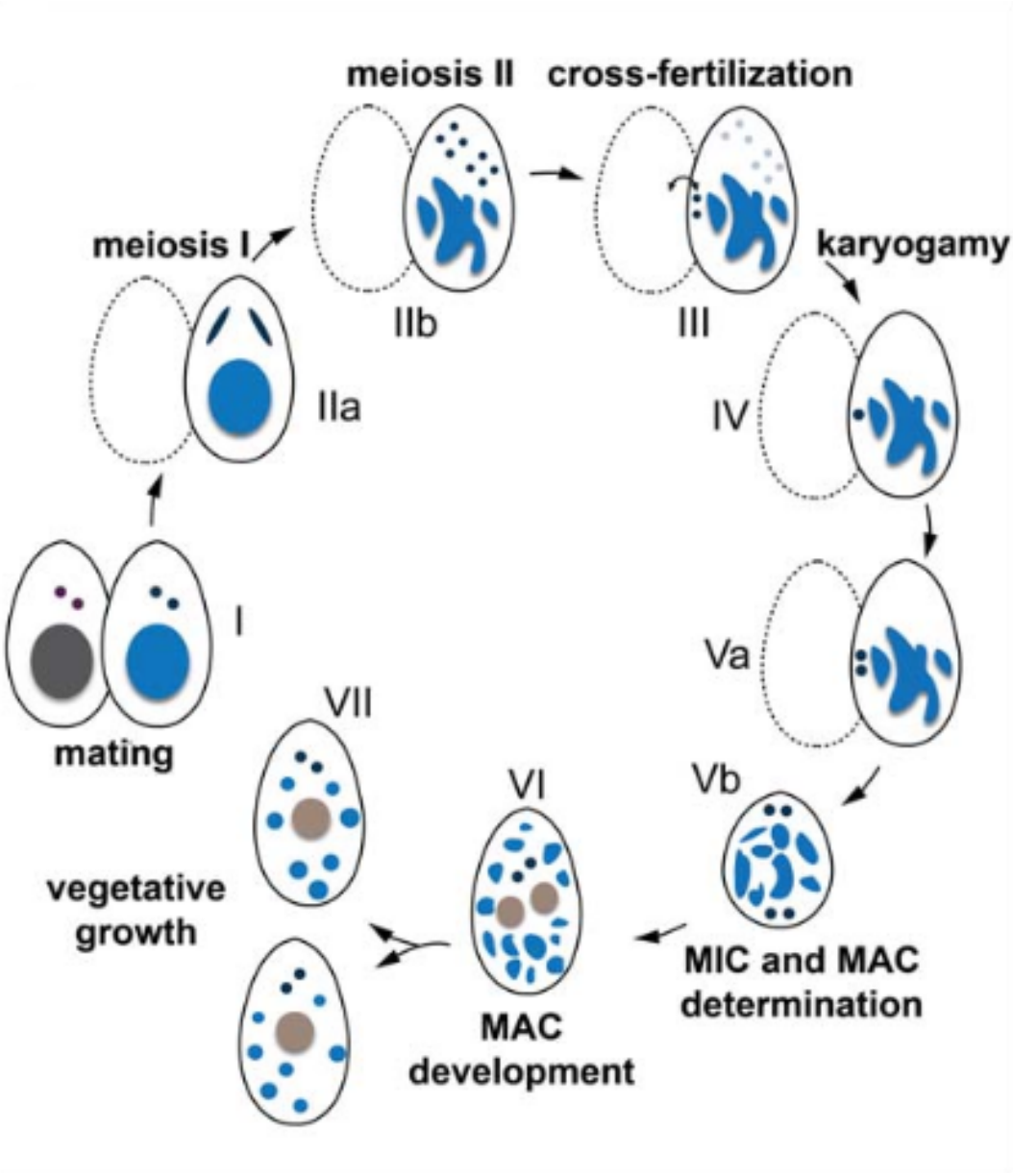


FIGURE III.6 – *Organisation nucléaire pendant le cycle sexuel de P. tetraurelia*

La conjugaison a lieu entre deux cellules de types sexuels compatibles. Après la formation de couples (I), chaque MIC subi une méiose (IIa et IIb). A partir l'étape IIb, le MAC parental, symbolisé en bleu, commence sa fragmentation. En III, chaque partenaire échange l'un de ses huit produits haploïdes tandis que les sept autres sont dégradés. Le noyau zygotique est formé par la fusion des deux pronoyaux (IV). Après deux divisions mitotiques (Va et Vb), les ex-conjugants se séparent. Les futurs noyaux MIC migrent au pôle cellulaire antérieur et les deux nouveaux MAC au pôle postérieur (Vb). A l'étape VI, l'amplification d'ADN et les réarrangements programmés du génome ont lieu dans les nouveaux MAC en développement. Les produits d'une nouvelle mitose des MIC et les MAC en développement se répartissent dans les cellules filles (VII) (division caryonidale). La ploïdie finale de 800n des MAC ne sera atteinte qu'après la division végétative suivante. Schéma tiré de BETERMIER AND DUHARCOURT (2014)

III.2.2.2 L'autogamie

L'AUTOGAMIE est un processus similaire à la conjugaison mais n'engageant qu'une seule cellule (avec un âge clonal d'au moins 20 divisions) : c'est une auto-fécondation. Il n'y a pas d'échange de noyaux. La fusion de deux pronoyaux identiques donne le noyau zygotique. Contrairement à la conjugaison, les noyaux résultant d'une autogamie sont 100% homozygotes à tous les *loci*. Pour cette raison, l'autogamie est un outil très puissant, en particulier pour des analyses génétiques ou pour des analyses génomiques où l'on peut séquencer et assembler l'équivalent d'un ADN "haploïde".

Plusieurs équipes, dont la mienne, étudient les réarrangements programmés de génomes pendant les processus sexuels. Pour des raisons techniques, l'induction de l'autogamie est préférée à la conjugaison. En effet, la conjugaison est plus compliquée à mettre en œuvre, et obtenir une grande quantité de matériel, pour des approches biochimiques ou passant par le séquençage, est problématique. Pourtant, l'autogamie n'a pas que des avantages. En laboratoire, l'autogamie est induite par carence alimentaire de cellules ayant subi une vingtaine de divisions végétatives. Même si toutes les paramécies de la culture vont finir par réaliser l'autogamie, toutes les cellules ne vont pas l'engager en même temps. Cette asynchronie ($\sim$5h) d'entrée en autogamie peut flouter des résultats d'études de cinétiques de cellules pendant les processus sexuels (voir **section IVB** p.89).

III.2.3 La génétique de la paramécie

III.2.3.1 Hérédité cytoplasmique

La **figure III.7a** (p.62) schématise, d'un point de vue génétique, un croisement entre deux cellules homozygotes de génotypes différents. L'une des cellules est sauvage et l'autre mutante pour un gène d'intérêt. Il y a échange de noyaux lors de la conjugaison entre deux paramécies de type sexuel compatible. Après croisement, les cellules de la génération F1 sont génétiquement identiques et hétérozygotes au *locus* d'intérêt grâce à une ségrégation mendélienne classique des allèles. Après auto-fécondation, 50% de la population F2 sera homozygote sauvage et 50% homozygote mutante.

Chez *Tetrahymena*, le type sexuel est déterminé au hasard parmi les 7 types sexuels possibles (Nanney 1953). Pendant la différenciation somatique, le réarrangement d'un *locus* MIC reconstitue un *locus* somatique correspondant à un type sexuel (Cervantes et al. 2013). L'espèce *Paramecium tetraurelia* possède deux types sexuels : O pour *Odd* et E pour *Even*. Comme illustré sur la **figure III.7b** (p.62), ce caractère phénotypique s'hérite maternellement (Sonneborn 1947, Epstein and Forney 1984). Autrement dit, le phénotype O ou E suit une hérédité cytoplasmique. L'expression des gènes mtA, mtB et mtC est requise pour exprimer le type E (Singh et al. 2014). Les produits des gènes mtB et mtC sont nécessaires à l'expression du gène mtA codant pour une protéine transmembranaire. Chez *P. tetraurelia*, la séquence promotrice du gène mtA des cellules du type O, est rendue non fonctionnelle pendant la maturation du nouveau MAC. En effet, une petite séquence de 195 nt, contenant le site de début de transcription et les 26 premiers nucléotides de la séquence codante de mtA, est excisée dans les MAC des cellules de type O, et pas dans les MAC des cellules du type E. Nous verrons dans la **section III.3.1** (p.64) que cette séquence est excisée comme une des 45 000 petites séquences (appelées IES pour *Internal Eliminated Sequence*) que compte le génome MIC de *Paramecium tetraurelia*, et que son élimination est guidée par des mécanismes impliquant une voie de petits ARN.

III.2.3.2 Outils moléculaires : transformation et extinction génique

Contrairement à *Tetrahymena* (Cassidy-Hanley et al. 1997, Gaertig and Kapler 2000), une transformation permanente du génome MIC n'est toujours pas opérationnelle pour la paramécie. En revanche, d'autres outils très puissants ont été développés. Il est possible de faire produire, à la paramécie, une protéine transgénique par microinjection d'ADN dans le noyau MAC. La machinerie cellulaire de la paramécie convertit le fragment d'ADN linéaire en véritable mini-chromosome MAC par télomérisation *de novo* des extrémités. En utilisant cette technique, une localisation subcellulaire de protéines peut être réalisée par injection d'une fusion d'un gène d'intérêt et d'une étiquette comme la GFP (*Green Fluorescent Protein*) (Beisson et al. 2010). Dans l'article de Guérin et al. (2017) (voir **section V.2** des résultats p.135), cette méthode est utilisée pour localiser le variant de l'histone H3 centromérique (CenH3) dans les noyaux germinaux de la paramécie

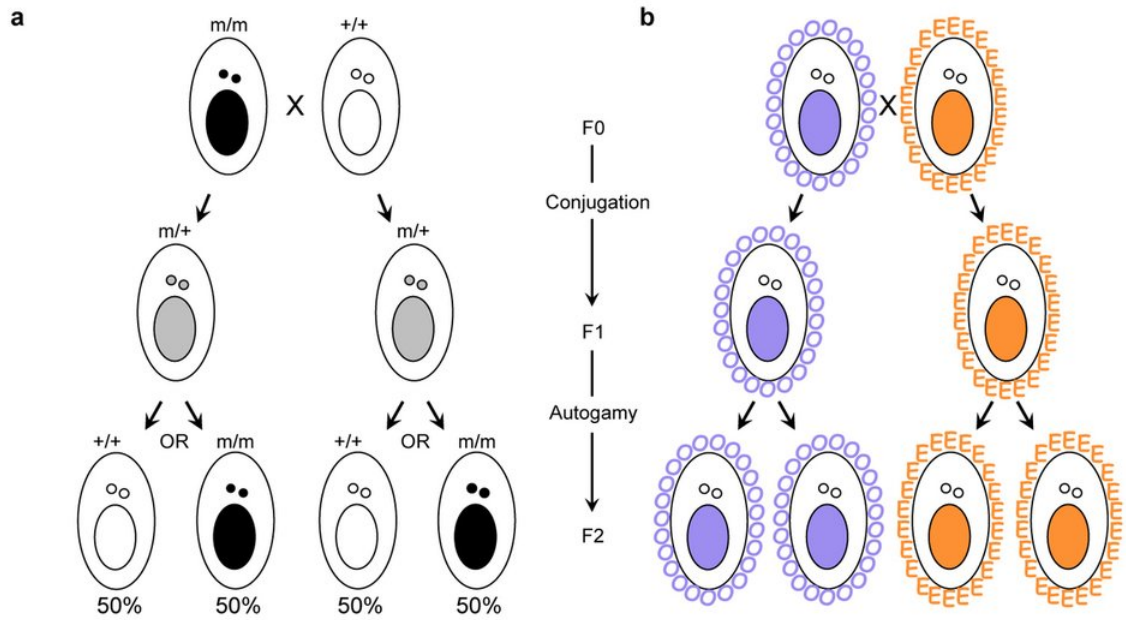


FIGURE III.7 – *Analyse génétique et type sexuel chez P. tetraurelia*

a | Ségrégation mendélienne d'une paire d'allèles. Les cellules différentes (m/m et +/+) échangent un pronoyau et effectuent la caryogamie. En conséquence, les cellules F1 sont génétiquement identiques. Durant l'autogamie suivante, chaque pronoyau fusionne avec une copie de lui-même formant ainsi le noyau zygotique, 100% homozygote sauvage ou mutant. **b** | Hérédité maternelle (cytoplasmique) du type sexuel. Les O et les E autour des cellules représentent le type sexuel exprimé par chaque cellule. Les MAC sont colorés en fonction de leur détermination pour l'un ou l'autre des types sexuels. Figure reproduite de SINGH ET AL. (2014)

(Lhuillier-Akakpo et al. 2016). Ruiz et al. (1998) ont remarqué qu'une injection d'un grand nombre de copies de transgène sans séquences régulatrices, entrainait la répression du gène endogène homologue, par une dégradation post-transcriptionnelle de l'ARNm (Galvani and Sperling 2001). Peu de temps après, la technique d'extinction génique par "alimentation" (*feeding*), initialement développée chez *C. elegans*, a été adaptée à la paramécie (Galvani and Sperling 2002). Un fragment d'ADN du gène d'intérêt entouré de promoteurs convergents inductibles est cloné dans la bactérie *E. coli*. Les bactéries, produisant de l'ARN double brin (ARNdb), sont ingérées par la paramécie et provoquent la répression du gène cible en quelques heures. Les voies d'extinction génique par injection ou par alimentation impliquent la présence de petits ARN de 23 nt, appelés siARN, produits suite au clivage de l'ARNdb précurseur. Des approches de mutagenèse, puis caractérisation de la mutation par séquençage global, ont permis de décrypter les acteurs impliqués, et notamment les protéines du type Dicer (endoribonucléase), Piwi (protéine liant l'ARN) et RdRP (ARN polymérase ARN-dépendante) (Marker et al. 2014).

La facilité de mise en œuvre de techniques comme l'extinction génique par alimentation ou la transformation par injection rend le modèle paramécie très attractif. Là ou plusieurs mois sont nécessaires pour obtenir l'invalidation génique par recombinaison chez *Tetrahymena*, 24 à 48 heures suffisent pour établir un phénotype (végétatif) chez la paramécie par ARN interférence.

III.3 Génomique de la paramécie

III.3.1 Les génomes micronucléaire et macronucléaire

Nous avons vu précédemment que les ciliés présentent un dimorphisme nucléaire. Le MIC est le noyau germinal transmettant le matériel génétique à la descendance. Le MIC est transcriptionnellement éteint pendant la vie végétative. En revanche, toute la transcription génique, nécessaire à la vie cellulaire, a lieu dans le noyau MAC somatique, Bien qu'étant originaire d'un même noyau zygotique (**Figure III.6** p.59), les noyaux MIC et MAC sont très différents. En effet, le petit MIC est diploïde, alors que le gros MAC est hautement polyploïde. Le nouveau MAC en développement subit plusieurs vagues d'endoréplication pour passer d'un stade diploïde à un stade polyploïde (**Figure III.8** p.65) (Berger 1973). L'ébauche endure également des réarrangements programmés de génomes (Coyne et al. 2012, Betermier and Duharcourt 2014). Ces réarrangements programmés de génome consistent en une fragmentation des chromosomes MIC et une perte de matériel génétique. Par exemple, contrairement aux chromosomes MIC, les chromosomes MAC sont dépourvus de centromères actifs, démontré par une absence de marquage des noyaux par le variant d'histone centromérique (Cervantes et al. 2006, Cui and Gorovsky 2006, Lhuillier-Akakpo et al. 2016) (voir **section I.2.1** p.7). La majorité des séquences perdues pendant la maturation du MAC correspondent à des séquences répétées, satellites et transposons. Si ces processus biologiques sont communs aux ciliés, les modalités et la nature des événements diffèrent. En effet, la situation est relativement différente si l'on étudie *Tetrahymena*, *Paramecium* ou *Oxytricha* (**Table III.1** p.64 et **Figure III.8** p.65).

	Tetrahymena thermophila	*Paramecium tetraurelia*	*Oxytricha trifallax*
Nombre de chromosome MIC	5	$\geq$ 50	?
Ploïdie du MIC	2	2	2
Taille du génome MIC	157 Mb	~100 Mb	~500 Mb
Contenu en G+C du MIC	22%	27%	28%
Nombre d'IES	12 000	45 000	150 000
Complexité éliminée (%)	34%	28%	90%
Nombre de chromosome MAC	181	~150	15600
Ploïdie du MAC	45	800	~2 000
Taille du génome MAC	103 Mb	72 Mb	50 Mb
Nombre de gènes	24 700	40 460	~18 400
Contenu en G+C du MAC	22%	28%	31%

Table III.1 – *Statistiques sur les génomes haploïdes MIC et MAC de Tetrahymena, Paramecium et Oxytricha*

Modifié de Wang et al. (2017)

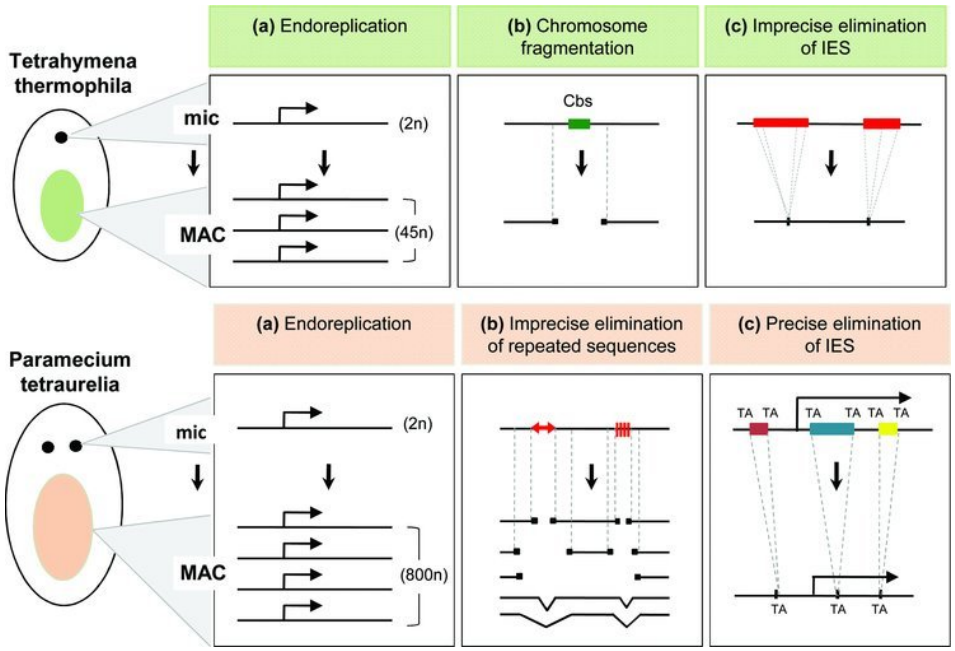


FIGURE III.8 – *Dimorphisme nucléaire et réarrangements de l'ADN chez Paramecium et Tetrahymena*

Ce schéma représente les réarrangements programmés de génome entre le noyau MIC et le noyau MAC chez *Tetrahymena thermophila* (en haut en vert) et *Paramecium tetraurelia* (en bas en saumon). La figure est commentée dans le texte. Figure tirée de COYNE ET AL. (2012)

Tetrahymena thermophila a un micronoyau contenant 5 chromosomes métacentriques avec une complexité de génome haploïde de 157 Mb (HAMILTON ET AL. 2016). Les chromosomes MIC sont coupés en 181 chromosomes MAC d'une taille comprise entre 20 kb et 3 Mb. La fragmentation des chromosomes s'effectue à des sites bien spécifiques de 15 pb, appelés Cbs (*Chromosome breaking site*) (YAO ET AL. 1990, FAN AND YAO 2000, HAMILTON ET AL. 2016). Aux abords des coupures, les extrémités sont dégradées puis télomérisées (FAN AND YAO 1996) (**Figure III.8** p.65). Pendant la différenciation macronucléaire, >30% du matériel génétique micronucléaire est éliminé. Environ 12 000 séquences, appelées IES (*Internal Eliminated Sequences*), sont excisées du futur génome MAC (HAMILTON ET AL. 2016). La très grande majorité de ces IES sont intergéniques et excisées de manière imprécise puis ligaturées. Uniquement 12 IES intragéniques ont été identifiées chez *Tetrahymena* (HAMILTON ET AL. 2016, CHENG ET AL. 2016, FENG ET AL. 2017). D'autre part, le MAC de *Tetrahymena thermophila* passe d'un état diploïde à une ploïdie ∼45n, pour une complexité de génome haploïde de 103 Mb (EISEN ET AL. 2006, COYNE ET AL. 2008) (**Table III.1** p.64).

Oxytricha trifallax est un autre organisme modèle cilié très distant phylogénétiquement de *Paramecium* ou de *Tetrahymena* (**Figure III.3** p.55 et **III.4** p.55). Cette espèce est cependant très intéressante. Le génome MIC a une complexité estimée d'au moins 500 Mb (CHEN ET AL. 2014). Les chromosomes micronucléaires sont fragmentés en ∼16 000 "nanochromosomes" somatiques de ploïdie variable (∼2000n), puis télomérisés (NOWACKI ET AL. 2010, SWART ET AL. 2013). Avec une taille moyenne de 3.2kb, les nanochromosomes

ne contiennent en général qu'un seul gène. Plus de 90% de l'ADN MIC est éliminé pendant la différenciation macronucléaire. On estime que ∼150 000 séquences (IES), bornées par de courtes répétitions appelées pointeurs, sont éliminées précisément (Chen et al. 2014). De petits ARN produits à partir de transcrits maternels ciblent les régions à protéger de l'élimination (Fang et al. 2012, Zahler et al. 2012). Nous verrons dans la **section III.3.2.3** (p.75) que les réarrangements programmés de génome de la paramécie impliquent également des petits ARN, mais ils ont un rôle de ciblage des régions à éliminer et non pas protecteur. En plus de cette élimination d'ADN, pour reconstituer ses nanochromosomes somatiques et reformer des gènes fonctionnels, *Oxytricha* utilise de longs transcrits maternels pour réordonner et réorienter les fragments de gènes portés par les chromosomes MIC (**Figure III.9** p.66) (Nowacki et al. 2008).

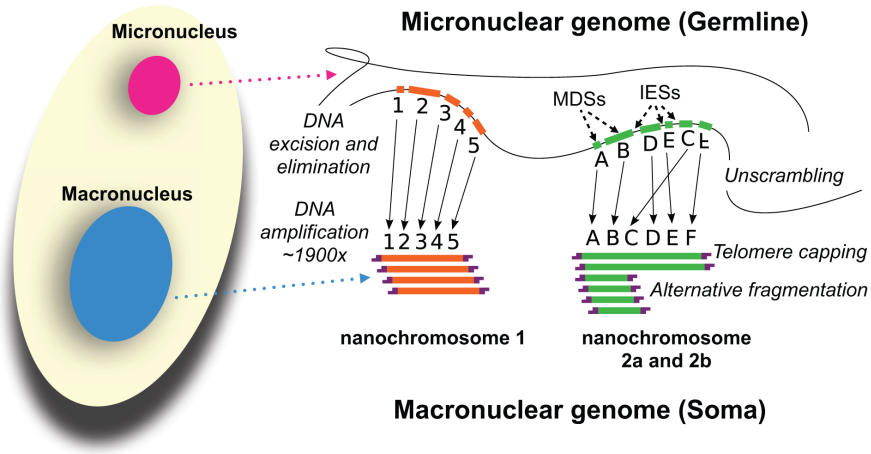


Figure III.9 – *Développement du génome macronucléaire chez Oxytricha*

Pendant la conjugaison des cellules d'*Oxytricha*, des segments du génome micronucléaire sont excisés et religués ensemble pour former les nanochromosomes du nouveau génome MAC (ou MDS pour *Macronucleus Destined Sequence*). Le reste du génome micronucléaire est éliminé. Les segments religués peuvent être soit en ordre (formant par exemple le nanochromosome 1, à gauche), soit en ordre ou inversés (formant par exemple les deux formes du nanochromosome 2), auquel cas ils doivent être ré-ordonnés. Une fragmentation alternative de l'ADN, pendant le développement MAC, produit des isoformes de nanochromosomes de tailles différentes (2a et 2b). Les nanochromosomes matures sont coiffés aux deux extrémités par des télomères. Figure de Swart et al. (2013)

Paramecium tetraurelia possède dans son cytoplasme deux micronoyaux (∼3µm) diploïdes et un macronoyau (∼30µm) polyploïde (∼800n) (Berger 1973). La complexité du génome MIC est estimée à ∼100 Mb (Guérin et al. 2017). Le nombre de chromosomes MIC n'est pas parfaitement connu, mais un marquage cytologique a détecté ∼60 paires de chromosomes germinaux (Jones 1956, Betermier and Duharcourt 2014). Contrairement à *Tetrahymena*, la fragmentation des chromosomes MIC de la paramécie ne s'effectue

pas à des sites spécifiques tels que les Cbs. Les points de cassures sont imprécis et localisés dans des zones particulières des chromosomes MIC contenant des minisatellites ou transposons. Cependant, aucun motif n'a été détecté. Le Mouël et al. (2003) ont observé chez *Paramecium primaurelia* une élimination variable d'ADN aux abords des coupures pouvant aller jusqu'à ∼20kb. Les extrémités peuvent être télomérisées ou religuées (**Figure III.8** p.65). On considère que *Paramecium tetraurelia* possède ∼120 chromosomes MAC (dénués de séquences répétées et sans centromères) pour une complexité haploïde de 72 Mb (Aury et al. 2006). La résolution variable des cassures chromosomiques entraine la présence de plusieurs versions différentiellement réarrangées d'un même chromosome MAC au sein d'une même cellule. L'assemblage du génome MAC est donc une version consensus et théorique de la diversité des molécules que l'on pourrait observer dans un macronoyau. Autrement dit, un chromosome MAC n'a pas de réelle existence. Environ 28% du génome MIC est éliminé pendant la différentiation macronucléaire. Parmi ces ∼28%, ∼3% de séquences, appelées IES, sont excisées du génome de manière précise. Contrairement à *Tetrahymena*, les 45 000 IES de paramécie sont en copie unique et bornées invariablement par deux di-nucléotides *TA* à chaque extrémité (Bétermier et al. 2000, Arnaiz et al. 2012). Dans la **section III.3.2** (p.68), je donnerai plus d'éléments sur les caractéristiques des IES de paramécie.

III.3.2 Élimination des séquences micronucléaires

Comme pour tous les ciliés, le bon développement du nouveau MAC est critique pour la survie de la paramécie. Nous avons vu que ce développement macronucléaire implique des réarrangements programmés de génome et une amplification de l'ADN de 2n à environ 800n (8 à 10 cycles de réplication Berger (1973)). Les réarrangements de génome chez la paramécie correspondent à une perte de matériel génétique (**Table III.1** p.64) que l'on peut séparer en deux catégories : une élimination hétérogène de séquences (∼25 Mb) plus ou moins répétées (transposons et satellites) et une excision précise de petites séquences (IES pour *Internal Eliminated Sequences*) représentant une complexité de ∼3.5 Mb (**Figure III.6** p.59). Ces deux processus complexes ont lieu dans une fenêtre de temps de quelques heures (Bétermier 2004). Les IES pouvant être intragéniques, le nouveau MAC doit être correctement réarrangé pour être fonctionnel. Pendant le développement macronucléaire, les fragments de l'ancien MAC (en dégradation) assurent la transcription génique nécessaire à la vie cellulaire. Le développement du MAC commence par 3 à 4 cycles de réplication discrets. La suite de l'amplification d'ADN sera continue et se poursuivra après la division caryonidale (Berger 1973). L'excision des IES commence après les premiers cycles de réplication et se terminera avant la division caryonidale (Gratias and Bétermier 2001). D'après des données préliminaires d'A. Le Mouël et A. Gratias, l'élimination hétérogène des séquences répétées s'amorcerait après les premières excisions d'IES.

III.3.2.1 Les *Internal Eliminated Sequences*

L'ensemble des éléments, que je vais introduire dans ces paragraphes, seront développés dans la **section V.1.1** des résultats (p.105) et dans l'article associé (Arnaiz et al. 2012).

Caractérisation des IES Les IES de paramécie sont des segments d'ADN éliminés précisément pendant le développement du génome MAC. A l'exception de l'IES du gène mtA (voir **section III.2.3.1** p.61 et la **discussion VII.2.2.3** p.178), les IES sont non codantes et globalement uniques dans le génome. Dans les années 1990-2000, seule une poignée d'IES avait pu être identifiée, majoritairement dans des gènes codant pour des antigènes de surface (Steele et al. 1994, Preer et al. 1992). Il a fallu attendre le séquençage du génome MAC et l'identification de la transposase domestiquée PiggyMac (Pgm), de la famille des transposases *piggyBac* (voir **section II.2.1.2** p.44), nécessaire aux réarrangements, pour identifier les 45 000 IES du génome de *Paramecium tetraurelia* (Aury et al. 2006, Baudry et al. 2009, Arnaiz et al. 2012). Les IES sont invariablement bornées par deux TA requis pour une excision correcte (Mayer and Forney 1999, Ruiz et al. 2000, Gratias et al. 2008). Les IES que nous avons identifiées sont de courtes séquences (entre 26pb et 5.5kb dont 93% avec une taille inférieure à 150 paires de bases) (**Figure III.10A** p.69) ne représentant que 3.5 Mb de complexité sur les 28 Mb éliminées pendant la maturation du

MAC (Guérin et al. 2017). La distribution des tailles d'IES est remarquable. Elle montre une périodicité de 10.2pb correspondant à un tour d'hélice de l'ADN double brin. Nous avons postulé que cette contrainte de taille traduirait une pression mécanistique pour une excision optimale des IES (Bischerour et al. 2018). En effet, le rapprochement des deux bornes de l'IES autoriserait l'introduction des cassures double brin, ensuite les deux TA résiduels s'apparieraient pour permettre la réparation (voir paragraphe suivant sur la *Mécanismes moléculaires d'excision des IES*). Donc deux TA spatialement proches minimiseraient l'énergie de torsion de l'ADN qui favoriserait la réaction (Arnaiz et al. 2012).

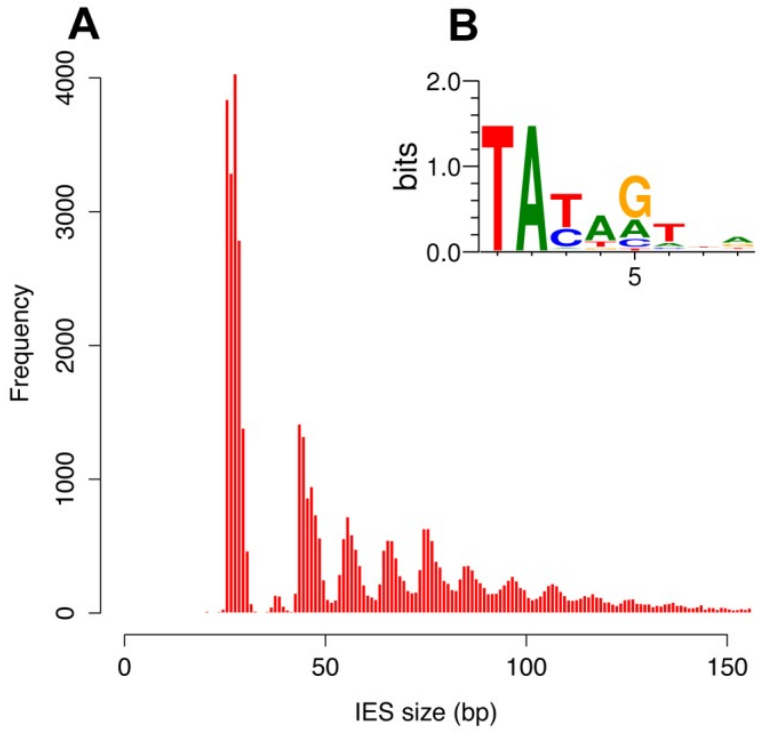


Figure III.10 – *Propriétés de séquence des IES de P. tetraurelia*

A. Histogramme de taille des IES **B.** Consensus aux bornes des IES. Figure tirée de Arnaiz et al. (2012)

Origine des IES Le consensus dégénéré de 6pb (TAYAGT) aux bornes des IES (**Figure III.10B** p.69) suggère qu'elles dériveraient de transposons de la famille Tc1/*IS630 mariner* (Klobutcher and Herrick 1995). Klobutcher and Herrick (1997) ont également proposé un scénario (*Invasion, Bloom, Adbicate and Fade*) expliquant l'origine des IES dans les génomes des ciliés. Dans ce modèle, revu par Dubois et al. (2012), un transposon PiggyBac aurait envahi le génome MIC d'un ancêtre des ciliés (antérieur à la divergence *Tetrahymena* et *Paramecium*). Le gène de la transposase aurait été domestiqué (Pgm). Le génome germinal de la paramécie aurait subi une seconde invasion par un ET Tc1/*mariner*. Grâce à la transposase domestiquée, ces séquences parasites peuvent être excisées précisément du génome somatique, sans impact sur la vie de la cellule. Sans pression de sélection, l'ET a pu se répandre, *via* sa transposase, pendant une longue période de temps, et dans tout

le génome. Dans un deuxième temps évolutif, les copies Tc1/*mariner* auraient perdu, par délétion, leur autonomie de transposition. Les IES observées aujourd'hui seraient le stade avancé de la dégradation des ET. L'hypothèse de l'origine transposon des IES a été confortée par la reconstruction d'un consensus de transposons Tc1/*mariner* (*Anchois*) à partir des rares longues séquences d'IES (Arnaiz et al. 2012). Une autre caractéristique de la distribution de taille des IES est une accumulation d'IES vers des tailles comprises entre 26 et 30 pb. En accord avec l'hypothèse de Klobutcher and Herrick (1997), cette observation peut être expliquée par la dégradation progressive des séquences, tout en maintenant les TA aux bornes essentiels à l'excision, jusqu'à une taille limite de 26pb. Par des analyses de génomique comparative, nous avons mis en évidence un lien entre la taille des IES et leurs "âges" d'insertion. En effet, les IES les plus grandes ont tendance à être les plus récemment insérées dans le génome (Arnaiz et al. 2012).

Mécanismes moléculaires d'excision des IES La première étape d'excision d'une IES est une cassure double brin (CDB) de l'ADN à chaque extrémité de la séquence (Gratias and Bétermier 2003). Avec 45 000 IES, 90 000 CDB sont introduites (par génome haploïde) et réparées dans un intervalle de temps réduit (voir **section III.2.2** p.58). Les IES sont séparées par 1.5Kb en moyenne, donc la moitié des gènes sont interrompus par au moins une IES. Ce processus d'excision doit donc être précis et efficace. Dans le cas d'une IES dans un exon codant, on imagine bien la nécessité de l'exciser précisément afin de reconstituer une ORF fonctionnelle. En théorie, il suffirait d'une IES, non excisée, dans un gène essentiel pour conduire à un phénotype de létalité. Le mécanisme d'excision est relativement complexe et implique de nombreux facteurs (Betermier and Duharcourt 2014). Des analyses de profils d'expression (puces à ADN et ARN-seq) des gènes pendant le développement du MAC (Arnaiz et al. 2010; 2017) ont participé à l'identification de certains de ces facteurs. La **Figure III.11** (p.72) schématise la géométrie de la CDB avec des extrémités de 4 bases 5' sortantes centrées sur le di-nucléotide TA. Les CDB sont réalisées par Pgm (Baudry et al. 2009). La présence de co-facteurs, homologues de Pgm (les Pgm-like), est également requise pour l'introduction des cassures de l'ADN (Bischerour et al. 2018). Les deux extrémités sont religuées en utilisant la voie de réparation du NHEJ (*Non-Homologous End Joining*) (Kapusta et al. 2011). Pour réparer la CDB, les deux extrémités de séquences MAC s'apparient sur le TA (voir **Figure III.11** p.72). Après maturation des extrémités 5', le complexe Ligase 4 (LIG4) et XRCC4 du NHEJ religue les deux extrémités, pour reformer la jonction chromosomique (Kapusta et al. 2011). Quant aux molécules excisées correspondant aux IES, elles peuvent être circularisées (si la longueur le permet) ou se concaténériser avec d'autres segments excisés en utilisant la voie du NHEJ (Bétermier et al. 2000, Gratias and Bétermier 2001, Allen et al. 2017). De manière intéressante, Marmignon et al. (2014) ont montré un couplage étroit entre la voie de réparation du NHEJ (par l'intermédiaire de l'hétérodimère Ku70/Ku80) et le processus de clivage de l'ADN accompli par la transposase domestiquée Pgm. En effet, la présence du dimère Ku80/Ku70 est requise pour engager les CDB.

Les erreurs d'excision des IES Nous venons de voir que les IES doivent être excisées précisément et de manière très efficace. Cependant, des erreurs d'excision peuvent apparaitre à des fréquences très faibles (Duret et al. 2008, Arnaiz et al. 2012, Swart et al. 2014). Le MAC ayant une ploïdie de 800n, il est concevable qu'une IES soit mal excisée sur quelques molécules, à condition que la rétention ne soit pas délétère. Une analyse fine de données de séquençage démontre l'existence de trois types d'erreurs d'excision d'IES (aussi appelés TA-indel par précaution, car tous les TA-indels ne correspondent pas forcément à une IES). (i) Un TA-indel "résiduel" est une rétention partielle d'une IES dans le MAC. (ii) L'excision d'une IES peut se faire avec des bornes TA alternatives. (iii) Une excision cryptique de séquence MAC peut également être détectée, non reliée à une IES.

Les TA-indels sont sur-représentées dans les séquences non-codantes (40% des TA-indels sont intergéniques contre 21% des nucléotides de l'ensemble du génome). Si elles

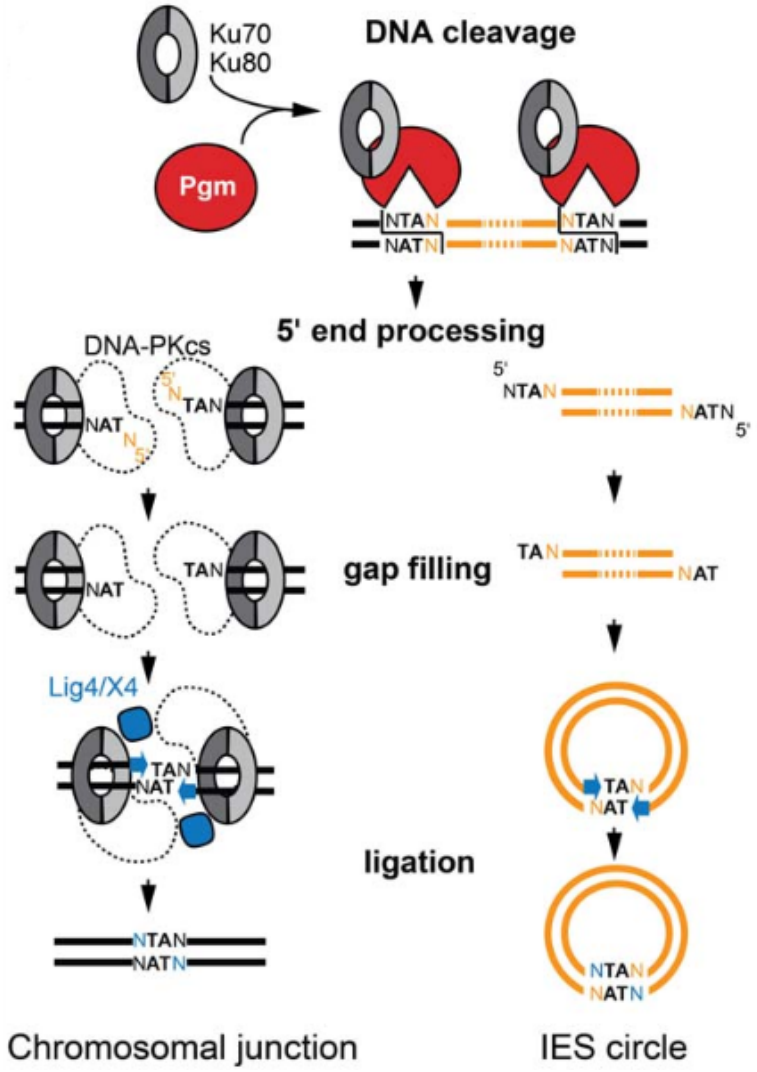


FIGURE III.11 – *Modèle d'excision des IES chez P. tetraurelia*

Mécanisme moléculaire et acteurs protéiques impliqués dans l'excision des IES. L'hétérodimère Ku70/Ku80 interagit avec Pgm pour autoriser ce dernier à introduire les CDB. Ku70/Ku80 protégerait les extémités et recruterait les autres acteurs de la voie de réparation du NHEJ, comme DNA-PKcs. Après élimination du nucléotide en 5' et addition d'un nucléotide complémentaire, le complexe Lig4/Xrcc4 referme la jonction MAC. A droite la séquence excisée serait circularisée par cette même voie de réparation du NHEJ. Figure tirée de BETERMIER AND DUHARCOURT (2014)

sont localisées dans les séquences codantes des gènes, la rétention de l'IES induit la formation d'un ARNm aberrant et potentiellement délétère. Comme pour les introns (Jaillon et al. 2008) (voir **section III.3.3** p.82), les IES semblent être soumises à une pression de sélection pour que l'éventuel ARNm non fonctionnel soit reconnu par les systèmes de contrôle qualité des ARNm, tel que le NMD (*None-sense Mediated Decay*, un système de dégradation des ARNm aberrants) (Maquat 2004, Saudemont et al. 2017). En effet, il existe une sous représentation de TA-indel d'une longueur multiple de 3 (3n, ne décalant pas le cadre de lecture) n'introduisant pas de codon stop prématuré en phase, et donc invisible au NMD (Duret et al. 2008).

III.3.2.2 Les séquences éliminées imprécisément

Le mécanisme d'élimination des séquences répétées est probablement analogue à celui des IES, mais reste moins bien caractérisé. En effet, l'hétérogénéité des coupures et des réarrangements compliquent les analyses. Les grandes régions à éliminer seraient principalement localisées aux bords des chromosomes MIC, isolées des zones génomiques contenant les gènes (**Figure III.12** p.75) (Guérin et al. 2017, Duharcourt and Sperling 2018). Cependant la structure réelle des chromosomes MIC est encore une question en suspens. L'élimination d'ADN est suivie par une télomérisation *de novo* des extrémités conduisant à la fragmentation des chromosomes MAC (Le Mouël et al. 2003, Duret et al. 2008). Les télomères MAC de paramécie sont constitués de 200 à 300 pb de répétitions 5'-G_4T_2-3' ou 5'-G_3T_3-3' (Baroin et al. 1987, McCormick-Graham and Romero 1996). Les sites de télomérisation ne sont pas précis et présentent une certaine variabilité. En effet, deux sites de télomérisation peuvent être séparés de plusieurs kilobases (Forney and Blackburn 1988, Caron 1992, Le Mouël et al. 2003, Amar and Dubrana 2004). Au lieu d'une télomérisation, les CDB peuvent être réparées par ligation des deux extrémités, conduisant à une délétion interne du chromosome (Le Mouël et al. 2003). L'analyse des jonctions résultantes suggère que les mécanismes impliqués seraient similaires à l'excision/réparation précise des IES (Coyne et al. 2012). Les ∼25 Mb d'ADN éliminés, pendant la maturation du MAC de *P. tetraurelia*, contiennent des ET (pour ∼2.5 Mb, 10%) et des répétitions simples en tandem (satellites) (pour ∼1.25 Mb, 5%) (Guérin et al. 2017). Dans la **section V.2** des résultats (p.135), nous verrons que les ET de *P. tetraurelia* sont majoritairement des LINE (70% des copies) et des TIR (20% des copies) (Guérin et al. 2017). Quant aux ∼21 Mb (85%) de génome MIC restant, ils ne sont pas annotés et restent au cœur de ma recherche actuelle (voir **discussion VII.2** p.172).

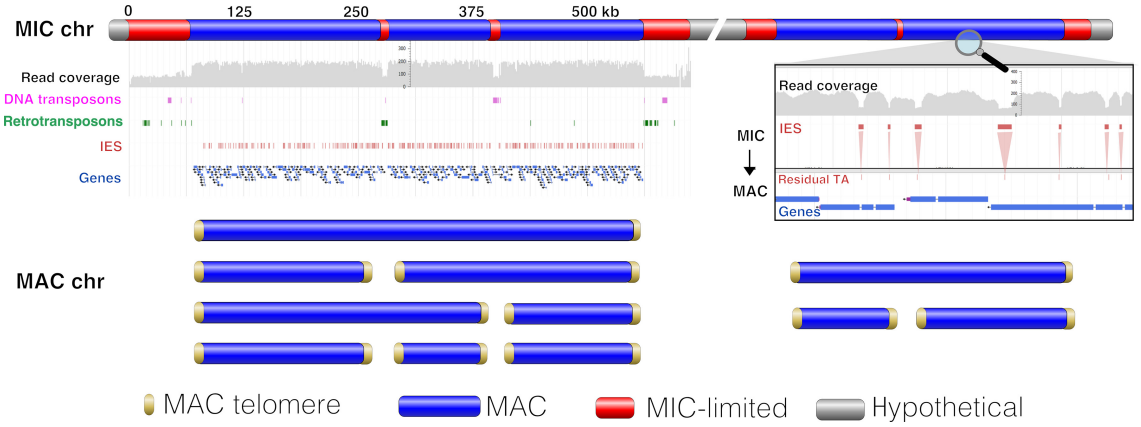


FIGURE III.12 – *Réarrangements génomiques et organisation chromosomique*

La première barre représente une vision spéculative de la structure d'un chromosome MIC de paramécie. Les séquences destinées au MAC sont indiquées en bleu et les grandes séquences à éliminer en rouge. Les zones de séquences grises sont hypothétiques car non présentes dans l'assemblage. DUHARCOURT AND SPERLING (2018) ont pris le parti de proposer un chromosome MIC à deux bras. Les pistes en dessous montrent un histogramme de couverture en lectures de séquençage MIC (200 × pour les régions destinées au MAC et 100 × pour les régions limitées au MIC), les transposons à ADN et ARN (GUÉRIN ET AL. 2017), les IES (ARNAIZ ET AL. 2012) et les gènes (ARNAIZ ET AL. 2017). Les barres bleues schématisent les différentes formes de chromosomes MAC générées par ligation ou télomérisation des bornes cassées (en jaune). Sur le côté droit, le zoom sur quelques Kb illustre bien la présence d'IES (segments rouges) dans les exons codants des gènes (segments bleus). Figure tirée de DUHARCOURT AND SPERLING (2018)

III.3.2.3 Le reconnaissance des séquences à éliminer

LA reconnaissance des séquences à éliminer est un problème d'actualité, même si des avancées notables ont été réalisées ces dernières années. Pour les ET de la superfamille TIR, la transposase utilise les répétitions inversées aux bornes de l'ET pour transposer (voir **section II.2.1.2** p.44). Chez la paramécie, l'excision des séquences implique également une protéine apparentée à une transposase (Pgm), mais aucun motif n'est clairement identifiable permettant de cibler ces séquences. Même si le di-nucléotide TA aux bornes est nécessaire à l'excision des IES, le consensus dégénéré de 6pb (TAYAGT, voir **Figure III.10B** p.69) n'est pas un signal suffisant pour une reconnaissance précise, en particulier dans un génome riche en A et T (voir **section III.3.1** p.64).

Une partie de la réponse est liée à de l'homologie de séquence et des mécanismes épigénétiques impliquant de petits ARN non codants. Des études ont montré que le contenu en ADN du MAC maternel influe sur les réarrangements du nouveau MAC. L'étude d'un mutant (d48) a montré qu'un gène, codant pour un antigène de surface A, est éliminé de manière reproductible du génome MAC, alors qu'il est intact dans le génome MIC (EPSTEIN AND FORNEY 1984, FORNEY AND BLACKBURN 1988). Ne suivant pas les lois mendéliennes, cette délétion du gène est héritée maternellement. L'injection du gène A dans le MAC de l'épi-mutant d48 restaure le phénotype sauvage à la génération suivante (KOIZUMI AND KOBAYASHI 1989). La présence de copies du gène dans le MAC maternel

entraine la non-élimination du *locus* correspondant (Meyer 1992). De manière similaire, l'introduction d'une séquence d'IES dans le MAC maternel peut conduire à la rétention de cette IES spécifiquement à la génération suivante (Duharcourt et al. 1995; 1998). L'effet sera d'autant plus fort que le nombre de copies injectées est élevé (Nowacki et al. 2005). Cependant, seul un tiers des IES testées (sur une quinzaine d'IES) possède ce genre de comportement suggérant que toutes les IES n'ont pas les mêmes caractéristiques (Duharcourt et al. 1998). Ces IES, dont l'excision est contrôlée maternellement, sont désignées comme *maternally controlled IES* (mcIES).

Modèle de scanning Le modèle du "*scanning*" a été proposé pour expliquer le ciblage des séquences à éliminer. Mochizuki et al. (2002) ont énoncé ce modèle pour *Tetrahymena thermophila*, puis il a été extrapolé à la paramécie (Lepere et al. 2008, Lepère et al. 2009, Coyne et al. 2012, Betermier and Duharcourt 2014). Ce modèle met en jeu plusieurs classes d'ARN non codants : des ARN non codants générés de l'ancien MAC avec un rôle protecteur, et des ARN produits à partir de la lignée germinale favorisant l'élimination de séquences. Une comparaison génomique entre ces deux populations d'ARN permettrait de déterminer les régions à déléter. Autrement dit, seules les séquences présentes dans le MAC maternel seront conservées et toutes les séquences présentes uniquement dans le MIC seront éliminées. La **figure III.13** (78), tirée de Betermier and Duharcourt (2014), schématise le modèle.

Au stade végétatif, le génome MAC somatique est transcrit à faible niveau, de manière constitutive et dans les deux sens (Lepere et al. 2008). Ces transcrits transmettent l'information du génome parental réarrangé. A la méiose, des transcrits précurseurs double brin sont produits à partir du génome germinal (version non réarrangée du génome) impliquant notamment le facteur Spt5 (régulateur de transcription ; Spt5 a probablement un autre rôle que la production de ces transcrits naissants) (Gruchota et al. 2017). Ces longs transcrits sont coupés en petits ARN de 25nt 5'UNG (2 nt sortant en 3') appelés scan ARN (scnARN) par les protéines Dicer-like 2 (DCL2) et Dicer-like 3 (DCL3) (Lepère et al. 2009, Sandoval et al. 2014). Dcl2 a une préférence de clivage pour des petits ARN de 25nt alors que Dcl3 préfère générer des petits ARN portant la signature 5'UNG (Sandoval et al. 2014, Hoehener et al. 2018). Même si les scnARN ressemblent aux piRNA des métazoaires (associés à des protéines Piwi et produits à partir de la lignée germinale), ces petits ARN couvrent l'ensemble du génome MIC (Lepère et al. 2009, Singh et al. 2014). Les scnARN sont chargés par des protéines de la famille *Argonaute*, Ptiwi01 et Ptiwi09 (Bouhouche et al. 2011) et transportés dans les fragments de l'ancien MAC. Par homologie de séquence, les transcrits maternels sont comparés au catalogue de scnARN. Ce processus implique notamment une hélicase PTMB.220 (Nowak et al. 2011) et les protéines Nowa1 et Nowa 2 (Nowacki et al. 2005). Bien que les modalités soient inconnues, le résultat de cette comparaison (*scanning*) est la sélection des scnARN ne pouvant pas s'apparier sur des transcrits du MAC maternel et donc spécifiques des régions à éliminer. La population de scnARN ayant pu s'apparier tend à disparaitre par dégradation, modification ou séquestration. Les scnARN sélectionnés, chargés sur les protéines Ptiwi1/9 Nowa1/2, sont transportés dans l'ébauche en développement. Pour guider l'endonucléase Pgm, l'homologie de séquence entre scnARN et séquences à éliminer, impliquant la protéine TfIIs4 (homologue du facteur d'élongation TFIIS) (Maliszewska-Olejniczak et al. 2015), conduirait au dépôt de marques épigénétiques H3K9me3 et H3K27me3 par Ezl1 (Taverna et al. 2002, Liu et al. 2007, Lhuillier-Akakpo et al. 2014, Frapporti et al. 2019) (voir **section I.2.1** p.8). Le système serait également amplifié par la production d'une autre classe de petits ARN. A partir des séquences IES excisées, la protéine Dicer-like 5 (DCL5) générerait des iesARN de 25 à 30nt de long (Sandoval et al. 2014), chargés par

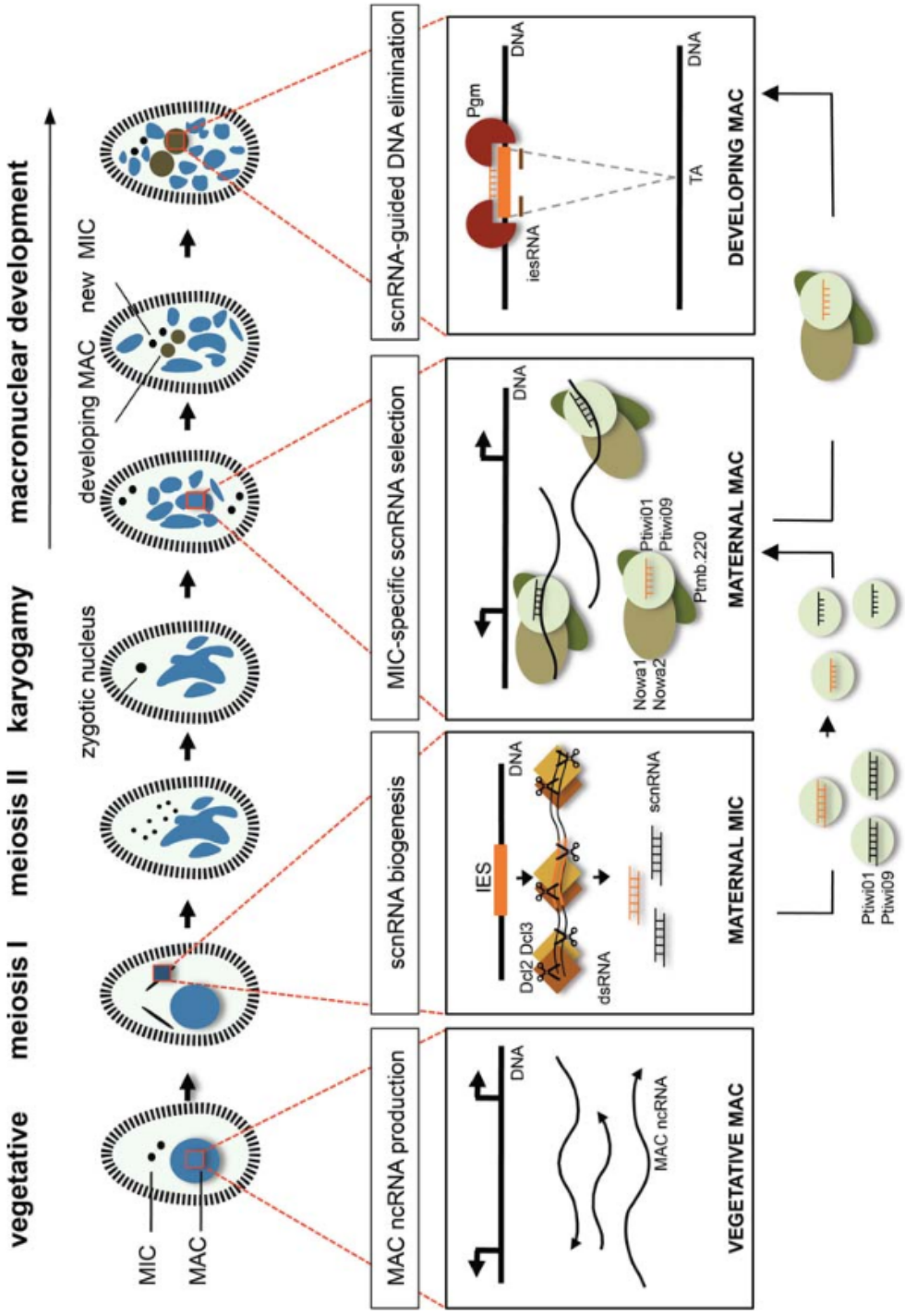


FIGURE III.13 – *Modèle de scanning chez P. tetraurelia*

La description de cette figure est donnée dans le texte. Figure tirée de BETERMIER AND DUHARCOURT (2014)

les protéines Ptiwi10 et Ptiwi11 (Furrer et al. 2017). Avec une signature 5'UAG, les ieSARN correspondent préférentiellement aux extrémités des IES avec un consensus aux bornes TAYAG (voir **section III.3.2.1** p.68).

Un modèle imparfait Le modèle du *scanning* a été imaginé pour expliquer l'hérédité maternelle des réarrangements de génome, en particulier des régions éliminées imprécisément et de l'excision précise des IES contrôlées maternellement (mcIES). Cependant, d'après des études moléculaires, seul un tiers des IES sont des mcIES, laissant une majorité d'IES incompatible avec le modèle du *scanning*. La déplétion par ARN interférence de l'endonucléase Pgm empêche l'excision de toutes les IES, et le séquençage de cet ADN a permis d'identifier les 45000 IES du génome de *Paramecium tetraurelia* (Arnaiz et al. 2012). En suivant la même stratégie, le séquençage d'ADN de cellules déplétées en divers facteurs a été réalisé ces dernières années (voir **Table III.2** p.80). Grâce au logiciel ParTIES (Denby Wilkes et al. 2016) (voir **section V.1.2** des résultats p.127), nous avons pu classifier les IES en fonction de la dépendance à tel ou tel facteur.

La **Table III.2** (p.80) et la **Figure III.14A** (p.81) montrent qu'il existe trois grandes catégories d'IES. Tout d'abord, les mcIES dont le comportement est explicable par le modèle du *scanning*. En effet, l'inactivation des facteurs DCL2/3/5 (co-inactivés) (Sandoval et al. 2014, Swart et al. 2014; 2017), NOWA1/2 (Nowacki et al. 2005, Swart et al. 2017) ou TFIIS4 (Maliszewska-Olejniczak et al. 2015), perturbant la genèse des transcrits maternels, la production/transport des scnARN ou la comparaison des séquences homologues (scnARN et ARN correspondant aux régions à éliminer), conduit à la rétention significative d'environ 50% des IES (en vert sur la **Figure III.14**). L'inactivation du facteur Ezl1, l'histone méthyltransférase qui dépose les marques H3K9me3 et H3K27me3, entraine la rétention de 70% des IES (Lhuillier-Akakpo et al. 2014, Frapporti et al. 2019). L'ensemble des mcIES sont des IES Ezl1-dépendantes. Il existe donc 20% des IES qui dépendent uniquement du marquage épigénétique mais pas des petits ARN (en bleu sur la **Figure III.14**). Lhuillier-Akakpo et al. (2014) ont également montré que plus une IES est petite plus la probabilité qu'elle soit dépendante d'Ezl1 est faible (voir **Figure III.14B** p.81). Et inversement, les grandes IES (>75nt) ont une grande chance d'être des IES "épigénétique" (vert ou bleu sur la figure). Enfin, les 30% d'IES restantes (en gris) sont des IES qui ne semblent etre affectées que par l'inactivation de facteurs impliqués directement dans l'introduction des CDB ou à leur réparation par le NHEJ (Baudry et al. 2009, Kapusta et al. 2011, Arnaiz et al. 2012, Marmignon et al. 2014, Bischerour et al. 2018).

Gène	Fonction	Nombre d'IES retenues après déplétion	Proportion du génome MIC couvert	Référence
PGM	Transposase domestiquée clivant l'ADN	100%	97%	[1,2]
PGM-likes	Partenaires de Pgm	∼100%	∼97%	[3]
SPT5	Facteur d'élongation	100%	98%	[4]
EZL1	H3K27 et H3K9 histone methyltransférase	∼70% (N=∼30000)	100%	[5]
CAF1	Modification de la chromatine	58% (N=26395)	98%	[6]
DCL2/3	Dicer-like ribonucléase pour la genèse des scnARN	∼6% (N=∼3000)	98%	[5,7]
DCL5	Dicer-like ribonucléase pour la genèse des iesARN	∼5% (N=∼2300)	84%	[7]
DCL2/3/5	Dicer-like ribonucléase pour la genèse des scnARN et des iesARN	37% (N=17070)	92%	[8,9]
NOWA1/2	Protéine interagissant avec les ARN	42% (N=∼19000)	98%	[9,10]
TFIIS4	Facteur d'élongation de la transcription	47% (N=21497)	96%	[11]
-	Extraction d'ADN de MAC végétatifs	∼0%	∼80%	[2,5]

TABLE III.2 – *Table récapitulative des IES sensibles*

Le tableau rapporte le nom du (ou des) gène(s) déplété(s), la fonction putative, le nombre d'IES retenu après la déplétion et la proportion du génome MIC couvert par les données de séquençage d'un ADN de cellules déplétées pour le(s) gène(s) d'intérêt. Le MIC retenu est donné en pourcentage de fenêtres de 1kb couvertes par les lectures (voir GUÉRIN ET AL. (2017) pour la méthode) Les références bibliographiques sont indiquées dans la dernière colonne : BAUDRY ET AL. (2009) [1], ARNAIZ ET AL. (2012) [2], BISCHEROUR ET AL. (2018) [3], GRUCHOTA ET AL. (2017) [4], LHUILLIER-AKAKPO ET AL. (2014) [5], IGNARSKI ET AL. (2014) [6], SANDOVAL ET AL. (2014) [7], SWART ET AL. (2014) [8], SWART ET AL. (2017) [9], NOWACKI ET AL. (2005) [10] et MALISZEWSKA-OLEJNICZAK ET AL. (2015) [11].

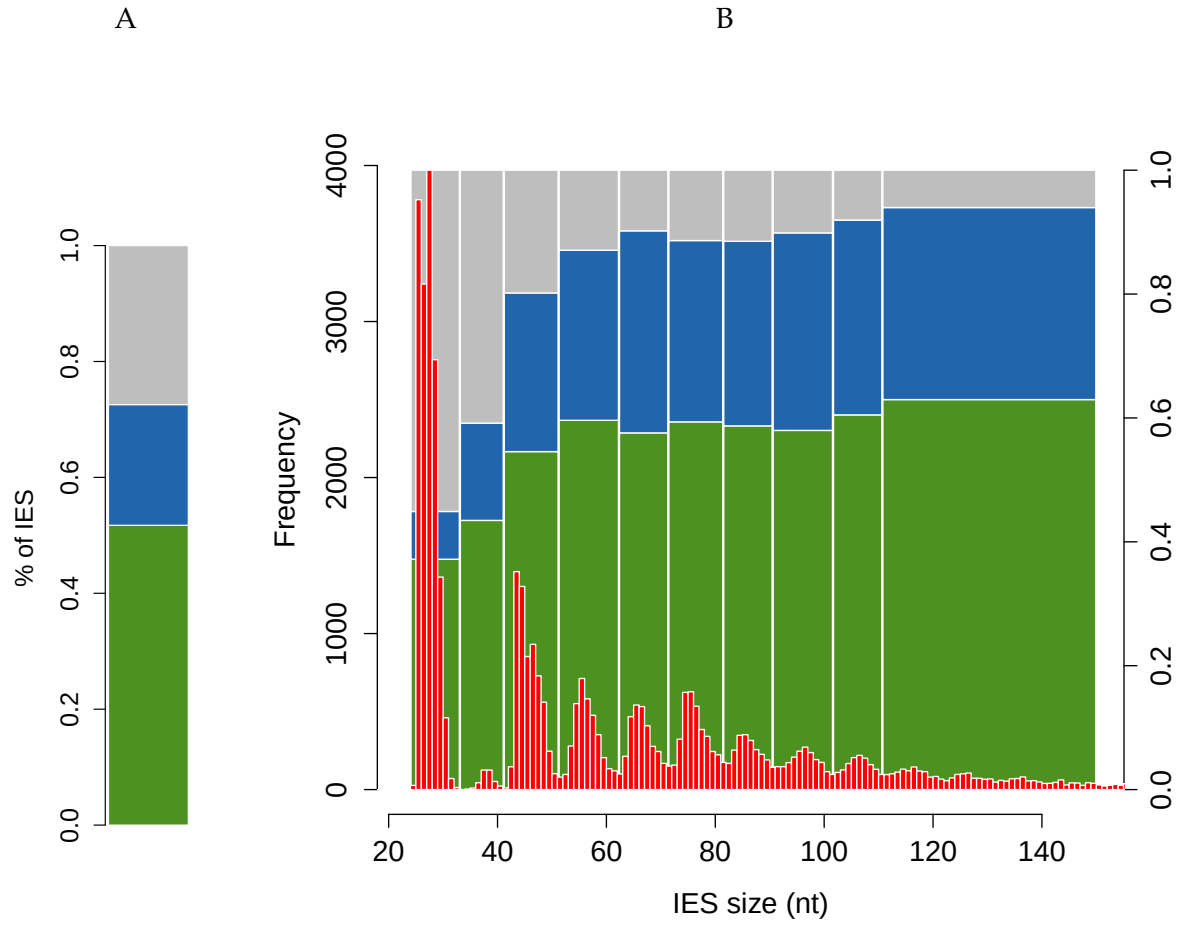


FIGURE III.14 – *Sensibilité des IES en fonction de leurs tailles*

A. Proportion des 45 000 IES de *Paramecium tetraurelia* dans les catégories : IES contrôlées maternellement (vert), les IES seulement sensibles au marquage épigénétique des histones par Ezl1 (bleu) les IES dont l'excision n'est pas explicable par le modèle du *scanning* (gris). **B :** Pour chaque pic de taille d'IES (histogramme rouge), la proportion d'IES dans chacune des trois catégories vu précédemment est indiquée en fond.

III.3.3 Les gènes codants

Les ciliés, comme nous l'avons vu, ont deux types de noyaux : un noyau germinal (MIC) et un noyau somatique (MAC). Excisé de ses IES, seul le génome MAC contient les gènes fonctionnels pouvant être exprimés. Les chromosomes MAC de *P tetraurelia* sont des molécules linéaires d'une taille comprise entre ∼150 kb et 1 Mb (voir **Table III.1** p.64). Grâce au processus d'auto-fécondation (autogamie, voir **section III.2.2** p.58)) et aidé par une ploïdie de 800n, il est aisé d'obtenir des échantillons de grande quantité d'ADN provenant de noyaux somatique 100% homozygote. En 2004, un premier projet de séquençage et d'annotation du plus grand chromosome MAC (de 1 Mb appelé le "Mégabase") de *P tetraurelia* a permis de donner de premières caractéristiques génomiques (Zagulski et al. 2004). Le génome de paramécie est très riche en nucléotides A et T (71%). Avec ses 460 gènes annotés manuellement sur 1 Mb, ce génome est très dense en gènes (∼74% codant). Ce succès a encouragé les premières collaborations avec le Génoscope (CNS). En 2003, un projet de grande envergure (pour l'époque) est lancé pour le séquençage et l'annotation de tout le génome MAC de *P tetraurelia* (Aury et al. 2006). Une banque plasmidique d'inserts a été séquencée par la méthode *Sanger*. Les 13X de lectures ont généré un assemblage *Arachne* (Batzoglou et al. 2002, Jaffe et al. 2003) de 72 Mb (413 kb de N50) avec 96% de l'assemblage dans les 188 *scaffolds* les plus grands (> 45kb). Des répétitions télomériques MAC sont retrouvées aux extrémités d'une grande partie des 188 scaffolds démontrant l'aspect complet de l'assemblage. Le génome est, en effet, riche en AT (28% GC seulement). Comme chez l'homme on observe une sous-représentation du dinucléotide CpG (humain 1% contre 4% en théorie ; paramécie 0.82% contre 1.9% en théorie). Cette observation peut être expliquée par une désamination spontanée de la 5-méthylcytosine en thymine tendant à convertir les m^5CpG en TpG (Scarano et al. 1967). Cependant, et contrairement à *Oxytricha* (Bracht et al. 2012), aucune trace de 5 méthylcytosine n'a été détectée jusqu'à présent chez la paramécie (Cummings et al. 1974, Wang et al. 2017, Singh et al. 2018).

En combinant, l'annotation du *Mégabase*, 90000 EST (*Expressed Sequence Tag*) et une combinaison de méthodes *ab initio* et de génomique comparative, 39642 gènes codants ont été prédits (Aury et al. 2006). Aucune évidence d'épissage alternatif n'a été identifiée chez la paramécie. La **Table III.3** (p.83) décrit les caractéristiques de génome et d'annotation pour deux ciliés (*Paramecium tetraurelia* et *Tetrahymena thermophila*), un autre alvéolé dont le génome est riche en AT (l'apicomplexe *Plasmodium falciparum*) et des données sur l'homme (*Homo sapiens*) (Aury et al. 2006, Arnaiz et al. 2017, Eisen et al. 2006, Coyne et al. 2008). Les résultats du *Mégabase* ont été confirmés à l'échelle du génome entier. Le génome de *Paramecium tetraurelia* est très riche en gènes (75% codant avec une distance intergénique moyenne de ∼350 nucléotides).

Les gènes de paramécie contiennent souvent au moins un intron (voir **Table III.3** p.83). Les minuscules introns de paramécie ont une taille comprise entre 20 et 30 nucléotides. Les quelques grands introns (∼90 nt) contiennent des ARN non codant de type snoARN

	P tetraurelia	*T thermophila*	*P falciparum*	*H sapiens*
Taille du génome	72 Mb	103 Mb	23 Mb	3,2 Gb
Contenu GC	28%	22%	19%	41%
Nombre de gènes	40460	26996	5460	22287
Pourcentage codant	75%	50%	53%	1%
Taille des gènes (pb)	1409	2400	2447	∼20 kb
Taille des protéines (aa)	443	627	752	509
Taille des exons (pb)	411	420	868	∼300
Taille des introns (pb)	25	140	167	∼3300
Introns par gène	2,9	5,12	2,6	8,1
Nombre de gènes avec intron(s)	80%	70%	54%	85%
Taille des 5'UTR (pb)	22	95	NA	150
Taille des 3'UTR (pb)	35	230	NA	520

TABLE III.3 – *Caractéristiques des gènes*

Le tableau est commenté dans le texte. Modifié de AURY ET AL. (2006) et ARNAIZ ET AL. (2017)

(CHEN ET AL. 2009). Sauf exception, ils commencent tous par GpT et se finissent pat ApG (**Figure III.15** p.84). Ce consensus, leurs tailles très contraintes et leurs taux de GC (∼21%) les rendent facilement reconnaissables dans une séquence génomique. De plus, les introns sont sous une forte pression sélective pour être visibles par le système de reconnaissance et dégradation des ARNm erronées (NMD), en cas de rétention. En effet, on observe une sous représentation d'introns d'une taille multiple de 3 (3n) ne changeant pas la phase de lecture de traduction, ou d'introns 3n n'introduisant pas de codon stop en phase en cas de rétention (JAILLON ET AL. 2008, SAUDEMONT ET AL. 2017). Le NMD est très efficace et compense les erreurs éventuelles d'épissage pouvant être délétères pour la cellule.

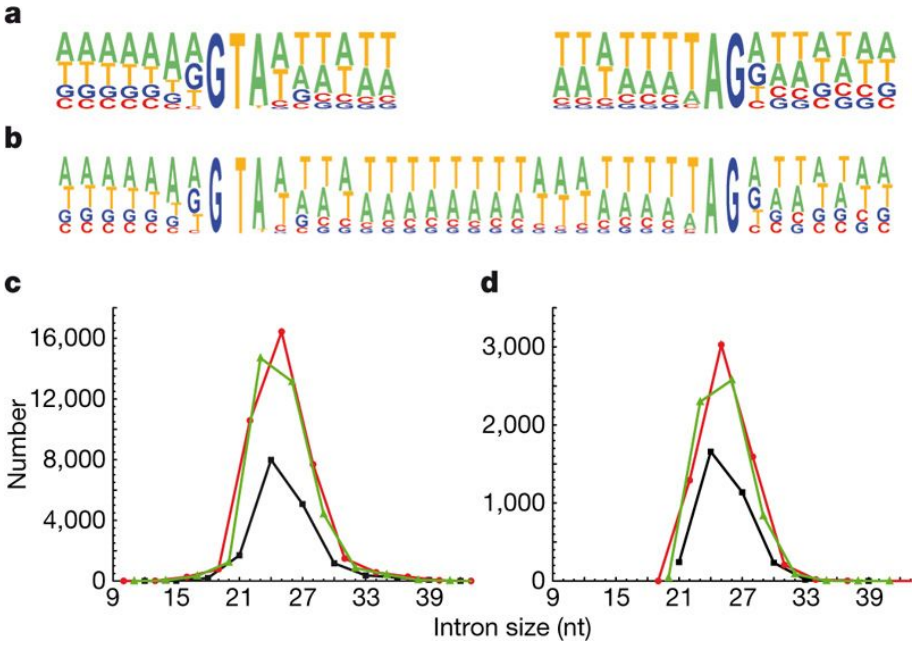


FIGURE III.15 – *Les introns chez P. tetraurelia*

a, Profils de composition des sites d'épissage 5 '(gauche) et 3' (droite), incluant sept nucléotides à l'extérieur et neuf nucléotides à l'intérieur de l'intron (n = 15 286 introns confirmés par EST). b. Profil de composition de toute la longueur des introns de 25 nt (la classe de taille la plus abondante), avec sept nucléotides des exons flanquants des deux côtés (n = 3 028 introns confirmés par EST). c, Distribution en taille des 90 282 introns annotés. Les introns 3n, 3n + 1 et 3n + 2 sont représentés en noir, rouge et vert, respectivement. d, Distribution de la taille des 15 286 introns confirmés par une EST. Figure tirée de JAILLON ET AL. (2008)

III.3.4 Duplication globale de génome

Avec ses 40 000 gènes, le génome MAC de *Paramecium tetraurelia* de 72 Mb est très riche en gènes. Ce nombre de gènes impressionnant peut s'expliquer par un pourcentage codant important (75%) et par des duplications globales de génome (WGD *Whole Genome Duplication*) (Aury et al. 2006). En effet, la moitié des gènes sont en deux copies. Après une WGD, on observe une perte massive de gènes pour revenir à l'état pré-duplication. La perte de gènes ne se fait pas au hasard. En effet, les gènes impliqués dans un même complexe ont tendance à avoir un profil de perte similaire afin de conserver une certaine stœchiométrie entre les facteurs (Aury et al. 2006). De plus, les gènes très exprimés ont tendance à être retenu préférentiellement. La perte d'un gène très exprimé aura un fort impact sur la cellule (Gout et al. 2010). Le génome de la paramécie n'a subi que peu de réarrangements chromosomiques, ce qui conduit à une conservation de la synténie entre chromosomes issus d'une WGD. Cette conservation a facilité la reconstruction successive des chromosomes ancestraux pré-WGD permettant d'identifier une série d'au moins trois WGD (appelées "récente", "intermédiaire" et "vieille" WGD) (voir **Figure III.16** p.86). Grâce à une évolution lente de génome ou une WGD arrivée relativement récemment dans l'histoire de la paramécie, la dernière duplication est encore très visible au niveau nucléotidique. Près de la moitié des gènes ont toujours une copie paralogue de la WGD la plus récente. Le pourcentage d'identité protéique moyen entre paralogues des WGD récente, intermédiaire et vieille est de 86%, 66% et ∼54% respectivement. Pour comparer, 82% correspond à la divergence homme-souris (100 Ma) et 66% à la divergence homme-poisson (Aury et al. 2006). Les duplications récente et intermédiaire pré-dateraient les paramécies du groupe *aurelia* (Aury et al. 2006). L'explosion de spéciation au sein du groupe *aurelia* pourrait être due à la dernière duplication (McGrath et al. 2014b;a) (voir **Figure III.5** p.57). *Paramecium caudatum* et *Paramecium multimicronucleatum* n'ont pas subi les deux dernières duplications. Plus d'informations sur les WGD sont fournies dans les **sections IV** (p.89) des résultats et dans la **discussion VI.2** (p.165) .

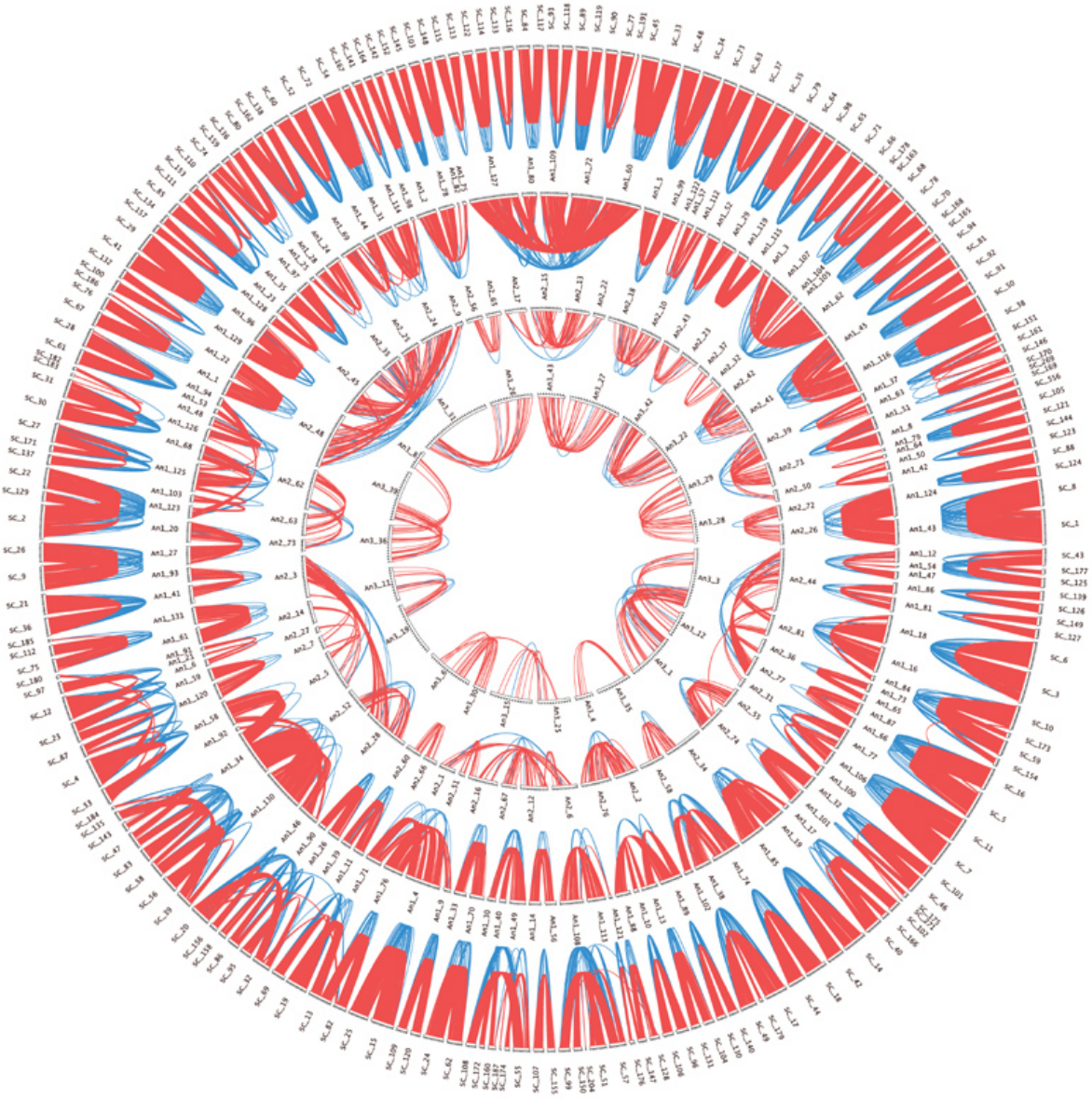


FIGURE III.16 – *Représentation des WGD successives du génome de P. tetraurelia*

Le cercle extérieur représente les scaffolds du génome MAC de *Paramecium tetraurelia* et la WGD plus récente. AURY ET AL. (2006) ont élaboré une procédure automatique pour identifier les gènes paralogues de WGD, basée sur des alignements de séquences protéiques et une analyse de la synténie chromosomique. Les arcs rouges montrent les liaisons entre les gènes avec une correspondance de meilleur match réciproque (BRH pour *Best Reciprocal Hit*), alors que les arcs bleus relient les gènes sans correspondance BRH mais avec une cohérence synténique. La combinaison des séquences appariées permet la reconstitution du génome ancestral pré-WGD. La même procédure de détection des paralogues WGD est appliquée sur cette nouvelle référence, montrant l'identification d'une autre WGD (la WGD "intermédiaire"). Répétée une nouvelle fois, la procédure montre une troisième "vieille" WGD. Le quatrième cercle pourrait montrer une quatrième WGD mais l'effectif n'est pas suffisant pour l'affirmer. Figure tirée de AURY ET AL. (2006).

Deuxième partie

Résultats

Chapitre IV

Annotation du génome macronucléaire

Un angle de réflexion de ce manuscrit porte sur l'impact qu'ont eu les technologies NGS sur l'annotation des génomes, et plus particulièrement sur les génomes des paramécies. L'article qui va suivre témoigne des bénéfices apportés par le séquençage de molécules d'ARN pour le processus d'annotation des gènes de paramécie (Arnaiz et al. 2017).

Avant tout, voici quelques éléments de contexte dans lequel cet article est paru. Le génome macronucléaire (MAC) de *Paramecium tetraurelia* fut le premier à être séquencé et annoté (Aury et al. 2006). Ce génome a notamment révélé au moins trois duplications globales de génome (WGD) successives (voir **section III.3.4** p.85). Les WGD sont des événements relativement courant dans l'évolution des espèces mais pour plusieurs raisons, la paramécie est un modèle très intéressant pour étudier les phénomènes biologiques après une WGD. Tout d'abord, très peu de réarrangements chromosomiques ont eu lieu depuis la dernière WGD. La **Figure IV.1** (p90) illustre la forte conservation de la synténie nucléotidique entre deux chromosomes MAC issus de la WGD la plus récente. Dans cette situation, la détection de paralogues de WGD est simplifiée. Par ailleurs, après une WGD, une des deux copies du gène dupliqué aura tendance à être perdue. Presque 90% des gènes dupliqués sont retournés à l'état de simple copie dans le génome de *Saccharomyces cerevisiae* (Kellis et al. 2004). La paramécie exhibe un génome où la moitié des 40 000 gènes sont encore en deux copies. Des analyses de données transcriptomiques par puces à ADN (Arnaiz et al. 2010) ont montré que le niveau d'expression est un déterminant de la perte et l'évolution des séquences géniques après une WGD (voir **section III.3.4** p.85 (Gout et al. 2010, Gout and Lynch 2015). Le modèle paramécie est également pertinent pour l'étude de l'évolution des espèces après une WGD. L'étude des génomes de *P. tetraurelia* (Aury et al. 2006), *P. sexaurelia*, *P. biaurelia* (McGrath et al. 2014b) et le groupe externe *P. caudatum* (McGrath et al. 2014a) a montré que la dernière WGD a précédé l'émergence du groupe d'espèces *aurelia*. Un événement brutal, tel qu'une WGD, pourrait être à l'origine de cette spéciation massive (voir **section III.1.3** p.56).

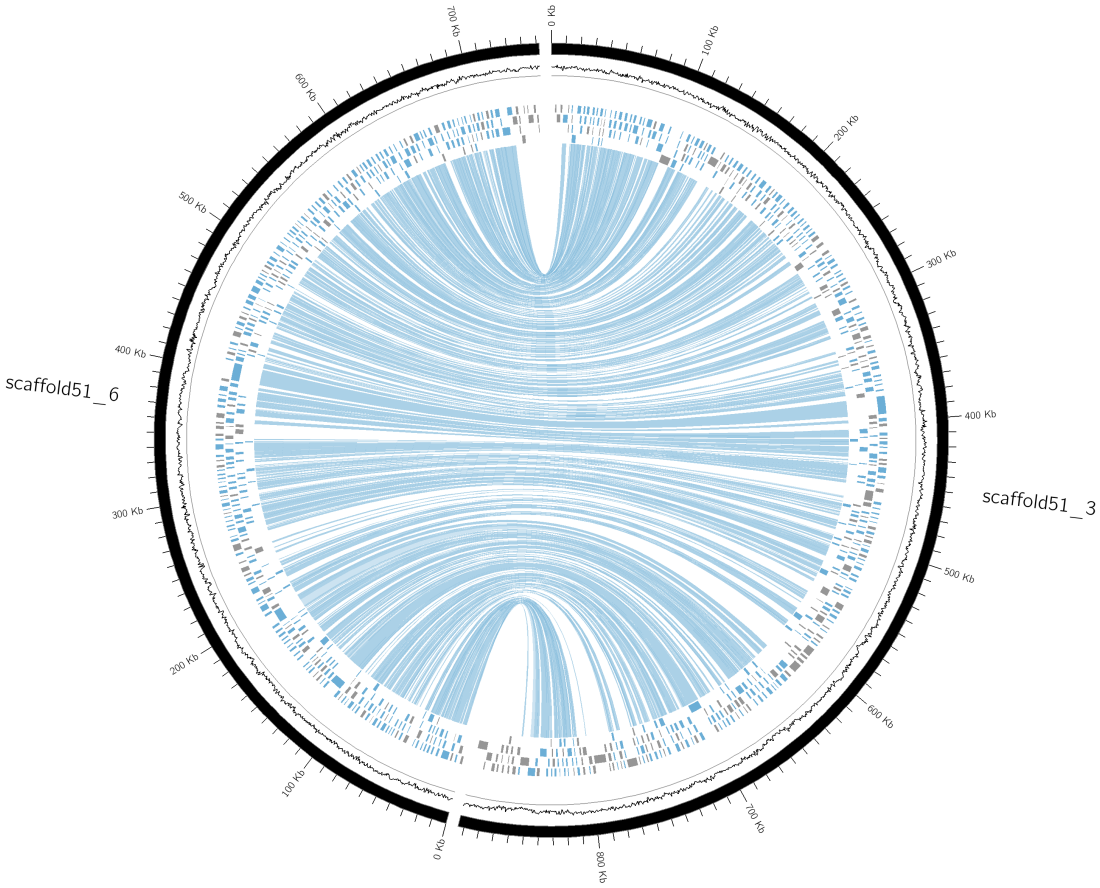


FIGURE IV.1 – *Représentation circulaire de la synténie entre 2 chromosomes de P. tetraurelia*

Les lignes épaisses noires représentent les scaffolds 3 et 6 du génome de *Paramecium tetraurelia* (v2.0). La ligne noire fine symbolise de taux de G+C. Les gènes sont affichés sous forme de rectangles de couleurs. Les gènes bleus ont un paralogue de la WGD la plus récente alors que les gènes gris n'ont pas de paralogues. Les arcs bleus représentent les liens entre deux gènes paralogues.

Toute la justesse de ces analyses est basée sur une annotation de qualité (voir **section II** p.31). Le message principal de l'article qui suit était de décrire une nouvelle chaine de procédures d'annotation de gènes, optimisée pour les génomes de paramécie, tirant profit d'évidences externes générées par du séquençage d'ARN. J'ai développé le logiciel TrUC (`https://github.com/oarnaiz/TrUC`) qui prédit la structure des unités de transcription à partir de données Cap-seq (analogue au principe du CAGE-seq) et de données ARNm-seq orientées (voir **section II.1.3.2** p.36). Ces évidences externes aident le logiciel de prédiction de gènes EuGene (v4.1) à annoter un génome de paramécie. Grâce à cette procédure, nous avons obtenu de meilleures versions de l'annotation des génomes de *P. biaurelia*, *P. caudatum*, *P. sexaurelia* et *P. tetraurelia*. Ces améliorations profiteront aux analyses génomiques futures. D'autre part, nous avons confirmé expérimentalement, pour la première fois, la position des débuts de transcription de 12 000 gènes révélant l'exceptionnelle petite taille des UTR de paramécie (21 nt en moyenne pour les 5'UTR et 44 nt en moyenne pour les 3'UTR, voir **section I.2.2.2** p.12). Par ailleurs, les ARN ont été prélevés de cultures cellulaires à différents stades des processus sexuels de *P. tetraurelia* (voir **section III.2.2** p.58). Une analyse d'expression différentielle des gènes de données ARNm-seq a permis de raffiner les résultats obtenus par puce à ADN (Arnaiz et al. 2010).

Ma contribution à cette étude : J'ai réalisé toutes les analyses et développements bioinformatiques. L'entrainement d'EuGene à reconnaitre les gènes de paramécie a été réalisé par E. Sallet et J. Gouzy. J'ai également participé à l'écriture de l'article.

Arnaiz et al. BMC Genomics (2017) 18:483
DOI 10.1186/s12864-017-3887-z

BMC Genomics

METHODOLOGY ARTICLE Open Access

Improved methods and resources for paramecium genomics: transcription units, gene annotation and gene expression

Olivier Arnaiz[1], Erwin Van Dijk[1], Mireille Bétermier[1], Maoussi Lhuillier-Akakpo[2,4], Augustin de Vanssay[2], Sandra Duharcourt[2], Erika Sallet[3], Jérôme Gouzy[3] and Linda Sperling[1*]

Abstract

Background: The 15 sibling species of the *Paramecium aurelia* cryptic species complex emerged after a whole genome duplication that occurred tens of millions of years ago. Given extensive knowledge of the genetics and epigenetics of *Paramecium* acquired over the last century, this species complex offers a uniquely powerful system to investigate the consequences of whole genome duplication in a unicellular eukaryote as well as the genetic and epigenetic mechanisms that drive speciation. High quality *Paramecium* gene models are important for research using this system. The major aim of the work reported here was to build an improved gene annotation pipeline for the *Paramecium* lineage.

Results: We generated oriented RNA-Seq transcriptome data across the sexual process of autogamy for the model species *Paramecium tetraurelia*. We determined, for the first time in a ciliate, candidate *P. tetraurelia* transcription start sites using an adapted Cap-Seq protocol. We developed TrUC, multi-threaded Perl software that in conjunction with TopHat mapping of RNA-Seq data to a reference genome, predicts transcription units for the annotation pipeline. We used EuGene software to combine annotation evidence. The high quality gene structural annotations obtained for *P. tetraurelia* were used as evidence to improve published annotations for 3 other *Paramecium* species. The RNA-Seq data were also used for differential gene expression analysis, providing a gene expression atlas that is more sensitive than the previously established microarray resource.

Conclusions: We have developed a gene annotation pipeline tailored for the compact genomes and tiny introns of *Paramecium* species. A novel component of this pipeline, TrUC, predicts transcription units using Cap-Seq and oriented RNA-Seq data. TrUC could prove useful beyond *Paramecium,* especially in the case of high gene density. Accurate predictions of 3′ and 5′ UTR will be particularly valuable for studies of gene expression (e.g. nucleosome positioning, identification of cis regulatory motifs). The *P. tetraurelia* improved transcriptome resource, gene annotations for *P. tetraurelia*, *P. biaurelia, P. sexaurelia* and *P. caudatum*, and *Paramecium*-trained EuGene configuration are available through ParameciumDB (http://paramecium.i2bc.paris-saclay.fr). TrUC software is freely distributed under a GNU GPL v3 licence (https://github.com/oarnaiz/TrUC).

Keywords: Ciliate, Cap-Seq, TSS, RNA-Seq, Gene annotation, Autogamy, Differential gene expression

* Correspondence: linda.sperling@i2bc.paris-saclay.fr
[1]Institute for Integrative Biology of the Cell (I2BC), CNRS, CEA, Univ. Paris-Sud, Université Paris-Saclay, 91198 Gif-sur-Yvette CEDEX, France
Full list of author information is available at the end of the article

Background

Ciliates are unique among unicellular eukaryotes in making a germ/soma distinction. The germline and somatic functions of chromosomes are respectively ensured by a germline micronucleus (MIC) which undergoes meiosis and fertilization and a somatic macronucleus (MAC) that contains a version of the germline genome stripped of parasitic sequences and optimized for gene expression. The MAC is lost at each sexual cycle and a new one differentiates from a copy of the zygotic nucleus, by reproducible DNA elimination under the control of meiosis-specific RNA interference pathways [1].

Genetics of the ciliate *Paramecium* was pioneered nearly a century ago [2] and this complex unicellular eukaryote has since served as model for a variety of biological processes commonly found in animals, from excitable membranes and swimming behavior [3] to programmed DNA elimination and its epigenetic control during somatic differentiation [4]. The early genetics studies led to the realization that *Paramecium aurelia* is a complex of morphologically identical but reproductively isolated sibling species, renamed *primaurelia, biaurelia, triaurelia, tetraurelia,* etc. [5]. *Paramecium tetraurelia* became the most widely used species for genetics and physiology, because of its convenient growth properties.

P. tetraurelia somatic genome sequencing revealed that the present diploid genome was shaped by a series of at least 3 whole genome duplications (WGDs), each WGD being slowly resolved by gene loss over evolutionary time [6]. It was suggested that the *P. aurelia* species complex emerged concomitantly with the most recent WGD [6], a hypothesis confirmed by sequencing two other *aurelia* genomes and *P. caudatum* as outgroup [7, 8]. Custom microarrays were designed to obtain gene expression data for the nearly 40, 000 *P. tetraurelia* protein-coding genes [9]. The data were used to show that gene expression level is a major determinant of gene dosage and protein evolution [10].

The *Paramecium aurelia* species complex is now recognized as an outstanding system to study the consequences of WGD in a unicellular eukaryote [11] and should also prove powerful for investigation of genetic and epigenetic mechanisms that drive speciation. In this context, the MAC genomes of many species are being sequenced. It thus became necessary to develop a pipeline optimized for the *Paramecium* lineage, able to make accurate gene predictions for AT-rich, compact (>80% coding) eukaryotic genomes with unusually small introns (20–30 nt). To this purpose, we generated oriented *P. tetraurelia* RNA-Seq and Cap-Seq data, as input for software we developed to predict transcription units (TrUC). The transcription unit predictions and other evidence were combined to produce gene annotations using EuGene software [12] trained for *Paramecium*. Annotations obtained for *P. tetraurelia* were used as evidence to improve the annotation of other *Paramecium* species. The *P. tetraurelia* RNA-Seq samples were also analyzed for differential gene expression during the sexual cycle of autogamy, generating an improved transcriptome resource.

Results and Discussion

Transcription units

Genome-guided transcription unit construction was pioneered by Denoeud et al. [13] (G-Mo.R-Se software) using short-read mapping. The widely used Cufflinks software [14] then adopted fragment alignment as part of its assembly strategy. To take into account alternative splicing, Cufflinks finds the minimal number of paths through the mapped fragments. We decided to develop our own transcription unit prediction software rather than use Cufflinks, because in our hands, Cufflinks did not always predict transcription units despite good fragment coverage in regions presenting strong evidence of protein-coding genes, probably because of overlapping UTRs in the very compact *Paramecium* genome. Our approach is conceptually similar to G-Mo.R-Se but takes into account improvements in library construction and sequencing, especially orientation information. We added detection of transcription termination sites (TTS) using the polyA signal and the optional use of Cap-Seq data to predict transcription start sites (TSS). The TrUC (TRanscription Units by Coverage) pipeline is schematized in Fig. 1. TrUC predicts TSS and TTS using the consensus position of Cap-Seq and polyA coverage, respectively. TrUC uses oriented data to predict oriented transcription units, however the software can predict introns from un-oriented RNA-Seq data, using the GT.. AG splice site consensus to determine orientation. TrUC does not consider alternative splicing since exon skipping is not found in *Paramecium* [15]. Like Cufflinks, TrUC can identify non-coding transcripts as it does not look at translation. TrUC multi-threaded Perl software is available from https://github.com/oarnaiz/TrUC.

Since not all genes are expressed at all stages of the life cycle, use of different life-cycle time points can help annotation by increasing the number of genes that are covered by RNA-Seq data. For *P. tetraurelia* transcription unit prediction, we sequenced polyA+ RNA from a time-course through the sexual process of autogamy as well as from vegetative cells. Combining all the RNA-Seq and Cap-Seq data (Additional file 1: Table S1), TrUC predicts 37, 847 transcription units greater than 300 nt in size, 12,389 TSS and 5967 transcription termination sites (TTS). We found 85% of the previously annotated *P. tetraurelia* genes [6] (hereafter called "v1" annotation) covered by a predicted transcription unit. The average size of the predicted transcription units, 1229 nt, is close

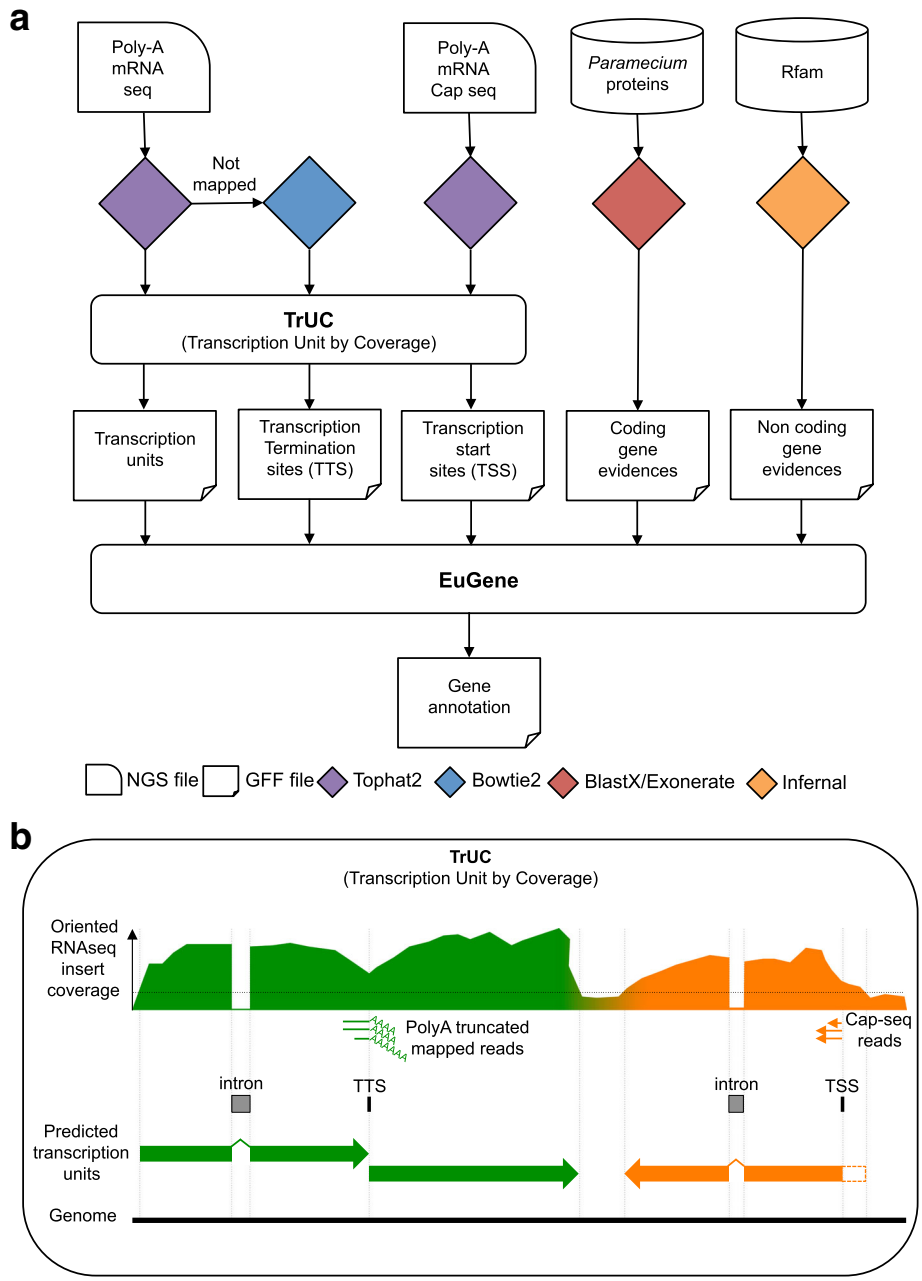


Fig. 1 Gene annotation strategy. **a** Overview of the workflow. EuGene software, using a Paramecium-trained matrix, combines (i) transcription unit predictions, (ii) TSS predicted positions, (iii) TTS predicted positions, (iv) *Paramecium* predicted proteins mapped on the reference genome using BLASTX then Exonerate, and (v) non-coding gene predictions obtained using the Rfam database. **b** Schema of the TrUC pipeline. TrUC is able to predict transcription units, TSS and TTS positions. To achieve this, the software uses oriented polyA$^+$ mRNA-Seq and Cap-Seq data. The upper part of the schema represents RNA-Seq insert coverage of the genome. A configurable threshold (horizontal dotted line) is used to determine the edges of the transcription units. The middle of the schema shows how intron, TSS and TTS positions are predicted. The transcription units predicted by combining all of the information are shown at the bottom of the schema. The TSS and the TTS are used to refine the structure of the transcription unit predictions. This can be particularly critical in a compact genome to avoid fusing adjacent transcription units. An example is shown in orange, where the TSS is used to shorten the predicted transcription unit, removing the open box. The example in green, shows how a TTS can prevent fusion of two adjacent transcription units

to the average size of the v1 genes (Table 1; Additional file 2: Figure S1) indicating that most of the predicted transcription units correspond to one gene. However, some transcription unit predictions may be split, owing to reduced RNA-Seq coverage within the unit. Alternatively, since genes in the compact *P. tetraurelia* genome

Table 1 Annotation statistics

	P tetraurelia		*P. biaurelia*		*P. sexaurelia*		*P. caudatum*	
	V1	V2	V1	V2	V1	V2	V1	V2
Genome size (nt)	72,094,543	72,102,941	75,777,660	74,348,537	67,662,147	67,662,147	29,932,356	29,932,356
Number of scaffolds	697	697	1426	1026	230	230	274	274
N50	412,884	412,881	156,715	159,040	430,207	430,207	312,370	312,370
Genomic GC content	0.28	0.28	0.25	0.25	0.24	0.24	0.28	0.28
Gene number	39,642	41,533	39,110	40,741	34,909	36,477	18,421	18,853
Gene length (nt)	1433.8	1409.91	1460.56	1430.22	1462.39	1430.91	1449.52	1448.91
Percent coding	75	75	71	73	71	73	85	86
Inter coding distance (TGA ↔ TGA)	285.23	244.88	304.13	275.83	362.68	331.12	109.29	87.73
Inter coding distance (ATG ↔ ATG)	451.27	388	462.77	410.36	565.88	501.34	143.39	120.63
Inter coding distance (TGA ↔ ATG)	332.03	286.77	357.91	333.08	433.5	393.35	131.64	116.84
Protein Coding Gene number	39,642	40,460	39,110	40,179	34,909	36,053	18,421	18,592
CDS length (nt)	1363.32	1330.35	1375.51	1359.78	1380.08	1367.54	1386.16	1385.68
Protein Coding Gene GC content	0.3	0.3	0.26	0.26	0.26	0.26	0.29	0.29
Non-coding gene number		1073		562		424		261
Exon number	130,216	136,527	141,873	135,427	126,637	124,022	63,789	62,173
Median exon length (nt)	230	222	200	221	202	213	216	232
Mean exon length (nt)	419.01	411.3	379.19	412.21	380.44	402.57	400.3	423.13
Exon per gene	3.28	3.29	3.63	3.32	3.63	3.4	3.46	3.3
Intron number	90,282	94,711	102,763	94,686	91,728	87,545	45,368	43,320
Median intron length (nt)	25	25	26	25	26	25	22	22
Mean intron length (nt)	25.14	25.31	32.37	25.82	31.32	25.91	25.72	23.3
Number of introns >40 nt	38	720	12,817	1366	9718	1218	2425	720

The genome assembly and v1 annotation of *P. tetraurelia* were published in [6]. The v2 annotation used the same assembly after polishing with Illumina reads, reported in [37]. The v1 genome assembly and annotation of *P. biaurelia* and *P. sexaurelia* were published in [7] and the v1 genome assembly and annotation of *P. caudatum* were published in [8]. The v2 annotations are those obtained in the present study. In the case of *P. biaurelia*, the reference genome was filtered to remove scaffolds of obvious bacterial origin before v2 annotation. All annotations are integrated into ParameciumDB and available for download as GFF3 files

sometimes overlap, some transcription units may be fused owing to continuous RNA-Seq coverage, a problem partially overcome by use of the predicted TSS and TTS (cf. Fig. 1b). Split or fused predicted transcription units do not compromise gene annotation (cf. next section), which also takes into account protein-coding potential to define gene models.

A total of 85,236 introns were identified in the predicted transcription units, corresponding to a mean of 2.25 introns per transcription unit, very close to the 2.3 introns per gene previously reported for *P. tetraurelia* [6, 15].

Gene annotation

Accurate, user-friendly gene annotation tools for eukaryotes, such as AUGUSTUS [16], would require code modifications to correctly identify tiny introns [17]. Indeed, ~ 98% of *P. tetraurelia* introns are 20–30 nt in size, with a mean of 25 nt. A handful of significantly larger introns, ~ 80 nt in size, contain snoRNAs [18]. We therefore decided to train the highly configurable EuGene annotation software [12] for *Paramecium*, using gene models completely confirmed by RNA-Seq coverage (see Methods).

Table 1 compares gene annotations predicted by EuGene for *P. tetraurelia*, using the transcription units assembled with TrUC and other lines of evidence (labeled 'v2'), with the v1 gene annotation for this species [6], long considered a gold standard for ciliate gene annotation. The statistical differences are slight, aside from the fact that the v1 statistics do not include any non-coding gene predictions. The v2 annotation contains about 800 more protein-coding gene models, probably because of the greater quantity of transcript evidence allowing prediction of short genes (the average CDS length is slightly smaller in v2 annotation). To determine sensitivity, a set of 1690 manually curated genes was constituted (available from ParameciumDB [19]). We found 95% of the gene structures (excluding UTRs) and 99% of the introns to be correctly annotated.

In order to analyze the impact of the Cap-Seq data, we ran the same gene annotation pipeline for *P. tetraurelia* without any Cap-Seq data and compared the two sets of gene models. We found 91.8% of the protein-coding gene models to be identical. Among the 3293 gene models that were different, there were 649 cases where addition of the Cap-Seq data split one gene model into two gene models. In most of the remaining cases (2149), the coding sequence was changed. The Cap-Seq data we generated thus have a modest but significant impact on the gene annotation. Knowledge of TSS for the *P tetraurelia* model will be particularly valuable for functional studies. We looked for possible alternative TSS by evaluating whether adjacent TSSs fall within the same gene. In >99% of the cases, adjacent TSS were in different genes. In only 29 cases could we find adjacent TSS in the same gene. We consider it likely that they represent technical noise given their occurrence mainly in highly expressed genes (Additional file 1: Table S2) but we cannot exclude biological significance. We conclude that alternative TSS are extremely rare in *Paramecium*.

The size distribution of 10,087 5′ UTRs that could be confidently predicted from the Cap-Seq data, shown in Fig. 2, confirms the unusually small size of 5'UTRs in *Paramecium* (mean 21.7 nt, median 19 nt), much smaller than in animals and fungi (100–200 nt average, [20]) or the ciliate *Tetrahymena thermophila* (n = 4149, mean = 95.6 nt, median = 88 nt) [21]. The 4641 3′ UTRs predicted from polyA tracts present in the RNA-Seq data (Fig. 2) display a nearly Gaussian distribution (mean = 44.6 nt, median = 44 nt) and are much smaller than in animals and fungi (200–1000 nt average, [20]) or *Tetrahymena thermophila* (n = 1290, mean = 231 nt, median = 163 nt, [21]).

The *Paramecium*-trained annotation pipeline was used to annotate *P. biaurelia, P. sexaurelia* and *P. caudatum* genomes, using the *P. tetraurelia* v2 predicted proteins as evidence as well as the unoriented RNA-Seq data previously used for the published annotation of these species [7, 8]. In all 4 species, we observe ~2.3 introns/gene and the same median intron size in v1 and v2 annotations (Table 1). However, the average intron size is larger in the published *P. biaurelia*, *P. sexaurelia* and – to a lesser extent – *P. caudatum* v1 annotations than in the corresponding v2 annotations (Table 1). This is owing to the prediction of significant numbers of introns larger than 40 nt in the published v1 annotations (Table 1, Fig. 3). Far fewer questionable "large" introns are found in the v2 annotations (10%, 12.5% and 30% with respect to the v1 annotations for *biaurelia*, *sexaurelia* and *caudatum*, respectively). This is because the v2 annotations integrate TrUC intron predictions. Most of these "large" introns are likely to be incorrect and as a consequence, so are the thousands of gene models that contain them. The difference between v1 and v2 annotations is less pronounced for *P. caudatum,* which has the smallest introns so far reported for the *Paramecium* genus (median size 22 nt, average size 23 nt). We conclude that TrUC predictions improve *Paramecium* gene models not only by predicting TSS and TTS if Cap-Seq and oriented mRNA-Seq data are available, but also thanks to the intron predictions, which can be made even from unoriented RNA-Seq data.

Paramecium genomes are intron-rich, an ancestral property of eukaryotes [22]. Not only is intron size very small, but no exon-skipping has been observed [15]. These properties helped discover translational control of eukaryotic intron splicing [15]. In *Paramecium*, splice

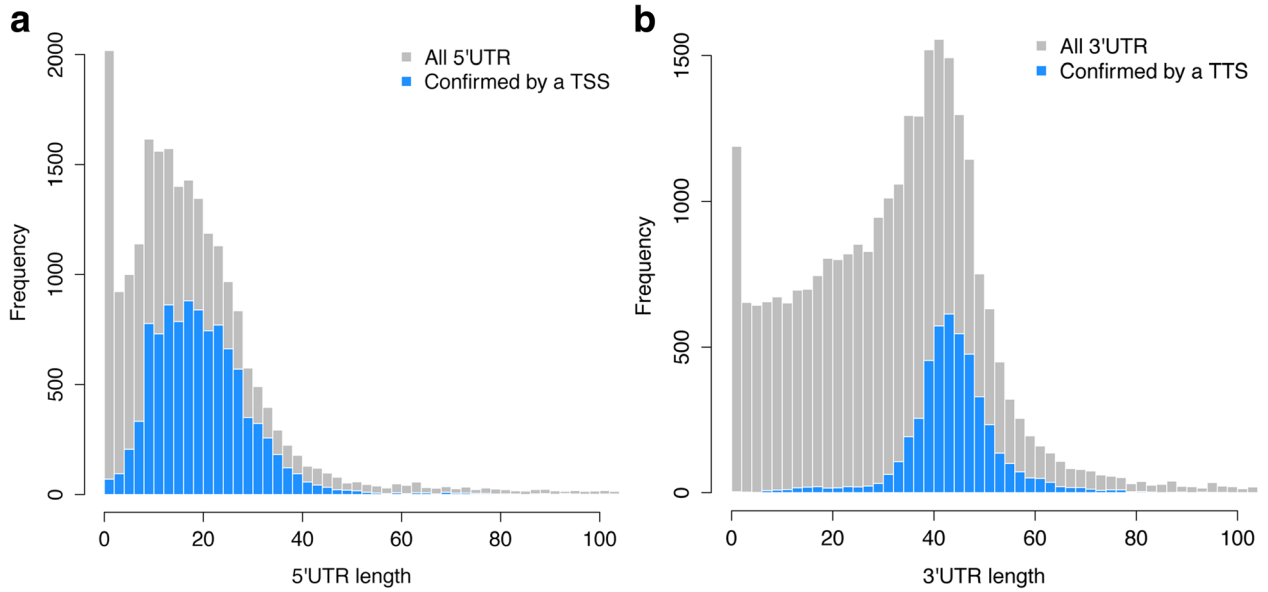


Fig. 2 Size distribution of *P. tetraurelia* 5′ and 3′ UTRs. **a** Histograms of the distribution of predicted 5′ UTR sizes (gray) and 5′ UTR sizes confirmed by a TSS prediction using Cap-Seq data. **b** Histograms of the distribution of predicted 3′ UTR sizes (gray) and 3′ UTR sizes confirmed by a TTS prediction using the polyA$^+$ RNA-Seq read pairs. The Cap-seq confirmed 5′ UTRs have a mean size of 21.7 nt and a median size of 19 nt. The 3′ UTRs predicted from polyA tracts have a mean size of 44.6 nt and a median size of 44 nt

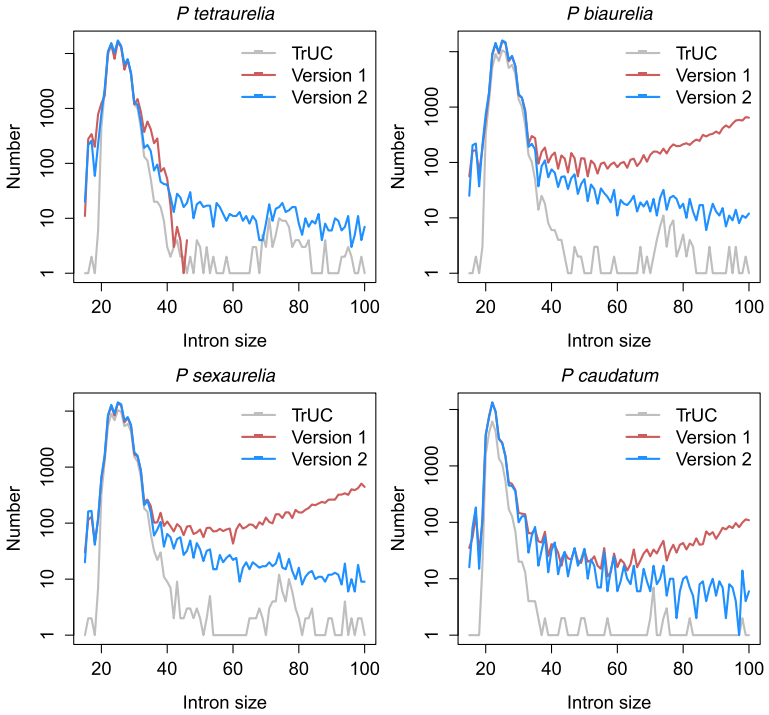


Fig. 3 Intron size distribution. The graphs show the number of introns as a function of their size for *Paramecium tetraurelia, biaurelia, sexaurelia* and *caudatum*. For each species, 3 curves are superimposed: introns predicted by TrUC (*gray*), introns in the v1 annotations (*red*) and introns in the v2 annotations (*blue*). The v2 introns are the subset of TrUC introns in gene models and additional introns predicted by Eugene without RNA-Seq support. The *P. tetraurelia* v1 introns were intentionally filtered when greater than 45 nt [6]. In all cases, the great majority of introns are smaller than 40 nt. In v1 annotations (except *P. tetraurelia*), there are more large introns than in TrUC or v2 annotations. Since the RNA-Seq (TrUC predictions) are expected to show the true distribution of intron sizes, we are more confident in the v2 annotations

site signals are weak, presumably because of the mutational burden, and splicing is not very accurate. The nonsense-mediated decay (NMD) pathway [23] cleans up the mess on the pioneer round of translation, provided that there is a Premature Termination Codon (PTC) in the poorly spliced transcript. That this is the case is easily visualized in *Paramecium*, thanks to the unique TGA stop codon, by comparing the frequency of introns that do or do not contain a STOP codon in phase with the upstream exon, as a function of their size modulus 3. As shown in (Additional file 2: Figure S2), stopless introns that are 3n in size are counter-selected in all 4 species: these introns cannot give rise to a PTC if retained in the transcript so are potentially deleterious. The observed deficit of 3n stopless introns in all 4 species (Additional file 2: Figure S2) validates the annotation and extends previous observation of translational control of intron splicing [15] to other *P. aurelia* species and to *P. caudatum*.

Differential gene expression

In *Paramecium*, the sexual cycle encompasses meiosis, fertilization and development of a new MAC. The latter process involves programmed elimination of at least 25% of the germline DNA. Custom microarrays were previously used to characterize differential gene expression during autogamy (auto-fertilization) in *P. tetraurelia* [9]. We now report use of the mRNA-Seq samples for differential gene expression. Since cells enter autogamy from a fixed point in the vegetative cell cycle [24], which is not synchronized in our cell cultures, there is an asynchrony of at least 5 h in the samples. Therefore, cytology data (Additional file 2: Figure S3) and gene expression levels were used to cluster samples into 6 stages: vegetative (Veg, n = 2), meiosis (Mei, n = 2), fragmentation (Frg, n = 3), early development (Dev1, n = 2), intermediate development (Dev2 and Dev3, n = 4) and late development (Dev4, n = 2) (see Methods, Additional file 1: Table S1 and Additional file 2: Figure S3). These stages are equivalent to those used previously with microarrays [9], with the addition of one later stage, Dev4.

For analysis of differential gene expression (DGE) we first counted mapped RNA-Seq fragments for each v2 gene model (see Methods). We found 99.8% of the mapped fragments in the sense orientation and 0.2% in the anti-sense (AS) orientation (see Additional file 2: Figure S4). This low level of AS transcription might reflect biological noise or pervasive transcription [25], especially as pervasive non-coding transcription is required for genome rearrangements in *Paramecium* [26, 27]. We cannot formally exclude errors in strand-specificity during construction of the sequencing libraries [28]. An intriguing possibility is regulation of gene expression by AS transcripts at specific loci. Higher coverage or long reads will be needed to interpret the AS transcription we have detected.

The DGE analysis was performed with the sense fragment counts and DESeq2 software. Requiring an adjusted p-value <0.01 and a fold-change of 2 in expression, we identified 17,190 genes whose expression varied during autogamy. We separated these genes into induced (n = 8220) or repressed (n = 8970) (cf. Methods), and then used hierarchical clustering to visualize the induced or repressed genes and to define 6 clusters: 'Early peak', 'Intermediate peak', 'Late peak', Late induction', 'Early repression', 'Late repression' (Additional file 2: Figures S5 and S6).

To validate the clusters, we used published genes involved in autogamy and some negative control genes not involved in autogamy (Additional file 1: Table S3). All of the genes known to be induced during autogamy were classified in an appropriate RNA-Seq induced cluster, although 4 of the 26 genes had not been found in the microarray clusters. Genes known to be expressed during meiosis in *Paramecium* (*SPO11*, *SPT5m*, *DCL2*, *DCL3*, *NOWA1*, *NOWA2, LIG4a* and *XRCC4*) [29–33] are found in the Early peak. Only one of the 12 negative control genes was induced during autogamy: *PTIWI14*, characterized as a component of the vegetative siRNA pathway [34], appears in the Late induction cluster. We also compared the distribution of the genes in the microarray clusters with respect to the RNA-Seq clusters (Table 2). The microarray resource was of good quality and essentially all of the genes in the microarray clusters are found in the RNA-Seq clusters. The RNA-Seq approach is more powerful and allows more genes to be identified as differentially expressed during autogamy. A modest qualitative difference between microarray and RNA-Seq classification concerns the microarray 'Intermediate induction' cluster. The genes in this cluster are found in either the RNA-Seq 'Intermediate peak' or the RNA-Seq 'Late induction' clusters. This may be a consequence of the additional late time points in the RNA-Seq experiment (Dev4) which change the gene expression profiles being clustered.

To estimate how well RNA-Seq data can discriminate the ~12,000 pairs of paralogs created by the recent WGD (~83% nucleotide identity on average), we computed the number of identical stretches of 100 nt shared by each pair. We estimate that >97% of the paralog pairs are devoid of common 100 nt stretches over >90% of their length and should therefore be well-discriminated in the present analysis. This is likely to be an underestimate, since the procedure uses mapping of pairs of 100 nt reads, not of single 100 nt reads. Therefore, in theory, RNA-Seq should better discriminate the paralogs than the previous microarray data [9]. In addition, RNA-Seq has greater dynamic range and sensitivity than

Table 2 Differential Gene Expression

	Early peak	Intermediate peak	Late peak	Late induction	Early repression	Late repression	none
Early peak	333	17	0	1	0	8	1
Early induction	49	33	1	9	0	0	0
Intermediate induction	5	315	17	211	0	0	1
Late induction	0	7	0	28	1	0	0
Early repression	0	0	0	1	209	30	4
Late repression	28	0	0	0	1	1022	3
none	1315	1398	395	2937	3986	3072	20,043

Distribution of differentially expressed (DE) genes in RNA-Seq clusters (columns) and microarray clusters [9] (rows). Overall, the two analyses provide similar results, but the RNA-Seq approach detects many more DE genes. Essentially all microarray DE genes were found by RNA-Seq. The main qualitative difference is that the genes in the microarray Intermediate induction cluster are now found in one or the other of 2 RNA-Seq clusters: Intermediate peak or Late induction. See Methods for details of the analysis

microarray technology ([35] and our data), which can also contribute to good paralog discrimination. The *P tetraurelia* microarrays involved hybridization of 6 probes (50 nt) per gene designed to optimize chances of discriminating the paralogs. However, it is difficult to evaluate the extent of microarray cross-hybridization. In order to compare the discrimination of WGD paralogs by the microarray and RNA-Seq methods, we calculated paralog expression level divergence (Additional file 2: Figure S7). Paralogs are more sensitively discriminated by RNA-Seq because of its greater dynamic range (Additional file 2: Figure S7a, b). When the paralogs share high nucleotide identity (with the greatest risk of cross-hybridization), there is little difference between their expression levels in the microarray data, but the same is true of the RNA-Seq data (Additional file 2: Figure S7a–d). We propose two, non-exclusive, explanations to account for this observation, the first technical, the second biological. The first explanation is that RNA-Seq cannot discriminate highly similar paralogs because reads are mapped randomly between 2 equivalent loci, which reduces the difference in the measured expression level of such loci. The second explanation is that the paralogs sharing high nucleotide identity probably tend to have similar expression levels. This could result from strong selective pressure on highly expressed genes [10] or from gene conversion between WGD paralogs, frequently observed for *P. aurelia* species [7]. Gene conversion leads to increased GC content and high nucleotide identity between paralogs, and can extend to promotor regions depending on the recombination breakpoint. Indeed, expression level and GC content are correlated with the nucleotide identity of *P. tetraurelia* paralogs (Additional file 2: Figure S7e, f).

We also looked at whether genes duplicated at the recent WGD have kept the same expression profiles. First, we removed paralogs that differ in length by more than 10%, a filter that removes potential pseudogenes (n = 10,323; 85% of the paralog pairs are retained after this stringent filtering, *P. tetraurelia* v2 annotation). For 22% of the pairs, we found both paralogs in the same autogamy cluster, and for 38%, neither paralog was differentially expressed during autogamy. Interestingly, we found 31.5% of the pairs had one paralog in an autogamy cluster and one paralog not differentially expressed. This might reflect an early stage of pseudogenization, shown to begin with changes in expression level [11], probably via mutations in promoter sequences. We also found 8.5% of pairs with paralogs in different autogamy clusters, a possible indication of neo- or sub-functionalization. The important point is that finding paralogs with different expression profiles, irrespective of the origin of the difference, is an indication that the paralogs are well-discriminated by the RNA-Seq data.

A qualitative picture of the biological processes turned on during autogamy can be obtained using Gene Ontology (GO) terms [36], with the caveat that *Paramecium* functional annotation has not been curated. We first made a high-confidence gene set by requiring a fold-change of 4 during autogamy (7065 genes). We then used GO Biological Process terms associated with protein domains (cf. Methods) to make word clouds (Additional file 2: Figure S8). The Early peak (Additional file 2: Figure S8a) covers meiosis and fertilization and is enriched in appropriate terms: *DNA*, *repair*, *chromosome*, *mismatch*, *condensation, homologous*, *recombination*, *Okazaki*, *replication*, *chromatid*, *cohesion*. The Intermediate peak (Additional file 2: Figure S8b) corresponds to development of the new MAC involving chromatin remodeling and shows enrichment in *repair, chromatin, DNA*, *methylation*, and *chromosome*. Many biological processes are activated in the Late peak and Late induction clusters so that the only over-represented informative words relate to signal transduction (*GTPase, signal, transduction, phosphorylation, inositol)* and membrane trafficking (*vesicle-mediated, autophagy)* (Additional file 2: Figure S8c, d). The cluster of genes that are turned off when cells leave vegetative growth and enter the sexual cycle (Early repression, Additional file 2: Figure S8e) is enriched for words that refer to translation and cellular homeostasis (*rRNA, redox, homeostasis,*

ribosome, oxido-reduction, translation, pseudouridine). For the Late repression cluster (Additional file 2: Figure S8f), the only informative words refer to *microtubule-based, movement* and *phosphorylation, dephosphorylation.*

The gene expression atlas is provided as a table (Additional file 3: Table S4).

Conclusions

Plummeting genome sequencing costs and rising interest in *Paramecium* species for studies of genome evolution in unicellular eukaryotes, prompted us to build a new pipeline for gene annotation that takes into account specificities of the genus, in particular high gene density and stereotyped small intron size. This has been achieved with new software to predict transcription units (TrUC) and specific training of the highly configurable EuGene gene annotation software. High quality gene annotations will be important for future comparative and functional genomics analyses of *Paramecium* species. The mRNA-Seq data used to predict *P. tetraurelia* transcription units for the gene annotation enabled us to generate an improved gene expression atlas and carry out differential gene expression analysis of the sexual cycle of autogamy that is more complete than previously possible with microarrays.

Methods

RNA samples and mRNA sequencing

Three time-course experiments for the sexual process of autogamy of *P. tetraurelia* wild-type strain 51 were used. Some of the samples had previously been used for microarray experiments [9] (see Additional file 1: Table S1). Total RNA was Trizol-extracted as previously described [9]. PolyA$^+$ RNA was extracted from each sample using the FastTrack MAG mRNA isolation kit (Thermo Fisher Scientific) following the manufacturer's instructions. Strand-oriented Illumina libraries were made with reagents from the Illumina TruSeq Small RNA library preparation kit using an adapted protocol. First, RNA was fragmented by incubation for 4 min at 94 °C with New England Biolabs' Mg2+ solution, yielding fragments with an average size of ~260 nt. The RNA was purified using RNeasy columns (Qiagen) followed by treatment with antarctic phosphatase and polynucleotide kinase (New England Biolabs) and another purification on RNeasy columns. Illumina adapter ligation and RT-PCR was done essentially following the Illumina protocol, except that for the final library purification step AMPure beads were used (Beckman Coulter). HiSeq paired-end sequencing was performed on the samples, yielding at least 2 × 30 million reads, 100 nt in length, for each of the samples.

Cap-Seq

Five samples, representing 3 time points (Veg, T0 and T11; [9]) were used for 5′ Transcription Start Site (TSS) mapping. Purified polyA$^+$ RNA was dephosphorylated using Calf Intestine Phosphatase (CIP) prior to 5′ cap removal with Tobacco Acid Pyrophosphatase (TAP), using a FirstChoice RLM-RACE kit (Life Technologies). Illumina 5′ adaptors were ligated to the 5′ monophosphate ends generated specifically at TSSs, followed by RNA fragmentation by incubation for 4 min at 94 °C with New England Biolabs' Mg^{2+} solution. This yielded an average fragment size of 260 nt. Following CIP treatment to convert the 3′ monophosphate ends generated by RNA fragmentation to 3'OH ends, Illumina 3′ adapters were ligated to the fragments. The libraries were subjected to RT-PCR amplification (18–20 cycles) before Illumina HiSeq paired-end sequencing that yielded approximately 2 × 13 million reads of 100 nt per sample. Every step in TSS library preparation ended with purification using RNAeasy columns (Qiagen) or phenol/chloroform extraction and isopropanol precipitation. The final PCR amplification products were purified using AMPure beads (Beckman Coulter) before sequencing.

Transcription unit determination

We developed the multi-threaded Perl software TrUC (Transcription Units by Coverage), dependent on the Bio::DB::Sam module. The software is organized in 3 independent modules (Fig 1). The TSS module uses Cap-Seq data, which need not be paired-end, to predict transcription start sites. The predicted TSS is the position with the highest Cap-Seq coverage in the interval defined by the size of the fragments. The TTS module uses oriented paired-end mRNA-seq reads. If one of the reads in the pair maps partially on the reference genome and ends in polyA, then the insert is used to specify a transcription termination site. The predicted TTS is the position with the highest polyA coverage in the interval defined by the size of the fragments. The transcript module takes paired-end TopHat2 mapping (BAM files; [14]) and optionally, the output of the TSS and TTS modules, to predict transcription units including intron positions, based on fragment coverage. TrUC was run with the following parameters for *P. tetraurelia* annotation:

truc TSS -min_coverage 15 -nb_replicates 2 -min_score 500; truc TTS -min_coverage 5 -nb_replicates 2 -min_score 10 -nb_min_A 5; truc transcript -min_splicing_rate 0.7 -no_overlap -min_coverage 15 \

-intron_consensus -min_intron_length 15 -max_intron_length 100 \
-min_intron_coverage 15 -min_length 300 -min_score 45 -tss [truc TSS GFF3 output file] -tts [truc TTS GFF3 output file].

For *P. biaurelia, P. sexaurelia and P. caudatum,* TrUC was used with unoriented RNA-Seq data reported in [7, 8] (Accessions PRJNA268243, PRJNA268244 and PRJNA268245, respectively), to predict introns, with the following parameters:

truc transcript -not_stranded -min_splicing_rate 0.7 -min_coverage 10 -intron_consensus -min_intron_length 10 \ -max_intron_length 100 -min_intron_coverage 3 -min_length 300 -min_score 10.
TrUC is distributed under a GNU GPL v3 license at https://github.com/oarnaiz/TrUC.

Gene annotation

The workflow for gene annotation is schematized in Fig. 1. EuGene software [12] was used to predict gene structure. EuGene was trained with 1597 curated *Paramecium tetraurelia* genes to generate a prediction matrix that takes into account the unusually small size of *Paramecium* introns [15]. The prediction matrix is available from http://paramecium.i2bc.paris-saclay.fr/download. The evidence sets used for annotation of the 4 *Paramecium* genomes are available on request.

Comparative genomics

UTR lengths for *Tetrahymena thermophila* were calculated using the June 2014 annotation available at http://www.ciliate.org/system/downloads/T_thermophila_June2014.gff3.

Differential gene expression

Paired-end RNA-Seq reads were mapped to the reference *P. tetraurelia* strain 51 genome [37] using TopHat2 (v2.0.12, −−mate-inner-dist 50 −mate-std.-dev 100 −min-intron-length 15 −max-intron-length 100 −coverage-search −keep-fasta-order −library-type fr-secondstrand −min-coverage-intron 15 −max-coverage-intron 100 −min-segment-intron 15 −max-segment-intron 100 −max-multihits 1 −read-mismatches 1 −max-deletion-length 1 −max-insertion-length 1). Raw fragment counts for the genes in each sample, determined using htseq-count (v0.6.0 −stranded = yes −mode = intersection-non-empty) on filtered BAM files (samtools v0.1.18 samtools view -q 30), were used as input for DESeq2 (v1.4.1) [38], an R Bioconductor package, which normalizes the fragment counts, calculates the dispersion in the data using the biological replicates, and then determines differential gene expression using negative binomial linear models. The samples were grouped into biological replicates (Veg, Mei, Frg, Dev1, Dev2_3, Dev4) using the cytology data and clustering of the sample normalized counts with a distance matrix (see Additional file 2: Figures S3 and S4 sample dendrogram; see Additional file 1: Table S1). We considered genes to be differentially expressed during autogamy if at least one contrast between Veg and any point in the autogamy time course had an adjusted *p*-value smaller than 0.01 and a fold-change (FC) of expression greater than 2. We filtered out genes if there was not at least one time point with more than 20 normalized counts. The genes were classed as induced (FC > 2) or repressed (FC < ½) before hierarchical clustering.

GO term enrichment and word cloud visualization

To gain qualitative appreciation of the processes associated with the groups of co-expressed genes, we focused on a subset of differentially expressed genes with a fold-change >4 (and an adjusted *p*-value <0.01). GO biological process terms were electronically inferred using InterProScan (v5.7.48) domain annotation of the corresponding proteins. If more than one protein domain was associated with a protein, the domain with the lowest InterProScan *P*-value was retained. All words in the terms were counted for all the protein-coding genes in the genome and for the protein-coding genes in each co-expression group. After removing non-discriminatory words ("protein", "process", "domain" and the "stop-words" defined by the wordcloud R package, v. 2.5), a Fisher exact test was used to determine the word enrichment ratio (*p*-value <0.05) in each co-expression group with respect to the word frequency for the whole genome. A score determined for each word (score = $\log_2(p\text{-value}^{-1})$) was used as weight to draw each word cloud (R wordcloud v2.5). The protein domains and GO terms used for this analysis can be found in the gene expression atlas (Additional file 3: Table S4).

Additional files

Additional file 1: Table S1. RNA sequencing. **Table S2.** Genes with potential alternative TSS. **Table S3.** Differential expression of genes with known autogamy expression profiles. (XLSX 38 kb)

Additional file 2: Figure S1. Comparison of the sizes of *P. tetraurelia* transcription units and genes. **Figure S2.** Intron size distributions. **Figure S3.** Autogamy time-course experiments. **Figure S4.** Anti-sense transcription. **Figure S5.** Hierarchical clustering of differentially expressed genes. **Figure S6.** Autogamy co-expression clusters. **Figure S7.** Paralog discrimination by microarrays and RNA-Seq. **Figure S8.** Word cloud analysis of biological processes in clusters. (PDF 8271 kb)

Additional file 3: Gene expression atlas. All *P. tetraurelia* v2 genes ('ID') with their normalized RNA-Seq counts (last 15 columns, sample labels as in Additional file 2: Table S1) are given. The mean value for biological replicates are given in the columns VEG, MEI, FRG, DEV1, DEV2/3, DEV4. The '*P*-value' 'Significant' and 'Expression profile' refer to the differential gene expression analysis (cf. Methods). 'Note' is the description of the best SwissProt BLASTP match. The GO ID and GO description were inferred electronically using InterProScan. The Biological Process GO term associated with the highest scoring protein domain is given. (TSV 12379 kb)

Abbreviations

AS: Anti-sense; DE: Differentially Expressed; DGE: Differential Gene Expression; GFF3: Generic Feature Format, version 3; GO: Gene Ontology; MAC: Macronucleus; MIC: Micronucleus; PTC: Premature Termination Codon;

TrUC: Transcription Units by Coverage; TSS: Transcription Start Site; TTS: Transcription Termination Site; WGD: Whole Genome Duplication

Acknowledgements
We thank Thomas Schiex for advice about training EuGene. This work has benefited from the facilities and expertise of the High-throughput Sequencing Platform of the I2BC. We are grateful to the INRA MIGALE bioinformatics platform (http://migale.jouy.inra.fr) for providing help and support. This work was carried out in the context of the CNRS-supported European Research Group "Paramecium Genome Dynamics and Evolution".

Funding
This work was supported by Centre National de la Recherche Scientifique intramural funding, the Agence Nationale de la Recherche Scientifique ANR-14-CE10–0005–04 'PIGGYPACK' to M.B., S.D. and L.S. and ANR-12-BSV6–0017-03 'INFERNO' to M.B., S.D. and L.S., the 'Comité d'Ile de France de la Ligue Nationale Contre le Cancer' to S.D. and an 'Equipe FRM DEQ20160334868' grant to S. D. The funding bodies had no role in the design of the study, analysis, or interpretation of data or in writing the manuscript.

Availability of data and materials
The 15 RNA-Seq and 5 Cap-Seq datasets were deposited in the ENA under the Project Accession PRJEB19343 (see Additional file 1: Table S1). All gene annotation versions are available as GFF3 files from http://paramecium.i2bc.paris-saclay.fr/download/.

Authors' contributions
EVD, MB, MLK, AV and SD prepared samples and acquired data; OA, ES, and JG developed methods and analyzed data; OA, MB, SD and LS conceived the study; OA and LS prepared the manuscript. All authors approved the manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Institute for Integrative Biology of the Cell (I2BC), CNRS, CEA, Univ. Paris-Sud, Université Paris-Saclay, 91198 Gif-sur-Yvette CEDEX, France. [2]Institut Jacques Monod, CNRS, UMR 7592, Université Paris Diderot, Sorbonne Paris Cité, F-75205 Paris, France. [3]LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France. [4]Current address: IRCM, CEA, INSERM UMR 967, Université Paris Diderot, Université Paris-Saclay, 92265 Fontenay-aux-Roses CEDEX, France.

Received: 20 February 2017 Accepted: 21 June 2017
Published online: 26 June 2017

References

1. Chalker DL, Meyer E, Mochizuki K. Epigenetics of ciliates. Cold Spring Harb Perspect Biol. 2013;5:a017764.
2. Sonneborn TM. Sex, Sex Inheritance and Sex Determination in *Paramecium Aurelia*. Proc Natl Acad Sci U S A. 1937;23:378–85.
3. Saimi Y, Kung C. Behavioral genetics of Paramecium. Annu Rev Genet. 1987; 21:47–65.
4. Betermier M, Duharcourt S. Programmed rearrangement in ciliates: Paramecium. Microbiol Spectr. 2014;2:MDNA3-0035-2014.
5. Coleman AW. *Paramecium aurelia* revisited. J Eukaryot Microbiol. 2005;52: 68–77.
6. Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, et al. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature. 2006;444:171–8.
7. McGrath CL, Gout J-F, Johri P, Doak TG, Lynch M. Differential retention and divergent resolution of duplicate genes following whole-genome duplication. Genome Res. 2014;24:1665–75.
8. McGrath CL, Gout J-F, Doak TG, Yanagi A, Lynch M. Insights into three whole-genome duplications gleaned from the *Paramecium caudatum* genome sequence. Genetics. 2014;197:1417–28.
9. Arnaiz O, Goût J-F, Bétermier M, Bouhouche K, Cohen J, Duret L, et al. Gene expression in a paleopolyploid: a transcriptome resource for the ciliate Paramecium tetraurelia. BMC Genomics. 2010;11:547.
10. Gout J-F, Kahn D, Duret L. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet. 2010;6:e1000944.
11. Gout J-F, Lynch M. Maintenance and Loss of Duplicated Genes by Dosage Subfunctionalization. Mol Biol Evol. 2015;32:2141–8.
12. Foissac S, Gouzy J, Rombauts S, Mathe C, Amselem J, Sterck L, et al. Genome Annotation in Plants and Fungi: EuGene as a Model Platform. Curr Bioinforma. 2008;3:87–97.
13. Denoeud F, Aury J-M, Da Silva C, Noel B, Rogier O, Delledonne M, et al. Annotating genomes with massive-scale RNA-sequencing. Genome Biol. 2008;9:R175.
14. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5.
15. Jaillon O, Bouhouche K, Gout J-F, Aury J-M, Noel B, Saudemont B, et al. Translational control of intron splicing in eukaryotes. Nature. 2008;451:359–62.
16. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinforma Oxf Engl. 2003;19(Suppl 2):ii215–25.
17. Slabodnick MM, Ruby JG, Reiff SB, Swart EC, Gosai S, Prabakaran S, et al. The Macronuclear Genome of Stentor coeruleus Reveals Tiny Introns in a Giant Cell. Curr Biol. 2017;27:569–75.
18. Chen C-L, Zhou H, Liao J-Y, Qu L-H, Amar L. Genome-wide evolutionary analysis of the noncoding RNA genes and noncoding DNA of Paramecium tetraurelia. RNA. 2009;15:503–14.
19. Arnaiz O, Sperling L. ParameciumDB in 2011: new tools and new data for functional and comparative genomics of the model ciliate Paramecium tetraurelia. Nucleic Acids Res. 2011;39:D632–6.
20. Pesole G, Mignone F, Gissi C, Grillo G, Licciulli F, Liuni S. Structural and functional features of eukaryotic mRNA untranslated regions. Gene. 2001;276:73–81.
21. Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, et al. Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. PLoS Biol. 2006;4:e286.
22. Csuros M, Rogozin IB, Koonin EV. A Detailed History of Intron-rich Eukaryotic Ancestors Inferred from a Global Survey of 100 Complete Genomes. PLoS Comput Biol. 2011;7:e1002150.
23. Popp MW-L, Maquat LE. Organizing principles of mammalian nonsense-mediated mRNA decay. Annu Rev Genet. 2013;47:139–65.
24. Berger JD. Autogamy in Paramecium. Cell cycle stage-specific commitment to meiosis. Exp. Cell Res. 1986;166:475–85.
25. Jensen TH, Jacquier A, Libri D. Dealing with pervasive transcription. Mol Cell. 2013;52:473–84.
26. Lepère G, Bétermier M, Meyer E, Duharcourt S. Maternal noncoding transcripts antagonize the targeting of DNA elimination by scanRNAs in Paramecium tetraurelia. Genes Dev. 2008;22:1501–12.
27. Maliszewska-Olejniczak K, Gruchota J, Gromadka R, Denby Wilkes C, Arnaiz O, Mathy N, et al. TFIIS-Dependent Non-coding Transcription Regulates Developmental Genome Rearrangements. PLoS Genet. 2015;11:e1005383.
28. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods. 2010;7:709–15.
29. Marmignon A, Bischerour J, Silve A, Fojcik C, Dubois E, Arnaiz O, et al. Ku-mediated coupling of DNA cleavage and repair during programmed genome rearrangements in the ciliate Paramecium tetraurelia. PLoS Genet. 2014;10:e1004552.
30. Gruchota J, Denby Wilkes C, Arnaiz O, Sperling L, Nowak JK. A meiosis-specific Spt5 homolog involved in non-coding transcription. Nucleic Acids Res. 2017;45:4722–32.
31. Lepère G, Nowacki M, Serrano V, Gout J-F, Guglielmi G, Duharcourt S, et al. Silencing-associated and meiosis-specific small RNA pathways in Paramecium tetraurelia. Nucleic Acids Res. 2009;37:903–15.

32. Nowacki M, Zagorski-Ostoja W, Meyer E. Nowa1p and Nowa2p: novel putative RNA binding proteins involved in trans-nuclear crosstalk in Paramecium tetraurelia. Curr Biol CB. 2005;15:1616–28.
33. Kapusta A, Matsuda A, Marmignon A, Ku M, Silve A, Meyer E, et al. Highly precise and developmentally programmed genome assembly in Paramecium requires ligase IV-dependent end joining. PLoS Genet. 2011;7: e1002049.
34. Marker S, Carradec Q, Tanty V, Arnaiz O, Meyer E. A forward genetic screen reveals essential and non-essential RNAi factors in Paramecium tetraurelia. Nucleic Acids Res. 2014;42:7268–80.
35. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Comput Biol. 2017;13:e1005457.
36. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
37. Arnaiz O, Mathy N, Baudry C, Malinsky S, Aury J-M, Wilkes CD, et al. The Paramecium germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. PLoS Genet. 2012;8:e1002984.
38. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

Chapitre V

Annotation du génome micronucléaire

Dans le chapitre précédent, je me suis intéressé à l'annotation des gènes du génome macronucléaire (MAC). Le présent chapitre est consacré à l'annotation du génome micronucléaire (MIC) et plus particulièrement aux *Internal Eliminated Sequences* (IES) et aux éléments transposables (ET) (voir **section III.3.1** p.64). Les technologies à haut-débit de séquençage ont fait progresser nos connaissances du génome MIC et fait évoluer la manière d'aborder les questions scientifiques.

V.1 Annotation des IES

V.1.1 Identification des IES de *Paramecium tetraurelia*

L'article Arnaiz et al. (2012), qui va suivre, décrit l'identification des IES de *Paramecium tetraurelia* à l'échelle du génome. Cette étude est une étape clé (54 citations relevées en janvier 2020 via PubMed) pour la communauté scientifique étudiant les réarrangements de génome chez la paramécie. En établissant un certain nombre de ressources et protocoles, elle autorise à imaginer des études fonctionnelles et comparatives. L'article formalise le protocole de préparation d'ADN de cellules en cours de développement macronucléaire et inactivées pour un gène impliqué dans les réarrangements. L'extraction d'ADN est suivie d'un séquençage haut débit (voir **Figure V.1** p.106). Nous l'avons dit, elle établit le jeu de référence d'IES du génome de *Paramecium tetraurelia*. Elle propose les fondements méthodologiques et bioinformatiques de traitement de lectures NGS pour la paramécie. En réalité, c'est la première publication utilisant des données NGS chez la paramécie.

Avant d'énoncer les grands résultats de cette étude, voici quelques éléments de contexte de notre connaissance des IES et des réarrangements de génome chez la paramécie au moment de la parution de l'article. Dans les **sections III.2.2** et **III.3.2** (p.58 et p.68), nous avons vu que pendant les processus sexuels de la paramécie, le génome MAC

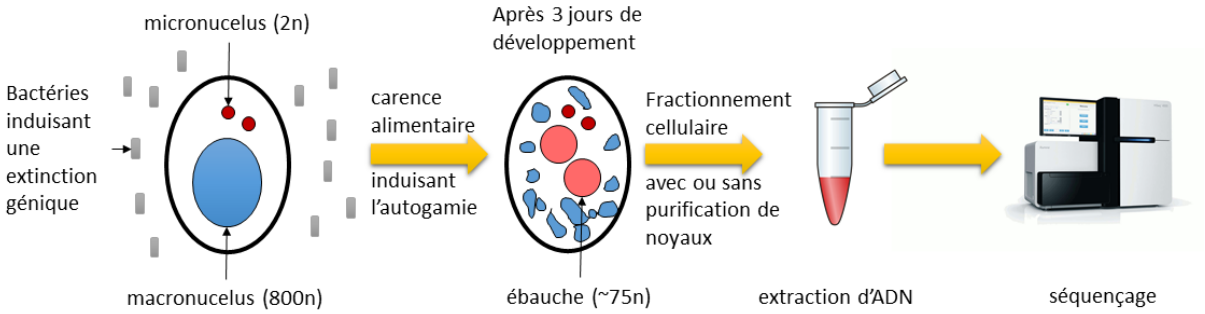


FIGURE V.1 – *Protocole d'extraction d'ADN de cellules inactivées pour un gène*

Les paramécies sont cultivées dans un milieu contenant des bactéries produisant de l'ARNdb et conduisant l'extinction du gène cible (voir **section III.2.3.2** p.61). Une fois les bactéries consommées, la carence alimentaire induit les processus sexuels. Les siARN, toujours présents dans les cellules, réduisent l'expression du gène cible pendant l'autogamie. Dans le cas d'une extinction de *PGM*, les CDB ne sont pas introduites mais la réplication n'est pas inhibée. Après 3 jours, le développement des ébauches n'est pas terminé et des fragments de l'ancien MAC sont toujours présents dans le cytoplasme. L'ADN est extrait, après fractionnement cellulaire suivi ou non d'une purification des ébauches, puis séquencé.

en développement subit des réarrangements génomiques : une endoréplication d'ADN fait passer le noyau d'un état diploïde à un état polyploïde. La complexité du génome MAC se réduit par une élimination imprécise de grandes régions génomiques, contenant notamment des ET ainsi qu'une élimination précise de petites séquences : les IES. En 2006, le génome MAC avait été séquencé (AURY ET AL. 2006) mais le génome MIC restait très largement inconnu car une purification de ce noyau était techniquement impossible à l'époque (voir **section V.2** p.135). La cinquantaine d'IES connues (complétée par 1800 TA-indels par DURET ET AL. (2008), voir **section III.3.2.1** p.71), a permis d'énoncer des caractéristiques qui s'avéreront correctes à l'échelle du génome. Les IES sont petites et peuvent être intragéniques ou intergéniques. Un consensus faible aux bornes des IES (5'TAYAGYNR3') suggère à KLOBUTCHER AND HERRICK (1995) un lien de parenté entre les IES et des ET de la famille des Tc1/Mariner (voir **section II.2.1.2** p.44). En 2009, BAUDRY ET AL. (2009) révèlent l'existence d'un gène codant pour une transposase domestiquée PiggyMac (Pgm) de la famille des ET PiggyBac dans le génome de la paramécie. La protéine Pgm est requise pour les deux types d'élimination d'ADN. L'inactivation de *PGM*, par ARN interférence (voir **section III.2.3.2** p.61), ne perturbe pas la réplication mais empêche l'introduction des CDB pendant les réarrangements programmés. Le séquençage d'ADN d'ébauches enrichies de cellules en cours de réarrangement, dans lesquelles l'expression du gène *PGM* a été déplétée (l'ADN "PGM"), nous donne un aperçu du génome germinal et donc des IES de *Paramecium tetraurelia*.

La première étape pour l'identification des IES est un alignement des lectures de séquençage *Illumina* de l'ADN "PGM" sur le génome MAC de référence. Comme discuté dans les **section I.3.4** (p.28) et **II** (p.31), la sensibilité de l'annotation dépend de la qualité du génome. Or, il faut savoir que le génome MAC de *Paramecium tetraurelia* publié

en 2006 a été fait sur des cellules de la souche *d4-2* (Aury et al. 2006), alors que l'ADN "PGM" a été extrait de cellules de la souche *51*. La souche *d4-2* est issue d'un croisement entre la souche *51* et la souche *29* suivit d'une série de rétrocroisements avec la souche *51*. Le polymorphisme entre les souches *51* et *d4-2* est considéré comme faible (∼2500 SNP d'après mes estimations) mais il était suffisant pour gêner une annotation exhaustive des IES. De plus, nous savions que l'assemblage MAC de *P. tetraurelia* contenait de nombreuses erreurs de séquençage, et notamment des *indels*. En 2009, Pilon (Walker et al. 2014) n'existait pas encore, j'ai donc développé une méthode analogue. A l'aide d'alignements de données de séquençage *Illumina* d'ADN macronucléaire sur le génome MAC, j'ai non seulement corrigé les erreurs d'assemblage (au moins ∼7500 *InDels* et ∼6900 substitutions) mais également obtenu un génome dans lequel les SNP entre les souches *d4-2* et *51* ont été modifiés. Bien qu'ayant la structure du génome de la souche *d4-2*, ce génome MAC de *P. tetraurelia* a été façonné pour ressembler à celui de la souche *51*. La nouvelle version de l'annotation des gènes présentée dans la section IV (p.89) précédente est faite sur ce génome *51* (Arnaiz et al. 2017). Cette référence MAC est encore largement utilisée aujourd'hui, car la souche *51* reste la souche de prédilection d'une grande partie des paraméciologues.

Des approches de détection de sites d'insertion (méthode MIRAA pour *Method of Identification by Read Alignment Anomalies*) et des approches par assemblage global ou local (méthode MICA pour *Method of Identification by Comparison of Assemblies*) ont déterminé la présence de 44 928 IES dans le génome de *P. tetraurelia* (voir plus de détails sur les méthodes dans la **section V.1.2** des résultats, p.127). Les ∼45 000 IES sont réparties dans tout le génome. Elles sont tout autant intergéniques qu'intragéniques, et 47% des gènes contiennent au moins une IES. En revanche, nous observons une asymétrie de densité en IES le long des chromosomes MAC (voir **Figure V.2** p.108), possiblement un indicateur de la structure, encore mal connue, des chromosomes MIC (voir **section III.3.1** p.64).

L'assemblage des lectures PGM génère un génome de ∼100Mb de complexité. Avec un génome MAC de 72Mb, la compléxité des séquences MIC éliminées pendant les réarrangements est donc estimée à au moins ∼28Mb dont 3.5Mb d'IES. Dans cet article, nous concluons que toutes les IES dépendent de PGM pour leur excision. En revanche, dans l'article de Guérin et al. (2017) (voir **section V.2** suivante p.135), nous montrons que le génome MIC n'est pas tout a fait équivalent au génome obtenu à partir de cellules déplétées pour Pgm (∼3 Mb de complexité seraient éliminées indépendamment de Pgm). Au sein des ∼25Mb de séquences éliminées imprécisement, nous avons caractérisé 3 types d'ET de la famille Tc1/*mariner* (voir **section II.2.1.2** p.44) : les éléments appelés *Sardine*, *Thon*, et *Anchois*. Le consensus d'*Anchois* a été reconstruit à partir de 28 longues séquences d'IES par des méthodes basées sur l'homologie (voir **section II.2.2.1** p.48). Ce résultat indique clairement un lien évolutif entre les IES et les ET.

Les IES sont petites : 93% sont inférieures à 150pb, et un tiers ont une taille entre 26 et 30pb. Dans cet article et dans la **section III.3.2.1** (p.68), je commente la distribution de taille des IES et son lien avec des contraintes mécanistiques d'excision. Tirant profit du

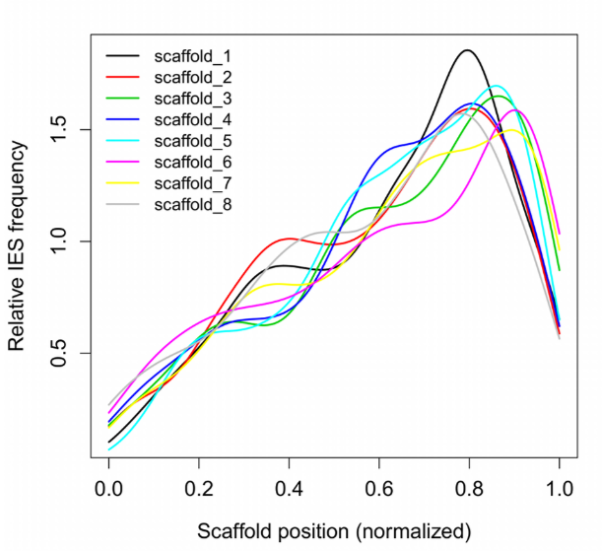


FIGURE V.2 – *Densité en IES sur 8 grands chromosomes MAC*

Figure supplémentaire 4 de ARNAIZ ET AL. (2012) montrant l'asymétrie de densité en IES de 8 scaffolds MAC

grand nombre de paralogues causé par les 3 WGD successives (voir **section III.3.4** p.85), j'ai développé une procédure pour analyser si les IES étaient insérées aux mêmes sites dans les gènes paralogues. Dans 85% des cas, le site d'insertion d'IES est conservé entre deux paralogues de la WGD la plus récente. Par contre, les IES ne montrent aucune trace de conservation nucléotidique. Cette observation est compatible avec une évolution caractéristique de séquences non-codantes, sans pression de sélection. De plus, nous montrons que les éléments à l'origine des IES ont envahi le génome de la paramécie avant et après les WGD. Grâce aux marqueurs temporels, que sont les WGD, nous concluons qu'une IES, après insertion, se décompose progressivement jusqu'à une taille limite de 26 pb. Autrement dit, plus une IES s'est insérée anciennement dans le génome, plus elle sera susceptible d'être courte. Dans la **discussion VII.2.2.1** (p.174), nous verrons que ces résultats ont été confirmés par une étude récente, utilisant une approche de génomique comparative entre IES de 9 espèces de paramécies (Sellis et al., en préparation).

Ma contribution à cette étude : Dans cette étude, mon travail a concerné les aspects computationnels. J'ai développé le logiciel MICA ainsi que la première version de MIRAA. Durant son stage de M2 et sa thèse, Cyril Denby Wilkes a repris et amélioré le code de MIRAA. J'ai réalisé les analyses sur la distribution des IES dans le génome ainsi que la distribution de tailles des IES. J'ai imaginé et conçu l'analyse sur la conservation des sites d'insertion d'IES dans les paralogues issus des DGG récente et intermédiaire. J'ai utilisé les gènes fortement exprimés pour démontrer la pression sélective sur l'insertion d'IES , ainsi que le biais de taille des IES ayant une longueur 3n sans codon terminateur en phase. J'ai participé à l'écriture de l'article.

OPEN ACCESS Freely available online

PLOS | GENETICS

The *Paramecium* Germline Genome Provides a Niche for Intragenic Parasitic DNA: Evolutionary Dynamics of Internal Eliminated Sequences

Olivier Arnaiz[1,2,3], Nathalie Mathy[1,2,3], Céline Baudry[1,2,3], Sophie Malinsky[4,5,6], Jean-Marc Aury[7], Cyril Denby Wilkes[1,2,3], Olivier Garnier[4,5,6], Karine Labadie[7], Benjamin E. Lauderdale[8], Anne Le Mouël[4,5,6¤], Antoine Marmignon[1,2,3], Mariusz Nowacki[9], Julie Poulain[7], Malgorzata Prajer[10], Patrick Wincker[7,11,12], Eric Meyer[4,5,6], Sandra Duharcourt[13], Laurent Duret[14], Mireille Bétermier[1,2,3]*, Linda Sperling[1,2,3]*

1 CNRS UPR3404 Centre de Génétique Moléculaire, Gif-sur-Yvette, France, **2** Département de Biologie, Université Paris-Sud, Orsay, France, **3** CNRS FRC3115, Centre de Recherches de Gif–sur-Yvette, Gif-sur-Yvette, France, **4** Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, Paris, France, **5** INSERM, U1024, Paris, France, **6** CNRS, UMR 8197, Paris, France, **7** Commissariat à l'Energie Atomique (CEA), Institut de Génomique (IG), Genoscope, Evry, France, **8** Methodology Institute, London School of Economics, London, United Kingdom, **9** Institute of Cell Biology, University of Bern, Bern, Switzerland, **10** Department of Experimental Zoology, Institute of Systematics and Evolution of Animals, Polish Academy of Sciences, Krakow, Poland, **11** Centre National de Recherche Scientifique (CNRS), UMR 8030, CP5706, Evry, France, **12** Université d'Evry, Evry, France, **13** Institut Jacques Monod, CNRS, UMR 7592, Université Paris Diderot, Sorbonne Paris Cité, Paris, France, **14** Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne, France

Abstract

Insertions of parasitic DNA within coding sequences are usually deleterious and are generally counter-selected during evolution. Thanks to nuclear dimorphism, ciliates provide unique models to study the fate of such insertions. Their germline genome undergoes extensive rearrangements during development of a new somatic macronucleus from the germline micronucleus following sexual events. In *Paramecium*, these rearrangements include precise excision of unique-copy Internal Eliminated Sequences (IES) from the somatic DNA, requiring the activity of a domesticated *piggyBac* transposase, PiggyMac. We have sequenced *Paramecium tetraurelia* germline DNA, establishing a genome-wide catalogue of ~45,000 IESs, in order to gain insight into their evolutionary origin and excision mechanism. We obtained direct evidence that PiggyMac is required for excision of all IESs. Homology with known *P. tetraurelia* Tc1/mariner transposons, described here, indicates that at least a fraction of IESs derive from these elements. Most IES insertions occurred before a recent whole-genome duplication that preceded diversification of the *P. aurelia* species complex, but IES invasion of the *Paramecium* genome appears to be an ongoing process. Once inserted, IESs decay rapidly by accumulation of deletions and point substitutions. Over 90% of the IESs are shorter than 150 bp and present a remarkable size distribution with a ~10 bp periodicity, corresponding to the helical repeat of double-stranded DNA and suggesting DNA loop formation during assembly of a transpososome-like excision complex. IESs are equally frequent within and between coding sequences; however, excision is not 100% efficient and there is selective pressure against IES insertions, in particular within highly expressed genes. We discuss the possibility that ancient domestication of a *piggyBac* transposase favored subsequent propagation of transposons throughout the germline by allowing insertions in coding sequences, a fraction of the genome in which parasitic DNA is not usually tolerated.

Citation: Arnaiz O, Mathy N, Baudry C, Malinsky S, Aury J-M, et al. (2012) The *Paramecium* Germline Genome Provides a Niche for Intragenic Parasitic DNA: Evolutionary Dynamics of Internal Eliminated Sequences. PLoS Genet 8(10): e1002984. doi:10.1371/journal.pgen.1002984

Editor: Harmit S. Malik, Fred Hutchinson Cancer Research Center, United States of America

Received March 16, 2012; **Accepted** August 9, 2012; **Published** October 4, 2012

Funding: This work was supported by the ANR BLAN08-3_310945 "ParaDice," the ANR 2010 BLAN 1603 "GENOMAC," a CNRS ATIP-Plus grant to MB (2010–2011), and an "Equipe FRM" grant to EM. The sequencing was carried out at the Genoscope - Centre National de Séquençage (Convention GENOSCOPE-CEA number 128/AP 2007_2008/CNRS number 028666). CDW and AM were supported by Ph.D. fellowships from the Ministère de l'Enseignement Supérieur et de la Recherche. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mireille.betermier@cgm.cnrs-gif.fr (MB); linda.sperling@cgm.cnrs-gif.fr (LS)

¤ Current address: UMR7216 Epigénétique et Destin Cellulaire, CNRS, Université Paris-Diderot/Paris 7, Paris, France

Introduction

Paramecium belongs to the ciliate phylum, a deep radiation of highly diverse unicellular eukaryotes. The hallmark of ciliates is nuclear dimorphism: each unicellular organism harbors two kinds of nuclei with distinct organization and function. A diploid "germline" micronucleus (MIC) undergoes meiosis and transmits the genetic information to the next sexual generation but is not expressed. A polyploid "somatic" macronucleus (MAC) contains a version of the genome streamlined for gene expression and determines the phenotype. A new MAC is formed at each sexual generation by programmed rearrangements of the entire zygotic, germline-derived genome, and the maternal MAC is lost. The MAC genome of *P. tetraurelia* has been sequenced [1] revealing a

Author Summary

Ciliates are unicellular eukaryotes that rearrange their genomes at every sexual generation when a new somatic macronucleus, responsible for gene expression, develops from a copy of the germline micronucleus. In *Paramecium*, assembly of a functional somatic genome requires precise excision of interstitial DNA segments, the Internal Eliminated Sequences (IES), involving a domesticated *piggyBac* transposase, PiggyMac. To study IES origin and evolution, we sequenced germline DNA and identified 45,000 IESs. We found that at least some of these unique-copy elements are decayed Tc1/mariner transposons and that IES insertion is likely an ongoing process. After insertion, elements decay rapidly by accumulation of deletions and substitutions. The 93% of IESs shorter than 150 bp display a remarkable size distribution with a periodicity of 10 bp, the helical repeat of double-stranded DNA, consistent with the idea that evolution has only retained IESs that can form a double-stranded DNA loop during assembly of an excision complex. We propose that the ancient domestication of a *piggyBac* transposase, which provided a precise excision mechanism, enabled transposons to subsequently invade *Paramecium* coding sequences, a fraction of the genome that does not usually tolerate parasitic DNA.

series of whole genome duplications (WGDs) in the lineage that provide a unique tool for evolutionary analyses.

Ciliate genome rearrangements and their epigenetic control by non-coding RNAs have been recently reviewed [2–4]. In *Paramecium*, genome rearrangements involve (i) endoreplication of the DNA to about 800 haploid copies, (ii) imprecise elimination of genomic regions that contain, in particular, transposons and other repeated sequences, usually leading to chromosome fragmentation and (iii) elimination of Internal Eliminated Sequences (IES) by a precise mechanism. The accuracy of this process is crucial for IESs located within coding regions, to correctly restore open reading frames. The characterization of fewer than 50 IESs identified by cloning MIC loci [5] showed that they are short (26–883 bp), unique copy elements that are located in both coding and non-coding regions of the genome. The IESs are invariably flanked by two TA dinucleotides whereas only one TA is found at the MAC chromosome junction after IES excision (Figure 1). IESs have also been discovered by *cis*-acting mendelian mutations that prevent their excision, conferring a mutant phenotype [6–10]. The mutations in almost all cases were found in one of the flanking TA dinucleotides, which seem to be an absolute sequence requirement for IES excision. Extrapolation of the number of IESs found mainly in surface antigen genes led to the estimation that there could be as many as 50,000 IESs in the *Paramecium* genome. Such massive presence of unique copy IESs inserted in genes is not a characteristic of all ciliates. The estimated 6,000 IESs of the related oligohymenophorean ciliate *Tetrahymena* [11] are excised by an imprecise mechanism [12], are usually multi-copy including recognizable transposons [13–15] and are rarely found in coding sequences [16,17].

Klobutcher and Herrick [18] first reported a weak consensus at the ends of 20 IESs from *Paramecium* surface antigen genes (5′-TAYAGYNR-3′) that resembles the extremities of Tc1/mariner transposons. These authors hypothesized a "transposon link" to explain the origin of IESs, suggesting that they are the decayed relics of a Tc1/mariner transposon invasion and that they are excised from the MAC DNA by a Tc1/mariner transposase encoded by a gene that has become part of the cellular genome [19]. In this model, IES excision represents the exact reversal of Tc1/mariner transposon integration into its TA target site with duplication of the TA dinucleotide, an evolutionary novelty that may have appeared more than once in the ciliate phylum. One problem with the model is that transposition catalyzed by Tc1/mariner transposases usually leaves a 2 or 3 bp "footprint" at the donor site [20] while IES excision is precise.

A decisive step towards understanding the mechanism of IES excision and validating a transposon link for the origin of the IES excision machinery was the identification of a domesticated *piggyBac* transposase in *Paramecium* [21]. Baptized PiggyMac (Pgm), the protein is encoded by the *PGM* gene which is expressed only late in sexual processes, at the time of genome rearrangements. Pgm, localized in the developing new MAC, was found to be required for the excision of all IESs tested and for the imprecise elimination of several regions containing transposons or cellular genes [21]. A similar *piggyBac*-derived transposase is found in *Tetrahymena* and is required for heterochromatin-dependent DNA elimination [22]. Since the *Paramecium* and *Tetrahymena* proteins appear to be monophyletic, based on a broad phylogeny of *piggyBac* transposases (L. Katz and F. Gao, personal communication), the domestication event may have preceded the divergence of these two ciliates, estimated at 500–700 Ma (million years ago) [23]. Most significantly, the *in vivo* geometry of IES excision, initiated by staggered double-strand breaks (DSBs) that generate 4-base 5′ overhangs centered on the TA at both ends of the IES

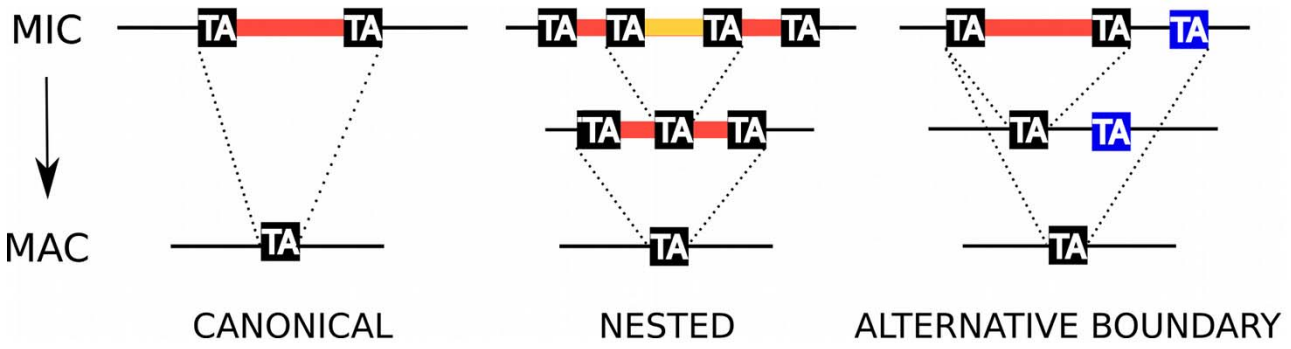


Figure 1. IES excision. Schematic representation of, from left to right, a canonical IES, a nested IES and an IES with an alternative boundary. In the case of the nested IES, the middle line represents either an intermediate in the excision pathway or an alternative final product. In the case of the alternative boundary IES, the middle line represents an alternative final product.
doi:10.1371/journal.pgen.1002984.g001

[24], is fully compatible with the *in vitro* reaction catalyzed by a *piggyBac* transposase isolated from an insect [25], whose target site is a 5′-TTAA-3′ tetranucleotide. *piggyBac* elements leave behind no scar when they jump to a new location: only ligation is required to join the fully complementary 5′ overhangs. Limited processing of 5′ and 3′ ends is further required for precise closure of the *Paramecium* IES excision sites since only the TA dinucleotides at the center of the 4-base 5′ overhangs are always complementary [24,26].

We report here a genomic approach to exhaustively catalogue the IESs in the *Paramecium tetraurelia* germline genome in order to study their evolutionary dynamics and seek evidence for a transposon origin of these elements. We obtained DNA highly enriched in un-rearranged germline sequences, from cells depleted in Pgm by RNA interference. Deep-sequencing of this DNA (hereafter called "PGM DNA") allowed us to identify a genome-wide set of nearly 45,000 IESs, by comparing contigs assembled using the PGM DNA (hereafter called "PGM contigs") with the MAC reference genome [1]. The hypothesis that Pgm is required for excision of all IESs was tested by genome-scale sequencing of a source of DNA from purified MICs [27], providing validation of the IES catalogue. The evolutionary dynamics of the IESs was studied by exploiting the series of WGDs that have been characterized in *Paramecium* [1]. The study provides, to our knowledge, the first genome-wide set of IESs, in *Paramecium* or any ciliate, and provides new evidence that IESs have deleterious effects on fitness and that at least a fraction of IESs do derive from Tc1/mariner transposons that have decayed over time. The IES sequences evolve rapidly. The constraints we could detect concern their size distribution, suggestive of the assembly of a transpososome-like excision complex and a weak consensus at their ends, which resembles the extremities of Tc1/mariner elements. We discuss the possibility that ancient domestication of the Pgm transposase favored subsequent propagation of transposons throughout the *Paramecium* germline genome, by providing a mechanism for their precise somatic excision, therefore allowing insertions in coding sequences.

Results

IES identification

An overview of the strategy for identification of a genome-wide set of IESs is presented in Figure S1. The first step was next-generation deep sequencing of DNA enriched in un-rearranged sequences, isolated from strain 51 cells that had undergone the sexual process of autogamy after depletion of Pgm protein by RNAi (Figure S2). In the absence of Pgm, the zygotic DNA is amplified but rearrangements are impaired. The sample that was sequenced contained a mixture of 60–65% un-rearranged DNA from the developing new MACs and 35–40% rearranged DNA from the fragments of the maternal MAC still present in the cytoplasm, as judged by Southern blot quantification of MIC and MAC forms at one locus (Figure S3). The PGM sequence reads (Table 1) were mapped to the MAC reference genome of strain 51 (see Materials and Methods), and putative IES insertion sites were defined as sites with a local excess of ends of read alignments (pipeline MIRAA for "Method of Identification by Read Alignment Anomalies"). This excess of ends of alignments arises when a read contains a MIC IES junction, since only part of such a read can align with a MAC chromosome, either starting or ending at the IES insertion site, expected to be a TA dinucleotide. Using MIRAA, we identified 45,739 potential IES insertion sites. Essentially all (99%) of the insertion sites contained a TA dinucleotide, even though this was not assumed by the pipeline.

In order to obtain the sequence of the IESs, the paired-end PGM DNA sequence reads were assembled into contigs (cf. Table S1 for assembly statistics) and compared to the MAC reference genome assembly (pipeline MICA for "Method of Identification by Comparison of Assemblies"). We looked for insertions in the PGM contigs with respect to the MAC reference assembly. Any insertion bounded by TA dinucleotides after local realignment was considered to be an IES. Using this pipeline we identified 44,928 IESs. The fact that 96% (n = 43,220) of the IESs identified by MICA correspond to an IES insertion site identified by MIRAA (Figure S1) testifies to the overall reliability of the procedure. Experimental validation of 6 IESs identified only by MICA and 17 insertion sites identified only by MIRAA was carried out by PCR amplification of an independent preparation of PGM DNA. The results (Table S2) show that the 6 IESs and at least 12 of 17 insertion sites tested do correspond to the presence of an IES. Interestingly, among the IES sites identified only by MIRAA, we found 8 examples of a pair of IESs separated by one or only a few nucleotides (in 5/8 cases, these tandem IESs are located in exons, a proportion similar to that found for the genome-wide IES set, see below). This case is not handled by the MICA pipeline since the initial global alignment with BLAT would have detected a single large insertion that would have been rejected by the local realignment filter, which requires the insertion to be flanked by TA dinucleotides. This is the first report of such closely spaced IESs, although nested IESs (Figure 1) have been previously documented [8].

In order to see whether the set of 44,928 IESs is likely to be exhaustive, we looked for the 53 previously characterized IESs identified directly by cloning MIC loci in *P. tetraurelia* strain 51 cells (Table S3). All 53 previously cloned IESs were found, with the exception of one IES that had been assembled into the MAC reference genome and one IES form that represents use of an alternative boundary. In addition, two small IESs, each of which is

Table 1. Sequencing and mapping statistics.

DNA	Insert size (bp)	Read length (bp)	Reads	Aligned reads	Aligned (%)	Genome coverage (%)
PGM	~500	108	130,266,728	110,189,736	84.6	99
Lambda-phage	~200	101	83,149,385	25,949,607	31	44

Paired-end Illumina sequencing was carried out as described in Materials and Methods, and reads were mapped to the *P. tetraurelia* MAC reference genome using the BWA short-read aligner. The genome coverage is the fraction of the genome covered by at least 1 read. The depth of coverage with the PGM DNA is on average 165×. The depth of coverage with the lambda-phage DNA is on average 75× for the part of the genome that is covered. The PGM reads that were not aligned contain *Paramecium* mitochondrial and rDNA sequences, contaminating bacterial sequences as well as sequences present only in the MIC genome. In addition, a large proportion of the unaligned lambda-phage reads are from bacterial contaminants with AT-rich genomes; this DNA was not eliminated by the cesium chloride density gradient separation step of the phage library construction [28].
doi:10.1371/journal.pgen.1002984.t001

nested within a larger IES, were found in PGM DNA but were not identified by our pipeline as IESs. Indeed, nested IESs can only be identified by time-course experiments or if the outer IES is retained in the MAC e.g. as the result of a point mutation [8]. Since 49 of 51 non-nested IESs were identified by MICA, the IES identification procedure has a sensitivity of at least 96%.

The entire IES identification approach is based on the assumption that the excision of all IESs in *Paramecium* requires the Pgm domesticated transposase activity. In order to test this assumption, we sequenced inserts from a lambda-phage library constructed some 20 years ago [28], using DNA from MICs that had been separated from MACs by Percoll gradient centrifugation [27]. This library has been extensively used to clone MIC loci with specific probes. Although the contigs assembled from the phage DNA reads only partially covered the MAC reference genome (Table 1), 98.5% of the 13,377 IESs that could be identified using the phage DNA and the MICA pipeline had also been identified using the PGM DNA. The difference of 1.5% is within the estimated sensitivity of the MICA pipeline. We conclude that all *Paramecium* IESs very likely require Pgm for excision, and that our data set does represent a genome-wide set of *P. tetraurelia* IESs.

IES distribution in the genome

The genome-wide set of IESs has an overall G+C content of 20%, significantly lower than the 28% G+C content of the MAC reference genome [29] but comparable to the G+C content of intergenic regions (21%). The IESs are found in exons (76.8%), introns (5.4%) and intergenic regions (17.8%), suggesting a nearly random distribution of IESs with respect to genes, since the MAC reference genome is composed of 76% exons, 3.2% introns and 20.8% intergenic DNA [1]. However, IESs are not randomly distributed along the chromosomes. Intriguingly, as shown in Figure S4 for the 8 largest MAC chromosomes, IESs tend to be asymmetrically distributed along MAC chromosomes. The MAC assembly (188 scaffolds >45 Kb constitute 96% of the 72 Mb assembly) contains 115 telomere-capped scaffolds, varying in size from ~150 Kb to ~1 Mb, that are considered to represent complete MAC chromosomes. For 70 of these telomere-capped scaffolds, IESs display non-uniform distributions ($p<0.002$, median scaffold size 417 Kb) while for the remaining 45 telomere-capped scaffolds, the IES distribution is uniform (median scaffold size 275 Kb). Thus the larger the MAC chromosome, the greater the chance of observing a non-uniform IES distribution. The distributions for all scaffolds are easily visualized using the ParameciumDB [30] Genome Browser. The significance of the asymmetry in IES distribution is not clear, but might be related to the global organization of MIC chromosomes, currently unknown (discussed in [29]).

Germline Tc1/mariner transposons

The genome-wide set of IESs covers 3.55 Mb (mean IES size 79 bp), compared to 72 Mb for the MAC reference genome assembly. The IESs thus add about 5% to the sequence complexity of the part of the MIC genome that is collinear with MAC chromosomes. The total complexity of the PGM contigs (after elimination of contigs with low PGM read coverage and high G+C content, assumed to represent bacterial contamination as confirmed in many cases by BLASTN matches against bacterial genomes) is ~100 Mb, however the use of a single paired-end sequencing library with small inserts (~500 bp) may have perturbed assembly of repeated sequences, possibly leading to underestimation of repeated sequence content. We infer that ~25 Mb of germline-specific DNA corresponds to the imprecisely eliminated regions located outside of the MAC-destined chromosomes i.e. the part of the MIC genome that is not collinear with MAC chromosomes.

We have not further characterized this fraction of the PGM DNA. However, we did identify the first germline *P. tetraurelia* Tc1/mariner transposons (Figure S5), by using the phage-lambda library of MIC DNA [28] to walk past the end of MAC scaffold_51, which bears the subtelomeric 51G surface antigen gene [31]. In all, 5 phage inserts and 4 cloned PCR products corresponding to part or all of different copies of the element downstream of the 51G surface antigen gene, named *Sardine*, were sequenced (EMBL Nucleotide Sequence Database accession numbers HE774468–HE774475) and a consensus for the ~6.7 Kb transposon was constructed (Figure S5 and Text S1). The ends of the *Sardine* copies contain intact or partially deleted 425 bp terminal inverted repeats (TIRs) which are themselves palindromic, containing a unique, oriented region nested within outer inverted repeats (Figure S5). *Sardine* contains up to 4 ORFs. One ORF is a putative DD35E transposase of the IS*630*-Tc1 family, like the DDE transposases of the TBE and Tec transposons found in stichotrich ciliates [32]. Another ORF, as in Tec transposons [33], encodes a putative tyrosine recombinase. The other two ORFs are hypothetical, though ORF2 shows some similarity (31.7% identity and 55.4% similarity over 202 aa) to the hypothetical ORF1 of the *Tennessee* element from *P. primaurelia* [34]. One of the *Sardine* copies (copy S6) is interrupted by the insertion, within the putative tyrosine recombinase gene, of a different but similar element, named *Thon* (French for "tuna"), which also contains a DD35E transposase, a tyrosine recombinase, possibly the two hypothetical ORFs, and palindromic TIRs of ~700 bp (Figure S5).

IES copy number and similarity to transposon sequences

For a handful of IESs, it has been shown experimentally that they are single copy elements [5]. In order to see whether this is generally the case, we looked for all IESs present in more than 1 fully identical copy (100% sequence identity). We found 44,210 IESs to be unique copy (98.4%). We examined all IESs present in 2 or more identical copies and found 39 cases of duplicate IESs as a result of errors in assembly of the MAC reference genome that had led to small, partially redundant scaffolds (4% of the MAC assembly is contained in scaffolds <45 Kb and some of these are partially redundant with the chromosome-size scaffolds [1]). The rest of the 319 IESs found in 2 copies were inserted in homologous genomic sites and appeared to be the result of recent segmental duplication or gene conversion. The 23 cases of IESs found in 3 to 6 copies correspond to expansion or recombination of repeated sequences such as tetratricopeptide repeat (TPR) domains or WD40 repeats.

We performed an all by all sequence comparison of the IESs and of their flanking sequences to see whether we could identify homologous IESs inserted at non-homologous sites in the genome. As shown in Table 2, we were able to identify 8 clusters of 2 to 6 IESs that share significant homology (BLASTN E-value $<10^{-10}$) over at least 85% of their length, inserted in non-homologous sites (cf. Text S2 for the alignments). Moreover, we found significant nucleotide identity (E-value 9×10^{-57} for the best match; nucleotide identity between 68 and 78% for the HSPs) between the IESs of cluster 5 and one of the Tc1/mariner-like transposons identified using the phage library (*Thon*, Figure S5). This is a strong indication that these IESs are derived from recently mobile elements.

However, the IES sequences of this cluster correspond to a single palindromic TIR. This might reflect assembly problems given use of a single insert size for the paired-end sequencing,

Table 2. Homologous IESs at non-homologous sites in the genome.

CLUSTER	IES SCAFFOLD	IES POSITION	SIZE (bp)	LOCATION	NUCLEOTIDE MATCH
2	scaffold51_25	381101	209	GSPATP00009750001	
	scaffold51_25	389332	213	intergenic	
3	scaffold51_117	944	608	intergenic	
	scaffold51_160	10020	577	intergenic	
	scaffold51_44	7711	555	intergenic	
5	scaffold51_109	40673	571	intergenic	TIR *Thon*
	scaffold51_128	266698	689	GSPATP00032295001	TIR *Thon*
	scaffold51_131	262422	630	intergenic	TIR *Thon*
	scaffold51_18	127217	770	GSPATP00007326001	TIR *Thon*
	scaffold51_34	280841	512	intergenic	TIR *Thon*
	scaffold51_58	302214	640	intergenic	TIR *Thon*
9	scaffold51_19	475992	666	intergenic	
	scaffold51_96	236752	665	intergenic	
12	scaffold51_124	248174	568	intergenic	
	scaffold51_27	275392	476	GSPATP00010339001	
13	scaffold51_155	211807	458	intergenic	
	scaffold51_20	46790	505	intergenic	
	scaffold51_27	294496	472	GSPATP00010351001	
14	scaffold51_184	21279	1024	GSPATP00038454001	
	scaffold51_21	430950	1038	GSPATP00008497001	
	scaffold51_58	200038	1010	GSPATP00018841001	
15	scaffold51_28	278632	262	GSPATP00010625001	
	scaffold51_4	361312	242	GSPATP00001801001	

A BLASTN internal comparison of all IESs, carried out with an E-value cutoff of 1e-10, was filtered for HSP coverage of at least 85% of the longest IES and for the absence of significant homology between 500 nt of MAC flanking sequence. The IESs were than transitively clustered and aligned using MUSCLE (Text S2). Some clusters were eliminated because of low complexity of the IES sequences. BLASTN homology searches at NCBI and against known Paramecium transposons ([34] and the present manuscript) were carried out using each IES in the clusters as query. *Thon* is a Tc1/mariner-like transposon. BLASTX similarity searches against the non-redundant protein database at NCBI did not yield any significant hits at an E-value cutoff of 0.001. The location of the IES, if in a coding sequence, is provided as a ParameciumDB accession number.

doi:10.1371/journal.pgen.1002984.t002

either because these IESs contain sequences repeated elsewhere in the genome or because the *Thon* TIRs are large (~700 bp) and palindromic so that the assembly might have jumped from one TIR to the other deleting the rest of *Thon*. We therefore used a long-range PCR strategy capable of amplifying large DNA fragments containing each of the IESs to verify their size and attempt to obtain sequences (detailed in Text S3). Amplification products of the expected sizes were obtained for all of the IESs from cluster 5, making it unlikely that these IESs correspond to a complete *Thon* element that had failed to be assembled from the paired-end sequencing reads. Three IESs were chosen for sequencing, and the sequences of the corresponding PCR products confirmed the IESs, indicating that they had been correctly assembled. Identification of 6 IESs (at non-homologous genomic sites) that share sequence identity with a *P. tetraurelia* Tc1/mariner solo TIR argues that at least a fraction of IESs do originate from Tc1/mariner-like elements.

We therefore adopted a complementary strategy, using the PFAM-A library of curated protein domains to search for domain signatures in the genome-wide set of IESs. Matches at a BLASTX E-value cutoff of 1 were inspected visually to filter out matches with PFAM-A protein domains from *Paramecium* and matches owing to compositional bias (high A+T content). This left 6 IESs, ranging in size from 2416 to 4154 bp, with a DDE_3 (PFAM accession number 13358) DDE superfamily endonuclease domain characteristic of IS*630*/Tc1 transposons. The peptides encoded by the IESs were subjected to an HMM search of the PFAM-A hmm profiles (http://pfam.sanger.ac.uk/search) for confirmation of the conserved residues and to validate the statistical significance of the match (E-values of 0.02 to 2.1×10^{-15} for the 6 peptides). The IESs were aligned with MUSCLE and a neighbor-joining tree grouped 4 of them together with good bootstrap values (not shown). The 4 IESs were used to search for sequence similarity with the genome-wide set of IESs and this allowed identification of 28 IESs ranging in size from 1251 to 4154 bp (Table S4). The IESs were aligned to provide the consensus sequence for 2 distinct Tc1/mariner-like 3.6 kb transposons from the same new family, baptized *Anchois* (Anchovy). Manually adjusted alignments used to reconstruct the *AnchoisA* and *AnchoisB* elements, consensus sequences and annotation are provided in Text S1.

Alignment of the DDE domains of the reconstituted *Anchois* transposons with the DDE domains from bacterial IS*630* elements, invertebrate Tc1 transposons and all known ciliate Tc1/mariner elements indicates that the *Anchois* transposase belongs to the IS*630*/Tc1 subfamily (Figure 2A). Unlike *Thon* and *Sardine* but like the *P. primaurelia Tennessee* element, *Anchois* TIRs are short and lack internal palindromes, moreover *Anchois* does not contain a putative tyrosine recombinase. *Anchois* has 2 hypothetical

ORFs in addition to the DDE transposase (Figure 2B; Text S1). The ORF2 of *Anchois* displays homology to ORF2 of *Sardine* (36.2% identity and 56.2% similarity over 210 aa) and to ORF1 of *Tennessee*. Interestingly, for 6 of the 28 IESs that initially identified the copies of *Anchois*, the *Anchois* TIRs do not correspond to the extremities of the IES, raising the possibility of *Anchois* insertions within pre-existing IESs. The discovery of the *Anchois* elements and the fact that several IESs appear to be full-length copies, provides a strong, direct link between IESs and transposons.

A remarkable IES size distribution

The size distribution of the genome-wide set of IESs is shown in Figure 3A, for the 93% of the IESs that are shorter than 150 bp. The most remarkable feature is a periodicity of ~10 bp, which corresponds to the helical repeat of double-stranded DNA. The first peak of the size distribution has maximal amplitude at 26–28 bp and includes 35% of all identified IESs. The abrupt cutoff at 26 bp represents the minimum IES size. A second peak appears to be forbidden and contains only a few IESs. The following peaks are centered at approximately 45–46, 55–56, 65–66 bp etc. and the distance between these peaks is best fit by a 10.2 bp sine wave (not shown). At the far end of the spectrum, 95 of the IESs are between 2 and 5 Kb in size. Similar periodic size distributions are found for IESs inserted in coding sequences and for IESs inserted in non-coding sequences (Figure S6). This indicates that the constraint on the distance between IES ends is

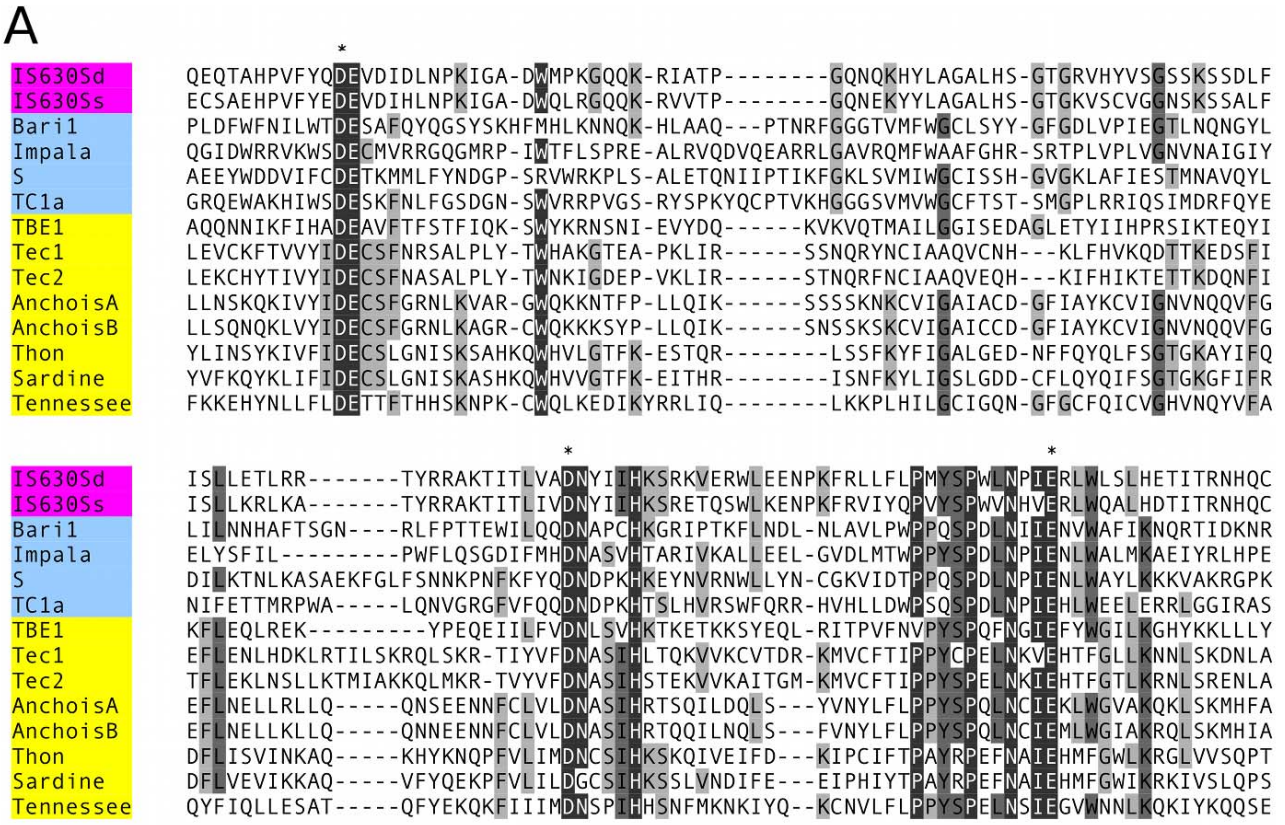


Figure 2. *Anchois* Tc1/mariner family transposon. A) Alignment of the DDE domains of bacterial IS*630* elements (IS*630*Sd, *Salmonella dublin*, GenBank Accession No. A43586; IS*630*Ss, *Shigella sonnei*, X05955), invertebrate and fungal Tc1 transposons (Baril, *D melanogaster*, Q24258; Impala, *Fusarium oxysporum*, AF282722; S, *D* melanogaster, U33463; Tc1, *C elegans*, X01005) and ciliate Tc1/mariner transposons (TBE1, *Oxytricha fallax*, L23169; Tec1 and Tec2, *Euplotes crassus*, L03359 and L03360; Anchois, Thon and Sardine, *Paramecium tetraurelia*, this article; Tennessee, *Paramecium primaurelia*, [34]). Asterisks mark the conserved catalytic DDE residues. B) Schematic diagram of the 3.6 Kb *Anchois* consensus, showing the position and orientation of the 3 ORFs. The yellow triangles represent the ~22 nt TIRs. Asterisks mark the position of residues of the catalytic DDE triad for the ORF encoding the DDE transposase.
doi:10.1371/journal.pgen.1002984.g002

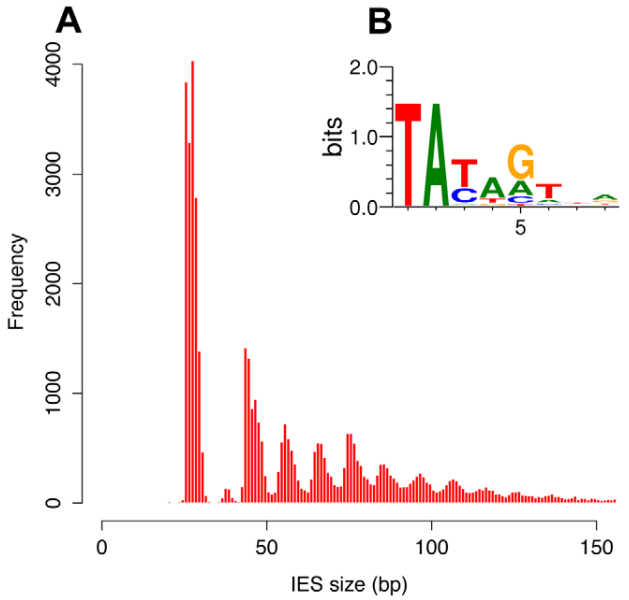


Figure 3. IES sequence properties. A) Histogram of the sizes of the genome-wide set of IESs that are shorter than 150 bp. B) Sequence logo showing information content at each position, corrected for a G+C content of 28%, for the ends of the genome-wide set of IESs.
doi:10.1371/journal.pgen.1002984.g003

an intrinsic property of the IESs and is not related to the locus in which they are inserted in the genome. Whatever their size, the IESs adhere to the weak, Tc1/mariner-like end consensus first reported for 20 IESs located in surface antigen genes [18], as illustrated in Figure 3B for the whole set. Differently sized subsets of the IESs all display essentially the same end consensus (data not shown).

We further examined constraints on IES size and sequence by evaluating IES conservation with respect to the 3 WGDs in the *Paramecium* lineage. We used the large number of paralogs (hereafter termed "ohnologs") of different ages (Table 3) that could be identified for each of the WGD events [1] to ask whether IESs are present, at the same position relative to the gene coding sequences, in ohnologs of the different WGD events. This analysis makes the assumption that IES insertions are rare events so that if IESs are present at the same position in ohnologous genes, then they must have been acquired before the WGD and can be considered to be "ohnologous" IESs. As shown in Table 3, we found 84.5%, 23.2% and 5.9% conservation of IESs with respect to the recent, intermediate and old WGDs respectively. For comparison, more than 99% intron conservation was found for 1,112 pairs of genes related by the recent WGD [35]. This indicates that the dynamics of IES insertion or loss over evolutionary time is relatively fast compared to that of introns. The only phylogenetic study of IESs, carried out for two loci in a few different stichotrich (formerly called hypotrich) ciliates, which are very distantly related to *Paramecium*, also concluded that the intragenic IESs in those species evolve very rapidly [36]. We found that the ohnologous IESs related by the recent WGD are highly divergent in sequence. In more than 90% of cases, the sequence identity was too low for detection by BLASTN (E-value threshold of 10^{-5}). This high level of sequence divergence is consistent with the pattern expected for neutrally-evolving non-coding regions, since the average synonymous substitution rate measured between ohnologous genes derived from the recent WGD is about 1 substitution per site [1]. However, if we compare the lengths of IESs that are conserved with respect to the recent WGD (Figure 4A), for ~55% of the pairs, both IESs are found in the same peak of the IES size distribution. The honeycomb appearance of the plot (Figure 4A), with off diagonal cells that result from ohnologous IESs in different peaks of the distribution, underscores the strong evolutionary constraint that is exerted on IES size.

Table 3. IES conservation in ohnologs produced by the different WGDs.

WGD event	Genes with ohnolog	IESs	Conserved IESs	% conserved
Recent	24052	20623	17430	84.5
Intermediate	12590	11561	2675	23.2
Old	3381	3646	215	5.9

The identification of ohnologs and the reconstitution of the pre-duplication genomes is described in [1]. For the most recent WGD, which preceded the appearance of the *P. aurelia* complex of 15 sibling species [95], 51% of the pre-duplication genes are still present in 2 copies. For the intermediate duplication, 24% of pre-duplication genes are still present in 2 or more copies. For the ancient duplication, which preceded the divergence of *Paramecium* and *Tetrahymena*, 8% of pre-duplication genes are still present in 2 or more copies. The significance of the column headers is as follows. Genes with ohnolog: the number of present day genes with at least one ohnolog from the indicated WGD event. IESs: the number of IESs found in the genes with at least one ohnolog from the indicated WGD event. Conserved IESs: number of IESs found at the same position in at least one other ohnolog, as determined by sequence alignment. The identification of ohnologs is described in [1] and the data are available through ParameciumDB [90]. Note that this analysis only concerns IESs that are within paralogous genes and not IESs found in intergenic regions.
doi:10.1371/journal.pgen.1002984.t003

Dynamics of IES gain and loss

In order to investigate the rate of IES insertions and losses during the evolution of the *Paramecium* lineage, we examined gene families, which we call "quartets", for which all 4 ohnologs issued from duplication of an ancestral gene at the intermediate and then the recent WGD are still found in the present day genome. Of the 1350 such quartets identified in the MAC genome [1], 878 contain at least one IES in at least one of the 4 duplicated genes. We evaluated the conservation of IESs at the same position with respect to the coding sequence for all members of each quartet (Figure S7), and identified 2126 IES groups, each group containing an IES conserved either in all 4 genes ($N_{1111} = 190$), in 3 genes ($N_{1110} = 64$), in 2 genes on the same intermediate WGD branch ($N_{1100} = 1304$), in 2 genes on different branches ($N_{1010} = 10$) or in only one of the 4 genes ($N_{1000} = 558$).

Under the assumption that two IESs present at the same location in ohnologous genes derive from a single ancestral IES (i.e. the probability of two insertion events occurring at the same site after a WGD is considered negligible), and that the rate of IES losses has remained constant, it is possible to estimate the rate of IES gain during the evolution of the *Paramecium* lineage (the model is developed in Text S4). The quartet analysis is fully consistent with a model whereby IES acquisition has been ongoing since before the intermediate WGD (15% of the IESs predating this WGD), with a peak in the period between the intermediate and the recent WGD events: 69% of IESs were acquired during the interval between these two WGDs, vs. 16% during the period since the recent WGD, which corresponds to about the same evolutionary time. Genome-wide IES data for other *Paramecium* species will be necessary in order to test the assumption of a

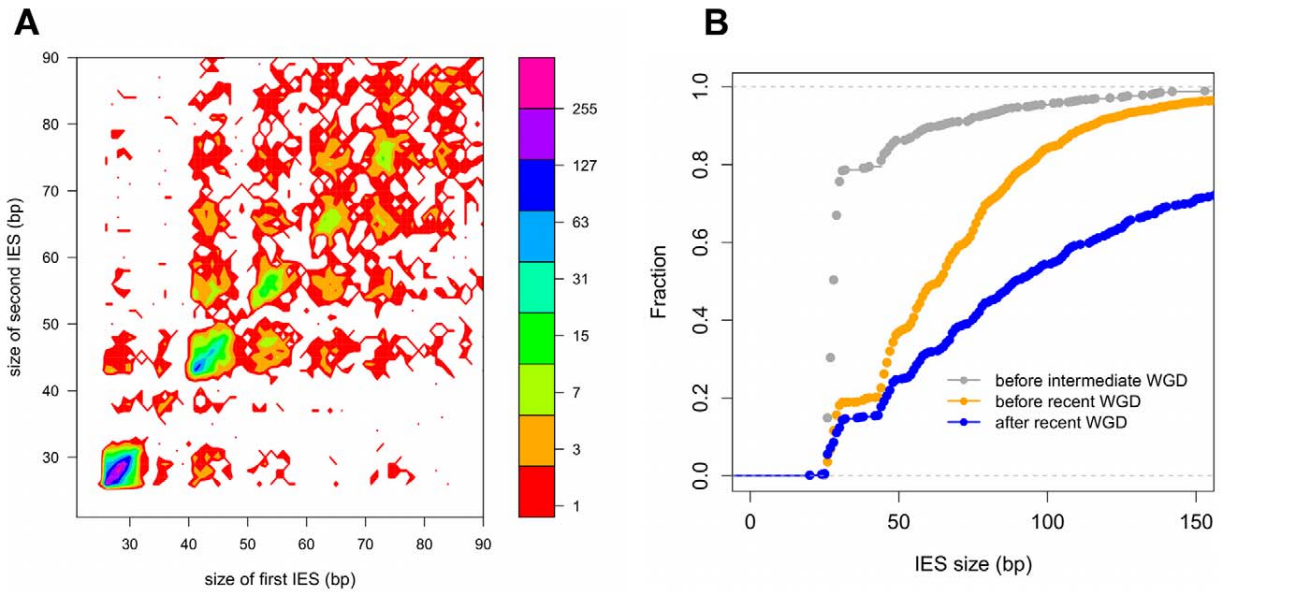


Figure 4. IES conservation in genes related by WGD. A) Filled contour plot of the correlation between the size of IES pairs that have been conserved with respect to the recent WGD. The x axis gives the size in bp of the first IES, the y axis gives the size in bp of the second IES found in the ohnologous gene and the color of each point indicates the number of times that combination of x,y values was found in the data set. The color legend is shown to the right of the figure, the numbers represent counts of the x,y value pairs; the rainbow colors are distributed according to a log2 scale. B) Size distribution of IESs conserved in "quartets" i.e. genes that are still present in 4 copies in the genome after duplication at both the intermediate and the recent WGD events. In order to compare size distributions for different classes of IES, they are represented as experimental cumulative distribution functions. The ripples in each curve correspond to the peaks of a histogram representation as in Figure 3A. The curves are for IESs that must have originated from an ancestral IES acquired before the intermediate WGD (grey, N_{1111} IESs), IESs that must have originated from an ancestral IES acquired before the recent WGD (orange, N_{1100} IESs) and the IESs that might have been acquired since the recent WGD (blue, N_{1000} IESs).
doi:10.1371/journal.pgen.1002984.g004

constant rate of IES losses. However, even if we relax this assumption (i.e. rates of IES losses are allowed to vary over time), the model still strongly rejects the hypothesis that all IESs were acquired before the intermediate WGD (cf. Text S4). Thus, with the presently available data and biologically reasonable assumptions, we conclude that IESs have been acquired in all 3 of the time periods delimited by the intermediate and recent WGD events.

We compared the cumulative size distributions of the N_{1111}, N_{1100} and N_{1000} IESs (Figure 4B). The N_{1111} IESs, which must have been acquired before the intermediate WGD, are much shorter than the IESs of the two other samples, with almost 80% of the IESs in the first peak, compared to 20% for N_{1100} IESs, which may mainly result from IES acquisition after the intermediate but before the recent WGD, and only 16% for N_{1000} IESs, at least some of which may have been acquired since the recent WGD. In addition, the curves are significantly shifted with respect to each other, in particular, 30% of the N_{1000} IESs are larger than 150 nt, compared to scarcely any IESs larger than 150 nt for the two other samples. This analysis shows that the older an IES, the shorter it is likely to be, consistent with a decay process involving progressive shortening of IESs by accumulation of small deletions, in addition to the accumulation of point mutations.

Quartet analysis is restricted to IESs in genes that have been retained in 4 copies (fewer than 10% of all IESs). Similar distributions of IES size are found if we consider all ohnologous IESs (45% of all IESs, cf. Table 3). IESs conserved with respect to the intermediate WGD (76% of IESs in first peak) are significantly shorter than IESs conserved only with respect to the recent WGD (30% of IESs in the first peak) (data not shown). The size distribution of IESs conserved with respect to the old WGD is poorly determined because of the small number of conserved IESs (Table 3), which are moreover often in genes that have undergone recent gene conversion judging from the nucleotide divergence of the ohnologs (data not shown). It is therefore uncertain that IESs were present in the genome before the old WGD, consistent with the absence of TA-bounded IESs in *Tetrahymena*, which diverged from *Paramecium* after the old WGD event [1].

Since we found essentially no IESs shorter than 26 bp, it seems likely that some mechanism(s) other than decay of the sequence through internal mutations and deletions is responsible for the complete loss of an IES. In order to explore this question, we examined case by case, using both nucleotide and conceptual protein alignments, all of the N_{1110} quartet IES groups (n = 64), which are most parsimoniously explained by insertion of an IES before the intermediate WGD followed by loss of an IES after the recent WGD. We examined the raw read alignments and PGM and phage contigs in order to be sure that there was sufficient read coverage and no evidence suggesting presence of an IES at any site of putative IES loss. We found 4 different explanations for the quartet triplets: precise loss of the fourth IES (n = 17), gain of the third IES by gene conversion between intermediate WGD ohnologs (n = 1), recruitment of the fourth IES into the exon sequence (n = 6), and deletion of the region that encompasses the fourth IES (n = 23), often testifying to the formation of a pseudogene. In addition, we found 5 errors in IES detection (the fourth IES probably exists as it can be found in the phage contigs or is predicted by the MIRAA pipeline). In the remaining cases (n = 12), annotation or alignment problems made it difficult to conclude. The observation of 17 cases of precise loss of an IES from the germline DNA raises the possibility that there is a mechanism for conversion of a MIC locus to the IES-free form

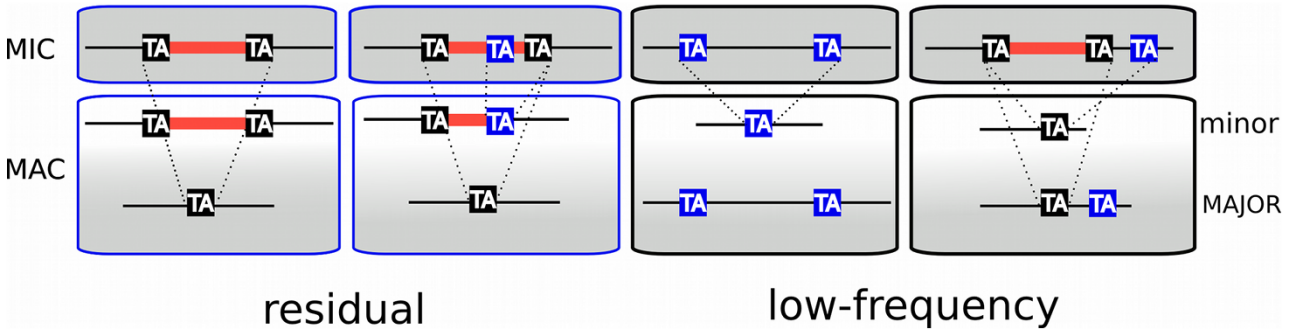


Figure 5. TA-indels are produced by IES excision errors. Schematic representation of the "residual" and "low frequency" TA-indels that were identified by comparing the MAC draft genome assembly (MAJOR form) with the 13× Sanger sequencing reads used to build the assembly [29]. The TA-indels were identified by one or more reads that differed from the assembly (minor form). The residual TA-indels were assumed to be the result of occasional failure to excise an IES and the low-frequency TA-indels to result from excision of MAC-destined sequences. Comparison of the genome-wide set of IESs with the TA-indels revealed that many TA-indels result from the use of alternative IES boundaries situated inside the corresponding IES in the case of residual TA-indels and outside the IES in the case of low-frequency TA-indels. In the schema, TA dinucleotides in black boxes are *bona fide* IES boundaries while TA dinucleotides in blue boxes are alternative IES boundaries.
doi:10.1371/journal.pgen.1002984.g005

using a MAC genome template. However, we cannot rule out the possibility that IESs can be precisely excised from the MIC DNA, and therefore lost, by the same Pgm-dependent mechanism as that involved in MAC genome assembly.

TA-indels reveal IES excision errors

The analysis of sequence variability in the polyploid (800n) MAC genome, carried out by comparing the MAC assembly representing a "consensus" sequence with the 13× Sanger sequencing reads used to build the assembly, revealed nearly 2000 "TA-indels" that were presumed to be produced by the IES excision machinery and to reflect excision errors [29]. As shown schematically in Figure 5, "residual" TA-indels (n = 739), that were suggested to represent occasional retention of IESs on some macronuclear copies, were absent from the assembly ("major" form in Figure 5), but present in at least one sequence read ("minor" form). For 689 of the residual TA-indels (93%), we found an IES at the corresponding site in the genome. Interestingly, in 134 cases (19.4%), the TA-indel was shorter than the IES and case by case inspection indicated that most of these TA-indels may be products of IES excision that used an alternative IES boundary located within the IES (Figure 5). In this case, the TA-indel would only correspond to part of a larger IES. A few cases of use of an alternative IES boundary that may confer a mutant phenotype have been reported [7,37].

"Low frequency" TA-indels (n = 1090), previously suggested to represent excision of MAC-destined sequences [29], were present in the assembly (major form, Figure 5), but absent from at least one sequence read (minor form). We could not look for the "low-frequency" TA-indels directly among the genome-wide set of IESs, since they are part of the MAC genome assembly. However, we examined the ends of the low-frequency TA-indels and found 249 cases (23%) where the TA dinucleotide at one of the ends corresponds to the insertion site of an IES in the genome-wide set (Figure 5), indicating that the TA-indel was generated by use of an alternative IES boundary located outside of the IES. The whole of the analysis supports the previous conclusion [29] that TA-indels are products of the IES excision machinery. The high incidence of alternative boundaries in both classes of TA-indels, revealed by comparing them with the genome-wide set of IESs, strengthens the previous conclusion [29] that TA-indels reflect IES excision errors (see below). Thus TA-indels cannot be considered to be IESs in the absence of further experimental support.

Evidence for selective pressure against IES insertion

IESs are tolerated in coding sequences and evolve under a strong constraint on their size and end-consensus, properties that are presumably important for their precise and efficient excision. However, the excision machinery can commit errors, as revealed by TA-indels (cf. above) and by the use of alternative IES

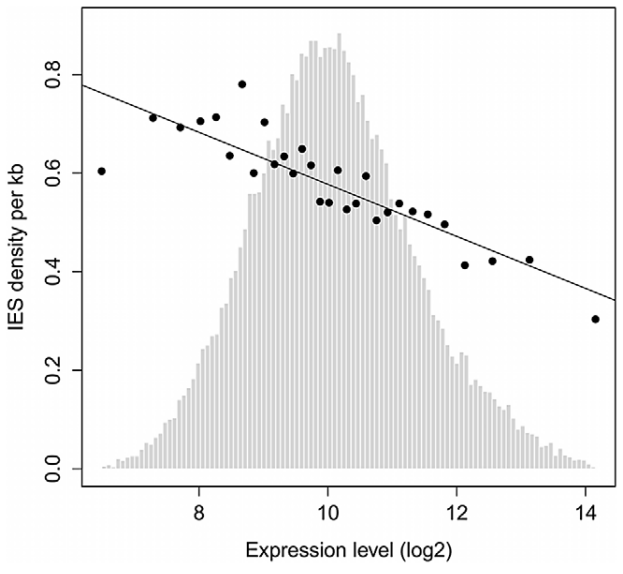


Figure 6. IES density is inversely proportional to gene expression level. Genes were binned according to their median expression level across 58 microarrays representing different cellular and growth conditions as described in [38,39]. The expression levels were divided into 30 bins as in [38]. The black points show the average IES density (per Kb) of genes in each bin. Linear regression was used to fit the points. Light gray bars show the distribution of genes according to their expression level (before binning).
doi:10.1371/journal.pgen.1002984.g006

Table 4. Deficit of 3n IESs in coding sequences.

IES Category	Number	3n	non-3n	χ^2	*P*-value
Non-coding	10304	3481 (33.78%)	6823 (66.22%)	-	-
Coding stopwith	11205	3700 (33.02%)	7505 (66.98%)	2.91	0.08
Coding stopless	23339	7095 (30.40%)	16244 (69.60%)	119.42	8.47×10^{-28}
Q1 stopless	6044	1892 (31.30%)	4152 (68.70%)	16.61	4.59×10^{-5}
Q4 stopless	5712	1615 (28.27%)	4097 (71.73%)	77.5	1.32×10^{-18}

For the calculation of χ^2, the observed numbers of IESs of length 3n and non-3n inserted in coding sequences are compared to the distribution found for IESs inserted in non-coding sequences under the null hypothesis that IES length is not under constraints related to translation. The null hypothesis is rejected only for those IESs inserted in coding sequences that do not contain a stop codon in frame with the upstream ORF (Sample "Coding stopless"). Microarray experiments [38] were used to group the IESs according to the expression level of the genes in which they are inserted. "Q1" designates IESs in exons of the 25% least expressed genes and "Q4" designates IESs in the exons of the 25% most expressed genes, those subject to the strongest selective pressure. The bias against 3n IESs is stronger in the Q4 sample than in the Q1 sample.
A more detailed analysis of the modulo 3 length distribution for IESs in coding and non-coding sequences, for each peak of the 10 bp periodic size distribution, is provided in Table S5.
doi:10.1371/journal.pgen.1002984.t004

boundaries [7,37]. We therefore looked for evidence that the rate of excision errors is high enough to represent a fitness burden for the organism. First, only 47% of genes contain at least one IES, and the IESs are less represented in strongly expressed genes. Figure 6 shows the density of IESs in genes as a function of gene expression level determined by microarray experiments [38,39]. The density varies from about 0.7 IESs per Kb (i.e. an IES on average every 1.4 Kb) in genes with low expression to less than 0.3 IESs per Kb (i.e. an IES on average every 3.3 Kb) for the genes with the highest expression. The inverse correlation observed across all levels of expression indicates that IESs are less-well tolerated the more a gene is expressed.

Second, IESs inserted in protein-coding exons display a characteristic bias in their size. There is a statistically significant deficit in IESs whose length is a multiple of 3, compared to IESs found in non-coding regions. Furthermore, this bias is only found for 3n IESs that do not contain a stop codon in phase with the ORF of the upstream coding sequence (Table 4; cf. Table S5 for a more detailed analysis). A similar 3n bias was reported for introns in eukaryotic genomes, and experiments in *Paramecium* showed that the Nonsense Mediated Decay (NMD) pathway destroys mRNAs containing unspliced introns, provided the intron retention leads to a premature stop codon [35]. Retention in mRNA of a 3n stopless intron would not be detected by NMD and therefore could lead to translation of potentially harmful proteins, explaining the deficit in 3n stopless introns. The fact that IESs display a similar deficit suggests that the rate of IES retention is high enough to represent a fitness cost, so that IESs in exons are under selective pressure to be detected by NMD in case they are retained in the MAC genome. We were able to test this hypothesis by looking at the size bias for IESs located in the exons of the 25% of *Paramecium* genes that are the most highly expressed hence subject to the strongest selective pressure. As shown by the last 2 lines of Table 4 (samples Q1 and Q4), the deficit in 3n IESs is the greatest for the IESs found in the most highly expressed genes (28.3%), where IES retention would be the most deleterious.

Discussion

An IES reference set for *P. tetraurelia*

Previous studies of *Paramecium* IESs all relied on a small reference set of about 50 IESs. For the first time in any ciliate genome, in so far as we are aware, we have carried out an exhaustive identification of IESs. Since it is not yet possible to isolate *Paramecium* MICs in the quantity and of the purity required for genomic sequencing, we relied on nuclear DNA isolated from cells depleted in Pgm, the domesticated transposase required for introduction of the DSBs that initiate IES excision [21]. We fortunately were able to use the only genomic library ever made from purified MICs [28] – but heavily contaminated by bacterial DNA – to obtain genome-scale evidence that Pgm is required for excision of all *Paramecium* IESs and to estimate that our IES reference set includes ~98.5% of all IESs.

Although this IES reference set will prove useful for a variety of studies, it is important to keep two things in mind. First, the IES definition used here is necessarily a genomic definition involving comparison of MIC and MAC sequences. Our procedure does not allow identification of nested IESs (unless the external IES is retained in the MAC), or of any IES located in part of the MIC genome that is not collinear with MAC chromosomes. The complexity of the assembled PGM DNA is almost 100 Mb, although we could not properly assemble repeated sequences. We thus estimate that at least 25% of the germline is not collinear with the MAC chromosomes, and might contain unique copy IESs or transposons, the excision of which could only have been detected if the flanking region were retained in the MAC.

Second, this reference set does not provide information about the variability in IES excision patterns that might exist between different, though genetically identical, cell populations. Many IESs are under maternal, epigenetic control [40,31,41]. The genome scanning model [42] posits that every time *Paramecium* undergoes meiosis, the scnRNA pathway compares the maternal MIC, in the form of 25 nt scnRNAs [43], with the maternal MAC, in the form of long non-coding transcripts [44]. The scnRNAs that cannot be subtracted by base pairing with the long maternal transcripts are licensed for transport into the new developing MAC [45] where they target homologous sequences for elimination, probably via deposition of epigenetic marks on the chromatin (cf [3,4] for recent reviews of genome scanning in *Paramecium* and *Tetrahymena*). The scnRNA pathway in theory provides a powerful defense mechanism against transposons that invade the germline and can explain the molecular basis of alternative MAC rearrangement patterns that are maintained across sexual generations [31,40,41,46,47]. Hence the following caveat: any genome-wide set of IESs is identified with respect to a particular MAC reference genome sequence. There can be no "universal" IES reference set for the species. Since IESs can be a source of genetic variation as discussed in [48], the IES catalogue we have established will make it possible to study this variation, for example by surveying IES retention in the MACs of geographic isolates and in stocks that have been experimentally subjected to different types of stress.

Constrained IES size distribution and the IES excision complex

The remarkable sinusoidal distribution of IES sizes retained by evolution reflects strong constraint on the distance between IES ends. We assume that the selection is exerted through the excision mechanism, since the retention of an IES in the MAC can impair gene function. An IES that cannot be efficiently excised is expected to be counter-selected. We propose an interpretation of the IES size distribution based on its similarity with data generated

by "helical-twist" experiments, which have provided evidence of DNA looping between distant protein-binding sites in various, mainly prokaryotic, DNA transaction systems (transposition, gene control, replication initiation, site-specific recombination, etc. reviewed in [49]). In these experiments, the distance between transposon ends [50,51], repressor binding sites [52–55] or site-specific recombination sites [56] is varied, on plasmids or on the bacterial chromosome, and the activity of the system is measured *in vivo*. The observed periodicity in the length-dependence of the activity corresponds to the helical repeat of the DNA, since the same face of the double helix must interact with the protein at each end, and given the prohibitive energetic cost of twisting the double helix to fit the binding site to the protein. This is especially true for DNA fragments whose size is close to the persistence length of double stranded DNA (~150 bp) or shorter. The persistence length, a physical measure of the bending stiffness of a polymer in solution, is the length above which there is no longer a correlation between the orientation of the ends of the molecule. For DNA longer than its persistence length, it becomes possible for the 2 ends to encounter each other to form a loop, without any external intervention.

Almost all (93%) of the IESs in the genome are shorter than the persistence length of DNA. The size distribution, which appears as a series of regularly spaced peaks, can be decomposed into three parts. The largest peak is centered on 28 bp but displays an abrupt minimum size cutoff at 26 bp. A second peak seems to be of forbidden size. Finally, there follow a series of peaks that are best fit by a sine wave with a ~10.2 bp periodicity. In the helical-twist experiments, the amplitude of the measured biological activity peaks tends to decrease with decreasing distance between interacting sites. However, for the IES size distribution, the decay of IESs over time imposes the opposite tendency: the peak heights increase as IES size decreases.

Our working model for assembly of an active IES excision complex is shown in Figure 7. We propose that, starting at the third peak (44–46 bp), the IESs assemble into the excision complex by forming a double-stranded DNA loop compatible with presentation of the same face of the double helix to the Pgm endonuclease at both IES ends. The near absence of the second peak, the minimum IES size of 26 nt and the 13 bp size of each *piggyBac* TIR [25] lead us to suggest that the IESs in the first peak are able to assemble an active excision complex without formation of a DNA loop. The IESs in the nearly absent second peak would not be efficiently excised, as they would be too short to form a DNA loop and too long to form an active excision complex without a DNA loop.

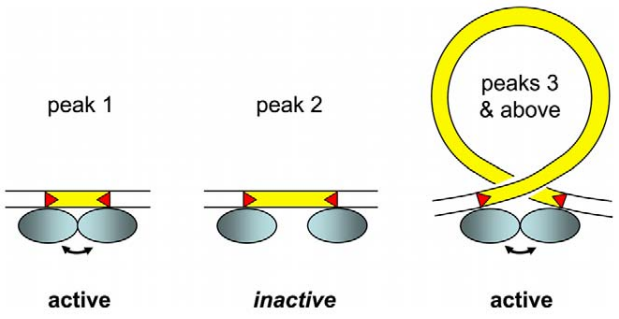


Figure 7. IES size constraint and the assembly of an active excision complex. Our working model is based on the assumption that oligomerization of the IES excisase (most likely the domesticated transposase PiggyMac) on DNA activates catalytic cleavage at IES ends (IESs are drawn in yellow and red triangles highlight the orientation of their ends). In the absence of any information on the stoichiometry of the complex, the excisase is represented by a shaded blue ellipse. For very short IESs from peak 1 (26–30 bp in length), the required contact between protein subunits may be established directly (double-headed arrow) and the complex is active. For IESs longer than 44 bp (peak 3 and above), we propose that looping of the intervening DNA double helix brings IES ends into close proximity and activates DNA cleavage. We have arbitrarily drawn the complex as an antiparallel arrangement of IES ends within a negatively supercoiled loop, but other conformations are possible. IESs from the "forbidden" peak 2 would be too long to allow direct contacts between protein subunits to be established, and too short to form an excision loop.
doi:10.1371/journal.pgen.1002984.g007

Molecular analysis of the IES excision mechanism supports the involvement of such a transpososome-type excision complex. First, the domesticated Pgm transposase, which has retained the catalytic site of *piggyBac* transposases [21], is very likely to be the endonuclease responsible for the cleavage reaction, involving the introduction of DSBs at each end of the IES [24]. Second, for IESs larger than 200 bp, covalently closed circular molecules containing the excised IES have been detected as transient intermediates during MAC development [57]. Third, if one end of an IES bears a mendelian mutation in the TA dinucleotide, no DSB occurs at either end of the IES. This indicates that the two IES ends must interact, directly or indirectly, before cleavage can occur [58].

It is worth noting that "canonical" TIRs of cut-and-paste transposons are often bipartite. They are composed of an internal sequence motif recognized and bound by the transposase, and of a few nucleotides at the termini that constitute the DNA cleavage site [59]. The obligatory conservation of a TA dinucleotide at IES ends is indicative of a requirement for DNA cleavage but is not sufficient for specific recognition, even if we take into account the weak consensus over the 6 internal nucleotides. The lack of a sufficiently long conserved motif in IESs makes it unlikely that Pgm recognizes IESs by binding to a specific sequence. For IESs under maternal control [31], it is currently thought that Pgm is recruited to its substrate via epigenetic marks deposited on the chromatin by the scnRNA pathway [3,21,42].

The picture of an IES excision complex that emerges from these considerations, which must of course be tested biochemically, requires very short pieces of DNA to form loops (Figure 7). Proteins that bend DNA, such as HMG proteins [60], could be involved. What is quite remarkable here, beyond the fact that evolution has performed such a nice "helical-twist" experiment, is that the DNA loops might be as short as ~45 bp, shorter than almost any reported case of DNA looping. The minimal *in vivo* value reported for cut-and-paste bacterial transposons is 64–70 bp [50,51] and this is also the minimum size reported for HMG assisted DNA loop formation *in vitro* [60]. The only indication of shorter loops comes from detection of a minor peak of activity *in vivo* and *in vitro* for ~50 bp DNA loops in the *E. coli* Hin invertasome, provided that invertasome assembly occurs in the presence of HU, a bacterial nucleoid protein that bends DNA [56]. Given the unusually high A+T content of IESs (80%), local melting might favor the deformations in the double helix required to make the very small looped structures of the postulated IES excision complex.

Evidence that IESs are remnants of transposons

Ciliate MICs have long been recognized as safe havens for transposons, since removal of the transposons from the somatic DNA during development would decrease the burden on host fitness, as discussed in [19]. Our study provides the first global vision of IESs in any ciliate germline and provides strong support for the "transposon link" hypothesis that present day IESs are remnants of transposons [18,19].

Although we do not yet have a complete picture of the transposon landscape of the *P. tetraurelia* germline genome, we have identified 3 families of Tc1/mariner elements, with 2 quite different structures. The *Thon* and *Sardine* transposons have long, palindromic TIRs, a tyrosine recombinase and a DDE transposase characteristic of the IS*630*/Tc1 subfamily, with a short spacer (32 aa) between the 2nd and 3rd catalytic residues. This clearly distinguishes these transposons from the *piggyBac* family characterized by a long spacer and a DDD catalytic triad. The IESs related to these elements that we were able to identify appear as solo TIRs. Given the presence of repeated, palindromic subsequences in each TIR, we can speculate that the solo TIRs result from recombination between short direct repeats present within the complex TIRs, as proposed to explain the incidence of solo LTRs derived from LTR retrotransposons in the genomes of some organisms [61,62]. The other transposon family we have identified, *Anchois*, is characterized by much shorter TIRs which do not contain internal palindromes, a similar DDE transposase and the absence of a tyrosine recombinase. This structure is similar to that of the *P. primaurelia Tennessee* transposon [34]. In the case of *Anchois*, we could find a number of IESs that appear to correspond to the entire transposon or large portions of it, including IESs with a recognizable but degenerate DDE transposase ORF.

It is possible that we have only scratched the tip of the iceberg since the germline genome is expected to contain other mobile elements. Indeed, we were able to identify 8 clusters of homologous IESs inserted at non-homologous genomic sites, suggesting recent mobility, and one of these clusters turned out to consist of IESs that are solo TIRs of the *Thon* element. The other clusters could be the remains of as yet unidentified elements. Both the *Thon* and the *Anchois* IES homologies were detected among the largest IESs in the genome-wide set (i.e. the 380 IESs >500 bp), and for none of them could we detect ohnologous IESs from the recent WGD, an indication that these IESs were recently acquired. Since over 90% of present day IESs have decayed to very short sizes (<150 bp) it is not surprising that internal transposon motifs can no longer be recognized. These very short IESs nonetheless display the short degenerate Tc1/mariner end consensus. The existence of this consensus at IES ends may testify to their evolutionary transposon origin. This end consensus would eventually have become a requirement for efficient cleavage by the IES excision machinery. We can imagine two instances of such convergent evolution: i) other families of mobile elements could be eliminated by the PiggyMac-dependent mechanism and ii) genomic sequences that adhere to the end consensus could be excised just like IESs. We conclude that at least a fraction of IESs are decayed Tc1/mariner transposons, and we consider highly probable that some IESs are derived from other mobile elements.

IESs are a burden for host fitness

Since IES excision is not 100% efficient, IES insertions are in general deleterious, consistent with the different kinds of selective pressure we have observed: (i) a constrained IES size distribution likely reflecting assembly of the excision complex; (ii) a bias against IESs that do not lead to premature stop codons in case of IES retention in the MAC; (iii) an inverse correlation between IES insertions and gene expression level. IESs can in addition be considered to constitute a mutational burden, in the same way as introns are considered to constitute a mutational burden in intron-rich eukaryotic genomes [63], since IESs are present in large number in *Paramecium*, and any mutation in a flanking TA dinucleotide abolishes IES excision. Nonetheless, the system can give rise to beneficial new functions, as attested by use of the IES excision machinery to provide a regulatory switch for mating type determination (D. Singh, personal communication).

Since IESs are in general deleterious and constitute a fitness burden for the organism, and since we have detected cases of probable clean IES loss from the germline DNA suggesting that a mechanism exists for precise IES excision in the MIC, we may ask why *Paramecium* has any IESs at all. This question can be easily answered if we consider that IESs arise from selfish genetic elements (SGEs, defined as elements – typically transposable elements or viruses – that can enhance their own transmission relative to the rest of the genome, with deleterious or neutral effects for the host [64]). The number of IESs reflects the balance between the number of IES insertions (e.g. invasion by SGEs that subsequently decayed to become unique-copy IESs) and the strength of selection against these insertions, which either prevents fixation of new insertions in the population or favors loss of already fixed insertions. This genetic conflict is mediated by an ''arms race'' between SGEs and the host as discussed by Werren [64].

Host defense mechanisms in ciliates

In all kingdoms of life, non-coding RNAs are used to defend host genomes against parasitic nucleic acids, as exemplified in eukaryotes by small RNA pathways involved in protection against viruses or in silencing transposons to ensure integrity of the germline genome [65–67]. In ciliates, nuclear dimorphism provides the potential for an additional layer of protection by physically separating the chromosomes that store the genetic information from the rearranged chromosomes that express the genetic information. Additional host defense machinery providing precise excision of transposons/IESs from somatic DNA, might have allowed the invasion of a fraction of the genome in which SGEs are not usually tolerated, namely the coding and regulatory sequences required for gene expression.

In the case of *Paramecium*, Pgm domestication has provided the mechanism for precise excision of TA-bounded insertions from the somatic DNA, allowing transposons/IESs to be cleanly excised from genes in the MAC. Since this would reduce the fitness burden caused by transposition, we presume that it allowed transposons to spread throughout the MIC genome. Recognition of the IESs is however ensured by the scnRNA pathway [3], itself an example of the more ancient mechanism of small RNA-based host immunity against foreign nucleic acids, and this epigenetic recognition may in part explain the less than 100% efficiency of IES excision.

In *Tetrahymena*, which has both a scnRNA pathway and domesticated *piggyBac*-like transposases [4,22], only excision of intergenic IESs has been studied for the moment and use of heterogeneous cleavage sites was found. This imprecise excision would not be compatible with insertion in genes since gene expression would be compromised. *Tetrahymena* has only about 6,000 IESs and indeed, they are not usually found within genes [17]. Why doesn't *Tetrahymena* have intragenic IESs? We can only speculate that a Tc1/mariner invasion after the divergence of *Paramecium* and *Tetrahymena* was instrumental in the evolution of a precise excision mechanism in *Paramecium*, necessary for spread of these elements throughout the genome. In support of this hypothesis, a recent genome-scale identification of hundreds of *Tetrahymena* IESs [17] revealed a new class of TTAA-bound IESs that are precisely excised. They were found to contribute 3′ exons to genes that are expressed from the zygotic genome during genome rearrangements. These elements might be derived from *piggyBac* transposons, which have TTAA target sites, and perhaps

testify to the ancient *piggyBac* invasion that led to domestication of the transposase.

A contrasting situation is found in some stichotrich ciliates. The stichotrich ciliates are very distantly related to the oligohymenophorean ciliates and are characterized by highly fragmented somatic genomes consisting of nanochromosomes that usually bear a single gene. Intragenic IESs are more abundant in the germline genomes of *Oxytricha* and related strichotrichs than in *Paramecium*, with an estimate of at least 150,000 IESs per haploid genome [68]. Both single-copy IESs and transposons are precisely excised and the precise IES excision is assured by guide RNAs transcribed from the maternal MAC [69], which are even capable of reordering the scrambled MAC-destined gene segments that occur frequently in *Oxytricha* and related stichotrichs [70]. There is also evidence that the endonuclease required for cleavage in *Oxytricha* is actually a transposase from germline TBE transposons [71]. However, there is currently no evidence for a scnRNA pathway specialized in the control of DNA elimination, although gene silencing by RNAi in *Oxytricha* testifies to the presence of small RNA machinery [69]. Thus the high precision and fidelity of the guide RNA mechanism for genome rearrangements in *Oxytricha* spp. seems to have tipped the balance even further in favor of intragenic IES insertions.

The case of *Euplotes*, a stichotrich ciliate distantly related to *Oxytricha* and probably lacking scrambled genes, merits special attention. Beautiful work carried out by the Jahn and Klobutcher labs in the 1990s showed (i) the existence of high copy number Tc1/mariner elements, Tec1 (2,000 copies per haploid genome) and Tec2 (5,000 copies), as well as lower copy number Tec3 elements (20–30 copies) [33,72,73]; (ii) at least a fraction of these Tec elements are precisely excised between TA dinucleotides [74]; (iii) an estimated 20,000 short TA-bounded IESs [33], bearing a Tc1/mariner end consensus just like the *Paramecium* IESs [18], are excised precisely between TA dinucleotides leaving a single TA at each excision site on the MAC destined chromosomes [33] and (iv) molecular characterization of excised circular forms of both Tec elements and short IESs revealed an unusual junction consisting of 2 TA dinucleotides separated by 10 bp of partially heteroduplex DNA, showing that both the Tec transposons and the short IESs are excised by the same mechanism [74,75]. The mechanism is moreover different from that of precise IES excision in *Paramecium* [24,57]. Neither the endonuclease responsable for IES cleavage nor the repair pathway has currently been identified in *Euplotes*. It will be fascinating to see whether the same actors, i.e. a domesticated *piggyBac* transposase and the NHEJ (non-homologous end-joining) pathway, are responsible for a mechanism that in its details is not the same as that found in *Paramecium*, or whether completely different cellular machinery has been recruited to carry out the same function i.e. the precise excision from somatic DNA of the Tc1/mariner family Tec transposons and of short TA-bounded IESs presumed to be their relics [19].

In conclusion, different ciliates have evolved different host defenses in response to germline SGE insertions. In all cases that have been examined at the molecular level, maternal non-coding RNAs are involved in programming genome rearrangements. In *Paramecium* and some other lineages, the co-evolution of host defense machinery and SGEs has provided mechanisms for precise somatic excision, uniquely allowing the colonization of coding sequences by Tc1/mariner and likely other transposable elements. This phenomenon is so far only paralleled by the spread of introns into eukaryotic coding sequences, also thought to result from domestication of precise excision machinery, derived in this case from mobile self-splicing ribozymes [76].

Materials and Methods

Purification of DNA enriched in un-rearranged sequences from isolated nuclei of cells depleted for PiggyMac

Cell growth and autogamy. *Paramecium tetraurelia* strain 51 was used for this study because the available phage-lambda library of purified MIC DNA was made using this strain. Strain 51 only differs at a few loci from strain d4-2 that was used for sequencing the MAC genome [77].

For gene silencing, we used the «feeding» method described in [78]. *Escherichia coli* HT115 [79] harboring plasmid L4440 [80], with the 567-bp *Hind*III-*Nco*I fragment of gene *PGM* inserted between two convergent T7 promoters [21], was induced at 37°C for the production of *PGM* dsRNA in WGP1X medium containing 100 µg/mL ampicillin. As a control, we induced HT115 bacteria for the production of dsRNA homologous to the *ND7* non essential gene (see plasmid description in [81]).

Paramecium tetraurelia strain 51new mt8 [24] was grown at 27°C in WGP1X inoculated with *Klebsiella pneumoniae* and supplemented with 0.8 µg/mL β-sistosterol. Following ~25 divisions, cells were washed and transferred to 4.1 L of freshly induced *E. coli* HT115. Cells were allowed to grow for 8 vegetative divisions, then starved to trigger autogamy. The progression of autogamy was monitored by DAPI staining (Figure S2A) and the viability of sexual progeny was tested to evaluate the efficiency of *PGM*-silencing (Figure S2B).

Cell lysis and purification of developing MAC DNA. Following prolonged starvation to favor the degradation of old MAC fragments (day 4 of autogamy), all cultures were filtered through eight layers of sterile gauze. Cells were collected by low-speed centrifugation (285× g for 1 min) and washed twice in 10 mM Tris-HCl pH 7.4. Particular care was taken to eliminate contaminating bacterial biofilms by letting them settle to the bottom of the tubes and removing them with a Pasteur pipette prior to all washing centrifugation steps. The final pellet was diluted 5-fold by addition of lysis buffer (0.25 M sucrose, 10 mM $MgCl_2$, 10 mM Tris pH 6.8, 0.2% Nonidet P-40) and processed as described in [82]. All steps were performed at 4°C. Briefly, cells (1 mL) were lysed with 100 strokes of a Potter-Elvehjem homogenizer and washing buffer (0.25 M sucrose, 10 mM $MgCl_2$, 10 mM Tris pH 7.4) was added to a final volume of 10 mL. Developing new MACs (together with cell debris, bacterial biofilms and the largest fragments of the old MAC) were collected by centrifugation at 600× g for 1 min and washed 3 times in washing buffer. To remove contaminating bacteria, the pellet was diluted in washing buffer, loaded on top of a 3-mL sucrose layer (2.1 M sucrose, 10 mM $MgCl_2$, 10 mM Tris pH 7.4) and centrifuged in a swinging rotor for 1 hr at 210,000× g. The nuclear pellet was collected and diluted 5-fold in 10 mM $MgCl_2$ 10 mM Tris pH 7.4 prior to addition of two volumes of proteinase K buffer (0.5 M EDTA pH 9, 1% N-lauryl sarcosine sodium, 1% SDS, 1 mg/mL proteinase K). Following 16-hr incubation at 55°C, genomic DNA was purified as described in [24], with three additional phenol:$CHCl_3$ extractions (1:1), one $CHCl_3$ extraction and a final ethanol precipitation [83]. Enrichment for non-excised IESs (IES^+ forms) was assayed by 1% agarose gel electrophoresis of *Pst*I-restricted DNA and Southern blot hybridization with ^{32}P-labeled Gmac probe [21], which corresponds to the MAC sequences just downstream of IES 51G4404 within the surface antigen G^{51} gene (Figure S3A). To measure the contamination with bacterial DNA, the same blot was dehybridized and probed with a ^{32}P-labelled fragment of *K. pneumoniae* 23S rDNA amplified by PCR using primers KP23S-U (5′-AGCGTTCTG-TAAGCCTGCGAAGGTG-3′) and KP23S-R (5′-TTCACCTA-CACACCAGCGTGCCTTC-3′) (Figure S3B). All radioactive

signals were scanned and quantified using a Typhoon phosphorimager (Figure S3C).

Purification of wild-type micronuclear DNA from a lambda-phage library

A lambda-phage library was provided by John Preer. This library had been made from DNA obtained after isolation of stock 51 wild type micronuclei [82] and further purified by cesium chloride density gradient centrifugation to eliminate G+C-rich DNA supposed to represent bacterial contaminants [28]. The library consisted of 70,000 recombinant phages (lambdaGEM11), expected to represent a 7× coverage of the MIC genome. We amplified the original library in 1995 and stored it at 4°C. Phage particles from 1 mL of the reamplified library (approximately 10^5 particles) were fully recovered by ultracentrifugation (42 min at 113898 g in a TLA-55 rotor; $S_{\text{lambda particle}} = 410$ according to [84]) and concentrated in ~30 µL. Given the limited amount of material (~18 pg of 40 Kb phage genomes corresponding to ~4.5 pg of inserts), the cloned DNA was amplified by PCR using primers located next to the cloning sites (LambdaL2 GGCCTAATACGACTCACTATAGG; LambdaR2 GCCATTTAGGTGACACTATAGAAGAG). Non-genomic sequences should only represent 0.6% of the total PCR-amplified DNA. As PCR inhibitors prevented direct amplification from the concentrated suspension of phage particles, 230 50 µL-PCR reactions were performed from 3 µL of a 30× dilution in SM. The Expand Long-Template PCR System (Roche) was used as recommended by the supplier with 23 amplification cycles, an annealing temperature of 60°C and 12 min for the extension time. PCR reactions were concentrated by ethanol precipitation and ~35 µg of 9 to 13 Kb PCR products were obtained after purification from 0.6% low-melting-temperature agarose gels and treatment with β-agarase (Sigma).

DNA sequencing

DNA was sequenced by a paired-end strategy using Illumina GAII and HiSeq next-generation sequencers. The shotgun fragments were ~500 bp and the paired-end reads 108 nt for DNA enriched in un-rearranged sequences (PGM DNA). The fragments were ~200 bp and the paired-end reads 101 nt for DNA prepared from the lambda-phage library. In the latter case, short reads that overlapped were merged.

Short read mapping

All Illumina short reads were mapped to the strain 51 reference genome (see below) using BWA [85] (version 0.5.8). Alignments were indexed using samtools [86] (version 0.1.11).

Strain 51 reference genome

The *P. tetraurelia* MAC genome [1] was assembled from 13× Sanger sequencing reads from different insert size librairies of strain d4-2 DNA. Strain d4-2 only differs from strain 51 at a few loci. We corrected sequencing errors in the scaffolds using Illumina deep sequencing in two stages, the first stage using the same strain d4-2 DNA sample that had been used for the original Sanger sequencing (84 million 75 nt paired-end reads), the second stage using two different samples of strain 51 MAC DNA (155 million 75 nt paired-end reads). The electronic polishing pipeline used for each stage consisted of the following steps. (i) Gap filling was achieved by assembling the Illumina reads into contigs using the Velvet [87] short read assembler (Kmer = 55 -ins_length 400 -cov_cutoff 3 -scaffolding no). The contigs were mapped to the draft assembly using BLAT [88] and locally realigned with Muscle [89]. If the contigs spanned a sequencing gap, then it was filled. (ii) The Illumina reads were mapped to the draft genome using BWA [85]. (iii) Alignments were indexed using samtools [86]. (iv) Samtools mpileup program and homemade Perl scripts were used to identify all positions covered by at least 10 reads and where at least 80% of the reads did not confirm the reference sequence. (v) The reference sequence was corrected using the list of errors. Steps (ii) through (v) were repeated a few times at the second stage of correction using the strain 51 reads, since BWA mapping has low error tolerance, and more reads could be mapped as the correction progressed. At the end of the process 442 of 861 sequencing gaps were filled, 13,758 substitutions were corrected, 929 deletions of 1–2 nt were filled and 10,339 insertions of 1–2 nt were removed, to yield the strain 51 reference genome that was used for IES identification. The strain 51 reference genome is available via ParameciumDB [90].

IES identification pipeline

MIRAA pipeline. All reads of the DNA enriched in un-rearranged sequences (PGM DNA), were mapped on the *P. tetraurelia* strain 51 reference MAC genome. Alignments indexed with samtools were analysed using custom perl scripts written with the BioPerl library (version 1.6) and the Bio::DB::Sam module (version 1.11). An IES site is characterized by an excess of ends of read alignments since reads that overlap IES junctions only map partially on the MAC genome and stop on the residual TA. These positions are considered to be IES sites if (i) the number of alignment ends is greater than 15 (10% of the average PGM DNA read coverage); (ii) if they are more than 500 bp from the ends of a scaffold, which avoids errors produced by heterogeneity in these regions; (iii) if the read coverage is lower than 300×, to avoid highly repeated sequences.

MICA pipeline. The IES detection pipeline consists of the following steps: (i) paired-end read assembly with Velvet [87] (version 1.0.18) using 3 different Kmer values (41, 45 and 55) and the parameters "-scaffolding no -max_coverage 500 -exp_cov auto -ins_length 500 -min_contig_length 100"; (ii) Only contigs with average G+C content less than 0.5 are retained; (iii) repeats are masked with RepeatMasker; (iv) masked contigs are aligned on the reference MAC genome with BLAT (version 34); (v) gaps are realigned locally with Muscle (version 3.7) and a custom Perl script is used to adjust the ends of the alignment. If the alignment is bound by TA dinucleotides, the insert in the contig is considered to be an IES. This pipeline was used to find IESs in the following sets of reads:

1. All the PGM reads after removal of known contaminants (bacteria, rDNA, mitochondrial DNA).
2. All pairs of reads in which at least one read does not align with the MAC reference genome, in order to enrich in MIC reads.
3. All PGM reads after removal of reads that correspond to the potential MAC IES junctions identified by the MIRAA pipeline.
4. Finally, all PGM reads after removal of reads that correspond to a MAC junction identified by the MICA pipeline using the above data sets.

The IES identification pipelines, datasets and overlap between IESs and potential IES sites are summarized in Figure S1. The statistics for each of the assemblies are provided in Table S1.

IES conservation

Determination of IESs that are conserved in genes duplicated by a WGD event involved identification of the position of the IES

with respect to the beginning of the alignment, either using a protein alignment of ohnologs, back translated into nucleotide sequence, or using nucleotide alignment of the 2 genes. In both cases, the alignments were carried out using Muscle [89] (version 3.7). If the relative positions of the IES is the same within a 2 nt tolerance, then the IESs are considered to be conserved.

Measurement of protein divergence

A phylogenetic tree was computed by concatenation of the alignment of 1350 protein families corresponding to quartets of ohnologs preserved after both the intermediate and recent WGD events. All gap-containing sites were excluded from the alignment, which is therefore robust with respect to possible annotation errors. The tree was constructed using BioNJ [91] with Poisson correction for multiple substitutions. The average length of the 2 branches between the intermediate and recent WGDs is 0.085 substitutions/site. The average length of the 4 branches between the recent WGD and the present is 0.0825 substitutions per site. Assuming a constant substitution rate, we can infer that the time between the intermediate and recent WGD events and between the recent WGD and the present are equivalent, although we cannot date the events since we do not know the substitution rate in *Paramecium*.

Availability of data

The MAC reference genome used for this study (strain 51) and the genome-wide set of IESs are available at http://paramecium.cgm.cnrs-gif.fr/download/. The IESs have also been integrated into ParameciumDB BioMart complex query interface and the ParameciumDB Genome Browser [90]. The short read datasets have been deposited at the European Nucleotide Archive (Accession numbers ERA137444 and ERA137420).

Validation by PCR of individual IES or IES insertion sites

Oligonucleotides were designed to flank the IES insertion site at a distance of 150–200 nt to allow detection of amplification products with or without an IES. All PCR amplifications were performed with an Eppendorf personal mastercycler. Standard PCR amplifications were performed with 1 unit of DyNazyme II with reagent concentrations according to instructions provided by Finnzyme (dNTP: 200 µM each, primers: 0.5 µM each) with 50 ng of template DNA. The program used is 2 min at 95°C, 10 cycles of 45 sec at 95°C, 45 sec at annealing temperature, and 1 min at 72°C, 15 cycles of 20 sec at 95°C, 20 sec at annealing temperature and 1 min at 72°C, followed by a final incubation for 3 min at 72°C. Amplified products were analyzed on 3% Nusieve (Lonza) in TBE 1×. Long and AT-rich PCR amplifications were performed with 1 unit of Phusion (Finnzymes) using the following concentrations of reagents (dATP and dTTP: 400 µM each, dCTP and dGTP: 200 µM each, primers: 0.5 µM each) with 50 ng of template DNA. The program used was 1 min at 98°C, 25 cycles of 10 sec at 98°C, 30 sec at annealing temperature and 5 min at 72°C, followed by a final incubation of 2 min at 72°C. Amplified products were analyzed on 1% UltraPure agarose (Invitrogen) in TAE 1×. The template DNA for the amplification reactions was an aliquot of PGM DNA enriched in un-rearranged sequences, prepared as described above.

Transposon identification

Isolation of inserts from the MIC lambda-phage library [28] was carried out as previously described [31]. Phage inserts and long-range PCR products obtained by amplification of total DNA from vegetative cells were isolated and subjected to Sanger sequencing as in [34]. The lambda-phage inserts and the cloned long-range PCR products used to characterize the *Sardine* and *Thon* transposons have been deposited in the EMBL Nucleotide Sequence Database with accession numbers HE774468–HE774475.

Several IESs with homology to the PFAM DDE_3 domain were used to find other IESs sharing nucleotide identity, leading to a set of 28 IESs that were aligned with Muscle [89] to identify 2 *Anchois* transposons. In a second step, the alignment was refined and manually adjusted in order to reconstruct the *AnchoisA* and *AnchoisB* transposons. These second step alignments were built using IESs along with some PGM contigs that correspond to germline-restricted, imprecisely eliminated regions of the genome containing *Anchois* copies (Text S1).

Data analysis

Statistical analyses and graphics were performed in the R environment for statistical computing [92] using standard packages, as well as the ape package [93] for phylogenetic analysis. Sequence logos were generated using weblogo software [94].

Supporting Information

Figure S1 IES identification. A. Schematic representation of the MIRAA pipeline for identification of IES sites by read mapping. B. Schematic representation of the MICA pipeline for identification of IESs by comparison of contigs with the reference genome assembly. C. PGM DNA datasets which were used with the MICA pipeline to identify the genome-wide set of IESs. As explained in Materials and Methods, the 4 datasets are (i) all PGM reads after filtering known contaminants, (ii) all filtered reads with at least one member of the pair that does not match the MAC reference genome, (iii) all filtered reads after removal of the read pairs with a perfect match to a MAC IES juction identified with the MIRAA pipeline and (iv) all filtered reads after removal of the read pairs with a perfect match to a MAC IES junction identified with MICA and the first 3 datasets. D.Venn diagram showing that 96% (n = 43,220) of the IESs identified with MICA correspond to IES insertion sites identified by MIRAA. The MICA pipeline was also used to identify IESs in the phage-lambda inserts: the sequence reads were assembled into 3 sets of contigs with Velvet, using 3 different kmer values (kmer = 45, 51 or 55).
(PDF)

Figure S2 Autogamy time-course of *P. tetraurelia* 51 mt8 submitted to RNAi against *PiggyMac*. A. Cells were transferred at day 0 into 4.1 L of freshly induced feeding bacteria producing dsRNA homologous to a 567-bp region of the *PGM* gene and incubated at 27°C. The progression of autogamy was monitored everyday (D1: day 1, D2: day 2, D3: day 3, D4: day 4) by DAPI staining of cells. V: vegetative cells, F: cells with fragmented old MAC and no clearly visible new developing MACs, A: cells harboring two developing new MACs, C: post-autogamous cells with one new MAC surrounded with fragments of the old MAC. B. Survival of post-autogamous progeny. At day 4, 30 autogamous cells were transferred individually to standard growth medium containing *K. pneumoniae* and incubated at 27°C to follow the resumption of vegetative growth. Survival of the progeny of autogamous cells obtained in standard (Kp) or in control RNAi medium (ND7) was also tested. Wt: normally-growing progeny, sick: slowly-growing cells, often with abnormal swimming behavior.
(PDF)

Figure S3 Purification of IES-enriched genomic DNA from PGM-silenced cells. Autogamous cells were collected at day 4 and

genomic DNA was extracted through several cell fractionation steps. Lys.1 and lys.2: independent samples of cells were lysed directly in proteinase K buffer; low sp.: DNA extracted from low speed pellets (600× g for 1 min followed by washing); suc.: DNA extracted from nuclear pellets obtained following centrifugation through a 2.1 M sucrose layer. Each DNA sample was digested by PstI and the digestion fragments were separated on a 1% agarose gel. A. Southern blot hybridization with the Gmac probe (shown as a grey box on the diagram). The position of size markers is shown on the left. IES^- and IES^+ bands were quantified separately. B. Southern blot hybridization with the *K. pneumoniae* 23S rDNA probe. Size markers are shown on the right. All rDNA bands were quantified together. C. Quantification of radioactive signals from the blots shown in A and B. The fraction of IES^+ form was normalized relative to the sum of IES^- and IES^+ signals (black histograms). Bacterial rDNA was normalized relative to the sum of IES^- and IES^+ signals (grey histograms).
(PDF)

Figure S4 IES distribution on the 8 largest MAC chromosomes. The 8 largest, telomere-capped scaffolds (~750 Kb to ~980 Kb in size) were normalized to length 1.0 and some were flipped so that the highest IES density is to the right. The curves represent histograms of IES position on each scaffold after Gaussian smoothing using the R "density" function [92]. IES distribution was evaluated using a Kolmogorov-Smirnov test of the null hypothesis that IESs are uniformly distributed on the scaffold. For the 8 largest scaffolds, the null hypothesis was strongly rejected ($p<10^{-8}$). The same statistical test was carried out for gene distribution on these chromosomes, and the null hypothesis was not rejected, consistent with a uniform distribution of genes on the chromosomes.
(PDF)

Figure S5 *Sardine* and *Thon* Tc1/mariner family transposons. From top to bottom: 1) *Sardine* transposon consensus sequence obtained by alignment of the lambda-phage and PCR copies (the latter were amplified from total DNA of vegetative cells using primers located within the *Sardine* TIRs), showing the presence of palindromic TIRs and 4 putative ORFs, including a DDE transposase and a tyrosine recombinase; 2) lambda-phage with the 51G flank that led to discovery of the *Sardine* element (the region of *de novo* telomere addition at the end of the MAC chromosome, following developmental breakage of the MIC chromosome, is indicated); 3) lambda-phage with the S5 copy of *Sardine*; 4) lambda-phage with the S6 copy of *Sardine*, containing an insertion of a different Tc1/mariner transposon, *Thon*, which has the same general organization as the *Sardine* element; 5) lambda-phage with the S7 copy of *Sardine*; 6) lambda-phage with the S8 copy; 7) PCR products (S46 and S103 copies) with nearly intact ORFs; 8) PCR products (S14 and S106 copies) with nearly intact ORFs. The sequences of the 5 lambda-phages and 4 PCR products have been deposited in the EMBL/GenBANK/DDBJ public nucleotide database with EMBL-Bank accession numbers HE774468–HE774475.
(PDF)

Figure S6 IES size distribution. The histograms represent A) IESs inserted in coding sequences. B) IESs inserted in non-coding sequences. IESs larger than 150 nt are not displayed. The fact that very similar periodic distributions are observed for IESs in both coding and intergenic regions is consistent with the hypothesis that the periodic size constraint is related to the IES excision mechanism. Indeed, IES retention in the MAC could be deleterious either by affecting ORFs (IESs in protein coding sequences) or by affecting regulatory signals (IESs in non-coding sequences).
(PDF)

Figure S7 IES evolution evaluated with quartet IES groups. A) schematic representation of the observable quartet IES groups, arranged from top to bottom according to the number of IESs that are conserved and from left to right, according to the most recent period in which the ancestral IES could have been acquired. B) Schematic representation of the parameters of a statistical model developed to test hypotheses about IES evolution (cf. Text S1). The three time periods delimited by the 2 WGD events and the present time are designated, from the oldest to the most recent, g_3, g_2 and g_1. The parameters ρ_3, ρ_2 and ρ_1 are the fraction of IESs that were acquired in each of these time period and the parameters of the form $\delta_{a,b}$ are the survival rates for an IES acquired in period g_a during the period g_b. The equations of the model express the observable IES counts as a function of these parameters.
(PDF)

Table S1 Assembly statistics.
(PDF)

Table S2 Molecular validation of some predicted IESs and IES insertion sites.
(PDF)

Table S3 Validation of the genome-wide set of IESs using previously characterized IESs.
(PDF)

Table S4 IESs with homology to *Anchois* transposons.
(PDF)

Table S5 Deficit of 3n IESs in coding sequences, for each peak of the 10 bp periodic size distribution.
(PDF)

Text S1 Transposon sequences. A). The sequences of *Sardine*, *Thon* and *Anchois* transposons reconstituted from manually adjusted multiple alignments of the different decayed copies, cloned from the lambda phage library of MIC DNA (*Sardine*, *Thon*) or found in the PGM DNA assembly (*Anchois*). The sequences of the *Thon* transposon are those of the only known copy, so that ORF annotation (based on homology with the *Sardine* element) is preliminary; the *Thon* ORF1 sequence apparently contains a frameshift. Predicted introns have been removed from the ORF sequences. B) Annotated comparison of *AnchoisA* and *AnchoisB*, showing the position and orientation of the ORFs, with a potential intron in the DDE transposase ORF. C). Manually adjusted alignment used to reconstitute the AnchoisA copy. See Text S4 for the IESs used in the reconstitution. D). Manually adjusted alignment used to reconstitute the *AnchoisB* copy. See Text S4 for the IESs used in the reconstitution. E) IESs used to obtain the final *AnchoisA* and *AnchoisB* consensus sequences based on the manually adjusted alignments in C) and D).
(PDF)

Text S2 Alignment of homologous IESs inserted at non-homologous genomic sites. The IESs of each cluster of homologous IESs (cf. Table 3 and its legend) and 200 bp of 3′ and 5′ flanking sequences were aligned using Muscle [89]. The IESs are in uppercase type and the flanking sequences are in lowercase type. For cluster5, consisting of IESs homologous to a solo TIR of the *Thon* transposon, the consensus sequence and the *Thon* TIR are included in the alignment and the palindromic repeats are highlighted.
(PDF)

Text S3 PCR approach to validate IESs with homology to *Thon* solo TIRs.
(PDF)

Text S4 A maximum likelihood framework for testing hypotheses about IES evolution.
(PDF)

Acknowledgments

The authors thank Laura Katz, Feng Gao, and Deepankar Pratap Singh for permission to cite their unpublished data and Jean Cohen, Emeline Dubois, and Julien Bischerour for critical reading of the manuscript. The project was carried out in the framework of the CNRS–supported European Research Group "Paramecium Genome Dynamics and Evolution" and the European Science Foundation COST network BM1102 "Ciliates as model systems to study genome evolution, mechanisms of non-Mendelian inheritance, and their roles in environmental adaptation."

Author Contributions

Conceived and designed the experiments: MB LD SD EM SM LS. Performed the experiments: MB CB SD CDW NM SM AM MN OG ALM MP EM. Analyzed the data: OA CDW LD LS. Wrote the paper: OA MB LD SD BEL EM SM LS. Mathematical model: BEL. DNA sequencing: J-MA KL JP PW.

References

1. Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, et al. (2006) Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature 444: 171–178. doi:nature05230.
2. Chalker DL, Yao M-C (2011) DNA elimination in ciliates: transposon domestication and genome surveillance. Annu Rev Genet 45: 227–246. doi:10.1146/annurev-genet-110410-132432.
3. Coyne RS, Lhuillier-Akakpo M, Duharcourt S (2012) RNA-guided DNA rearrangements in ciliates: is the best genome defense a good offense? Biol Cell Accepted manuscript online. doi:10.1111/boc.201100057.
4. Schoeberl UE, Mochizuki K (2011) Keeping the soma free of transposons: programmed DNA elimination in ciliates. J Biol Chem 286: 37045–37052. doi:10.1074/jbc.R111.276964.
5. Bétermier M (2004) Large-scale genome remodelling by the developmentally programmed elimination of germ line sequences in the ciliate Paramecium. Res Microbiol 155: 399–408.
6. Ruiz F, Krzywicka A, Klotz C, Keller A, Cohen J, et al. (2000) The SM19 gene, required for duplication of basal bodies in Paramecium, encodes a novel tubulin, eta-tubulin. Curr Biol 10: 1451–1454.
7. Haynes WJ, Ling KY, Preston RR, Saimi Y, Kung C (2000) The cloning and molecular analysis of pawn-B in Paramecium tetraurelia. Genetics 155: 1105–1117.
8. Mayer KM, Mikami K, Forney JD (1998) A mutation in Paramecium tetraurelia reveals functional and structural features of developmentally excised DNA elements. Genetics 148: 139–149.
9. Mayer KM, Forney JD (1999) A mutation in the flanking 5′-TA-3′ dinucleotide prevents excision of an internal eliminated sequence from the Paramecium tetraurelia genome. Genetics 151: 597–604.
10. Matsuda A, Forney JD (2005) Analysis of Paramecium tetraurelia A-51 surface antigen gene mutants reveals positive-feedback mechanisms for maintenance of expression and temperature-induced activation. Eukaryotic Cell 4: 1613–1619. doi:10.1128/EC.4.10.1613-1619.2005.
11. Yao MC, Choi J, Yokoyama S, Austerberry CF, Yao CH (1984) DNA elimination in Tetrahymena: a developmental process involving extensive breakage and rejoining of DNA at defined sites. Cell 36: 433–440.
12. Saveliev SV, Cox MM (2001) Product analysis illuminates the final steps of IES deletion in Tetrahymena thermophila. EMBO J 20: 3251–3261. doi:10.1093/emboj/20.12.3251.
13. Fillingham JS, Thing TA, Vythilingum N, Keuroghlian A, Bruno D, et al. (2004) A non-long terminal repeat retrotransposon family is restricted to the germ line micronucleus of the ciliated protozoan Tetrahymena thermophila. Eukaryotic Cell 3: 157–169.
14. Wuitschick JD, Gershan JA, Lochowicz AJ, Li S, Karrer KM (2002) A novel family of mobile genetic elements is limited to the germline genome in Tetrahymena thermophila. Nucleic Acids Res 30: 2524–2537.
15. Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, et al. (2006) Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. PLoS Biol 4: e286. doi:10.1371/journal.pbio.0040286.
16. Yao M-C, Chao J-L (2005) RNA-guided DNA deletion in Tetrahymena: an RNAi-based mechanism for programmed genome rearrangements. Annu Rev Genet 39: 537–559. doi:10.1146/annurev.genet.39.073003.095906.
17. Fass JN, Joshi NA, Couvillion MT, Bowen J, Gorovsky MA, et al. (2011) Genome-Scale Analysis of Programmed DNA Elimination Sites in Tetrahymena thermophila. G3 1: 515–522. doi:10.1534/g3.111.000927.
18. Klobutcher LA, Herrick G (1995) Consensus inverted terminal repeat sequence of Paramecium IESs: resemblance to termini of Tc1-related and Euplotes Tec transposons. Nucleic Acids Res 23: 2006–2013.
19. Klobutcher LA, Herrick G (1997) Developmental genome reorganization in ciliated protozoa: the transposon link. Prog Nucleic Acid Res Mol Biol 56: 1–62.
20. Plasterk RH, Izsvák Z, Ivics Z (1999) Resident aliens: the Tc1/mariner superfamily of transposable elements. Trends Genet 15: 326–332.
21. Baudry C, Malinsky S, Restituito M, Kapusta A, Rosa S, et al. (2009) PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. Genes Dev 23: 2478–2483. doi:10.1101/gad.547309.
22. Cheng C-Y, Vogt A, Mochizuki K, Yao M-C (2010) A domesticated piggyBac transposase plays key roles in heterochromatin dynamics and DNA cleavage during programmed DNA deletion in Tetrahymena thermophila. Mol Biol Cell 21: 1753–1762. doi:10.1091/mbc.E09-12-1079.
23. Parfrey LW, Lahr DJG, Knoll AH, Katz LA (2011) Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc Natl Acad Sci USA 108: 13624–13629. doi:10.1073/pnas.1110633108.
24. Gratias A, Bétermier M (2003) Processing of double-strand breaks is involved in the precise excision of paramecium internal eliminated sequences. Mol Cell Biol 23: 7152–7162.
25. Mitra R, Fain-Thornton J, Craig NL (2008) piggyBac can bypass DNA synthesis during cut and paste transposition. EMBO J 27: 1097–1109. doi:10.1038/emboj.2008.41.
26. Kapusta A, Matsuda A, Marmignon A, Ku M, Silve A, et al. (2011) Highly precise and developmentally programmed genome assembly in Paramecium requires ligase IV-dependent end joining. PLoS Genet 7: e1002049. doi:10.1371/journal.pgen.1002049.
27. Preer LB, Hamilton G, Preer JR Jr (1992) Micronuclear DNA from Paramecium tetraurelia: serotype 51 A gene has internally eliminated sequences. J Protozool 39: 678–682.
28. Steele CJ, Barkocy-Gallagher GA, Preer LB, Preer JR Jr (1994) Developmentally excised sequences in micronuclear DNA of Paramecium. Proc Natl Acad Sci USA 91: 2255–2259.
29. Duret L, Cohen J, Jubin C, Dessen P, Goût J-F, et al. (2008) Analysis of sequence variability in the macronuclear DNA of Paramecium tetraurelia: a somatic view of the germline. Genome Res 18: 585–596. doi:gr.074534.107.
30. Arnaiz O, Sperling L (2010) ParameciumDB in 2011: new tools and new data for functional and comparative genomics of the model ciliate Paramecium tetraurelia. Nucleic Acids Res. Available:http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/20952411. Accessed 14 December 2010.
31. Duharcourt S, Keller AM, Meyer E (1998) Homology-dependent maternal inhibition of developmental excision of internal eliminated sequences in Paramecium tetraurelia. Mol Cell Biol 18: 7075–7085.
32. Doak TG, Doerder FP, Jahn CL, Herrick G (1994) A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif. Proc Natl Acad Sci USA 91: 942–946.
33. Jacobs ME, Sánchez-Blanco A, Katz LA, Klobutcher LA (2003) Tec3, a new developmentally eliminated DNA element in Euplotes crassus. Eukaryotic Cell 2: 103–114.
34. Le Mouël A, Butler A, Caron F, Meyer E (2003) Developmentally regulated chromosome fragmentation linked to imprecise elimination of repeated sequences in paramecia. Eukaryotic Cell 2: 1076–1090.
35. Jaillon O, Bouhouche K, Gout J-F, Aury J-M, Noel B, et al. (2008) Translational control of intron splicing in eukaryotes. Nature 451: 359–362. doi:nature06495.
36. DuBois ML, Prescott DM (1997) Volatility of internal eliminated segments in germ line genes of hypotrichous ciliates. Mol Cell Biol 17: 326–337.
37. Dubrana K, Le Mouël A, Amar L (1997) Deletion endpoint allele-specificity in the developmentally regulated elimination of an internal sequence (IES) in Paramecium. Nucleic Acids Res 25: 2448–2454.
38. Gout J-F, Kahn D, Duret L (2010) The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet 6: e1000944. doi:10.1371/journal.pgen.1000944.
39. Arnaiz O, Gout J-F, Betermier M, Bouhouche K, Cohen J, et al. (2010) Gene expression in a paleopolyploid: a transcriptome resource for the ciliate Paramecium tetraurelia. BMC Genomics 11: 547. doi:10.1186/1471-2164-11-547.
40. Duharcourt S, Butler A, Meyer E (1995) Epigenetic self-regulation of developmental excision of an internal eliminated sequence on Paramecium tetraurelia. Genes Dev 9: 2065–2077.
41. Meyer E, Keller AM (1996) A Mendelian mutation affecting mating-type determination also affects developmental genomic rearrangements in Paramecium tetraurelia. Genetics 143: 191–202.

42. Duharcourt S, Lepère G, Meyer E (2009) Developmental genome rearrangements in ciliates: a natural genomic subtraction mediated by non-coding transcripts. Trends Genet 25: 344–350. doi:10.1016/j.tig.2009.05.007.
43. Lepère G, Nowacki M, Serrano V, Gout J-F, Guglielmi G, et al. (2009) Silencing-associated and meiosis-specific small RNA pathways in Paramecium tetraurelia. Nucleic Acids Res 37: 903–915. doi:10.1093/nar/gkn1018.
44. Lepère G, Bétermier M, Meyer E, Duharcourt S (2008) Maternal noncoding transcripts antagonize the targeting of DNA elimination by scanRNAs in Paramecium tetraurelia. Genes Dev 22: 1501–1512. doi:10.1101/gad.473008.
45. Nowacki M, Zagorski-Ostoja W, Meyer E (2005) Nowa1p and Nowa2p: novel putative RNA binding proteins involved in trans-nuclear crosstalk in Paramecium tetraurelia. Curr Biol 15: 1616–1628. doi:10.1016/j.cub.2005.07.033.
46. Epstein LM, Forney JD (1984) Mendelian and non-mendelian mutations affecting surface antigen expression in Paramecium tetraurelia. Mol Cell Biol 4: 1583–1590.
47. Meyer E (1992) Induction of specific macronuclear developmental mutations by microinjection of a cloned telomeric gene in Paramecium primaurelia. Genes Dev 6: 211–222.
48. Sperling L (2011) Remembrance of things past retrieved from the Paramecium genome. Res Microbiol 162: 587–597. doi:10.1016/j.resmic.2011.02.012.
49. Schleif R (1992) DNA Looping. Annual Review of Biochemistry 61: 199–223. doi:10.1146/annurev.bi.61.070192.001215.
50. Lane D, Cavaillé J, Chandler M (1994) Induction of the SOS response by IS1 transposase. J Mol Biol 242: 339–350. doi:10.1006/jmbi.1994.1585.
51. Goryshin IYu, Kil YV, Reznikoff WS (1994) DNA length, bending, and twisting constraints on IS50 transposition. Proc Natl Acad Sci USA 91: 10834–10838.
52. Müller J, Oehler S, Müller-Hill B (1996) Repression of lac promoter as a function of distance, phase and quality of an auxiliary lac operator. J Mol Biol 257: 21–29. doi:10.1006/jmbi.1996.0143.
53. Bellomy GR, Mossing MC, Record MT Jr (1988) Physical properties of DNA in vivo as probed by the length dependence of the lac operator looping process. Biochemistry 27: 3900–3906.
54. Bond LM, Peters JP, Becker NA, Kahn JD, Maher LJ (2010) Gene repression by minimal lac loops in vivo. Nucleic Acids Research 38: 8072–8082. doi:10.1093/nar/gkq755.
55. Lee DH, Schleif RF (1989) In vivo DNA loops in araCBAD: size limits and helical repeat. Proceedings of the National Academy of Sciences 86: 476–480.
56. Haykinson MJ, Johnson RC (1993) DNA looping and the helical repeat in vitro and in vivo: effect of HU protein and enhancer location on Hin invertasome assembly. EMBO J 12: 2503–2512.
57. Bétermier M, Duharcourt S, Seitz H, Meyer E (2000) Timing of developmentally programmed excision and circularization of Paramecium internal eliminated sequences. Mol Cell Biol 20: 1553–1561.
58. Gratias A, Lepère G, Garnier O, Rosa S, Duharcourt S, et al. (2008) Developmentally programmed DNA splicing in Paramecium reveals short-distance crosstalk between DNA cleavage sites. Nucleic Acids Res 36: 3244–3251. doi:10.1093/nar/gkn154.
59. Chandler M, Mahillon J (2002) Insertion Sequences Revisited. Mobile DNA II. Washington, D.C.: ASM Press. pp. 305–366.
60. Paull TT, Haykinson MJ, Johnson RC (1993) The nonspecific DNA-binding and -bending proteins HMG1 and HMG2 promote the assembly of complex nucleoprotein structures. Genes Dev 7: 1521–1534.
61. Tian Z, Rizzon C, Du J, Zhu L, Bennetzen JL, et al. (2009) Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Res 19: 2221–2230. doi:10.1101/gr.083899.108.
62. Garfinkel DJ, Nyswaner KM, Stefanisko KM, Chang C, Moore SP (2005) Ty1 copy number dynamics in Saccharomyces. Genetics 169: 1845–1857. doi:10.1534/genetics.104.037317.
63. Lynch M (2006) The origins of eukaryotic gene structure. Mol Biol Evol 23: 450–468. doi:10.1093/molbev/msj050.
64. Werren JH (2011) Selfish genetic elements, genetic conflict, and evolutionary innovation. Proc Natl Acad Sci USA 108 Suppl 2: 10863–10870. doi:10.1073/pnas.1102343108.
65. Bourc'his D, Voinnet O (2010) A small-RNA perspective on gametogenesis, fertilization, and early zygotic development. Science 330: 617–622. doi:10.1126/science.1194776.
66. Malone CD, Hannon GJ (2009) Molecular evolution of piRNA and transposon control pathways in Drosophila. Cold Spring Harb Symp Quant Biol 74: 225–234. doi:10.1101/sqb.2009.74.052.
67. Baulcombe D (2004) RNA silencing in plants. Nature 431: 356–363. doi:10.1038/nature02874.
68. Prescott DM, Prescott JD, Prescott RM (2002) Coding properties of macronuclear DNA molecules in Sterkiella nova (Oxytricha nova). Protist 153: 71–77.
69. Nowacki M, Vijayan V, Zhou Y, Schotanus K, Doak TG, et al. (2008) RNA-mediated epigenetic programming of a genome-rearrangement pathway. Nature 451: 153–158.
70. Prescott DM (1999) The evolutionary scrambling and developmental unscrambling of germline genes in hypotrichous ciliates. Nucleic Acids Res 27: 1243–1250.
71. Nowacki M, Higgins BP, Maquilan GM, Swart EC, Doak TG, et al. (2009) A functional role for transposases in a large eukaryotic genome. Science 324: 935–938. doi:10.1126/science.1170023.
72. Jahn CL, Klobutcher LA (2002) Genome remodeling in ciliated protozoa. Annu Rev Microbiol 56: 489–520. doi:10.1146/annurev.micro.56.012302.160916.
73. Jahn CL, Doktor SZ, Frels JS, Jaraczewski JW, Krikau MF (1993) Structures of the Euplotes crassus Tec1 and Tec2 elements: identification of putative transposase coding regions. Gene 133: 71–78.
74. Jaraczewski JW, Jahn CL (1993) Elimination of Tec elements involves a novel excision process. Genes Dev 7: 95–105.
75. Klobutcher LA, Turner LR, LaPlante J (1993) Circular forms of developmentally excised DNA in Euplotes crassus have a heteroduplex junction. Genes Dev 7: 84–94.
76. Lambowitz AM, Zimmerly S (2011) Group II Introns: Mobile Ribozymes that Invade DNA. Cold Spring Harbor Perspectives in Biology 3. Available:http://cshperspectives.cshlp.org/content/3/8/a003616.abstract.
77. Sonneborn TM (1974) Paramecium aurelia. Handbook of Genetics. R. King. New York: Plenum Press, Vol. 11. pp. 469–594.
78. Galvani A, Sperling L (2002) RNA interference by feeding in Paramecium. Trends Genet 18: 11–12.
79. Timmons L, Court DL, Fire A (2001) Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in Caenorhabditis elegans. Gene 263: 103–112.
80. Timmons L, Fire A (1998) Specific interference by ingested dsRNA. Nature 395: 854. doi:10.1038/27579.
81. Garnier O, Serrano V, Duharcourt S, Meyer E (2004) RNA-mediated programming of developmental genome rearrangements in Paramecium tetraurelia. Mol Cell Biol 24: 7370–7379. doi:10.1128/MCB.24.17.7370-7379.2004.
82. Preer LB, Hamilton G, Preer JR Jr (1992) Micronuclear DNA from Paramecium tetraurelia: serotype 51 A gene has internally eliminated sequences. J Protozool 39: 678–682.
83. Sambrook J, Fritsch EF, Maniatis T (1989) Molecular Cloning: A Laboratory Manual. 2nd ed. Cold Spring Harbor Laboratory Pr. 1659 p.
84. Weigle J (1966) Assembly of phage lambda in vitro. Proc Natl Acad Sci USA 55: 1462–1466.
85. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754–1760. doi:10.1093/bioinformatics/btp324.
86. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. doi:10.1093/bioinformatics/btp352.
87. Zerbino D, Birney E (2008) Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs. Genome Res: gr.074492.107. doi:10.1101/gr.074492.107.
88. Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12: 656–664. doi:10.1101/gr.229202.
89. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797. doi:10.1093/nar/gkh340.
90. Arnaiz O, Sperling L (2011) ParameciumDB in 2011: new tools and new data for functional and comparative genomics of the model ciliate Paramecium tetraurelia. Nucleic Acids Res 39: D632–636. doi:10.1093/nar/gkq918.
91. Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14: 685–695.
92. R Development Core Team (2011) R: A language and environment for statistical computing. Available:http://www.R-project.org.
93. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20: 289–290.
94. Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188–1190. doi:10.1101/gr.849004.
95. Catania F, Wurmser F, Potekhin AA, Przybos E, Lynch M (2009) Genetic diversity in the Paramecium aurelia species complex. Mol Biol Evol 26: 421–431. doi:10.1093/molbev/msn266.

V.1.2 ParTIES, *Paramecium Toolbox for Interspersed DNA Elimination Studies*

L'article précédent décrit notamment deux méthodes pour l'identification des IES : MIRAA (pour *Method of Identification by Read Alignment Anomalies*) qui détecte les sites d'insertion et MICA (pour *Method of Identification by Comparison of Assemblies*) qui utilise plusieurs stratagèmes d'assemblage pour déterminer la séquence insérée (Arnaiz et al. 2012). Dans le cadre de sa thèse, Cyril Denby Wilkes a développé deux autres méthodes : MIRET (pour *Method of Ies RETention*) pour mesurer le taux de réarrangement des IES et MILORD (pour *Method of Identification and Localization of Rare Deletion*) qui détecte les rares erreurs d'excision d'IES (voir **section III.3.2.1** p.71). Avec Cyril, nous avons décidé d'intégrer ces 4 programmes dans une suite logicielle appelée ParTIES (pour *Paramecium Toolbox for Interspersed DNA Elimination Studies*) et de la rendre disponible à la communauté (`https://github.com/oarnaiz/ParTIES`). La **Figure V.3** (p128) schématise les 4 modules de ParTIES.

La **Figure V.1** (p106) décrit le protocole d'extraction d'ADN de cellules dans un milieu conduisant à l'inactivation d'un gène cible, pendant le cycle sexuel. L'ADN d'intérêt est contenu dans les ébauches en développement. Or, il faut garder en tête que malgré un enrichissement (ou un tri, voir **section V.2** p.135) des ébauches à ∼75n au moment de l'extraction d'ADN, un certain niveau de contamination par les fragments MAC (80on) est probablement inévitable. Donc, l'échantillon d'ADN analysé avec ParTIES contient des molécules d'origine nucléaire et de séquence différentes : des molécules non réarrangées provenant des ébauches et des molécules réarrangées du cycle sexuel précédent provenant des fragments MAC.

MIRAA Au niveau d'un site d'insertion, MIRAA va détecter un excès local d'arrêt d'alignements des lectures de séquençage. En effet, les lectures contenant une partie de l'IES vont pouvoir s'aligner sur la séquence génomique MAC flanquante, mais s'arrêteront aux abords du TA bornant les IES (jonction MAC) (voir **section III.3.2.1** p.68). Sur la **Figure V.3** (p128), cette population de lectures (numéro "4") est baptisée "lecture IES+".

MICA MIRAA ne donne que la localisation de l'IES sur le MAC. Pour accéder à la séquence de l'IES, MICA réalise plusieurs assemblages avec Velvet (Zerbino and Birney 2008) en utilisant différents jeux de lectures : (i) l'ensemble des lectures (1,2,3,4) (ii) l'ensemble des lectures, en soustrayant les lectures traversant parfaitement la jonction MAC (numéro "3" sur la figure) et provenant probablement des fragments parentaux contaminants. Cette filtration a pour but d'aider l'assembleur en limitant les contradictions entre séquences (1,2,4) (iii) uniquement les lectures susceptibles de contenir une partie de la séquence IES (1,4) (voir **Figure V.3** p128). Finalement, par comparaison entre les *contigs* générés (voir **section I.3.3** p.26), contenant en principe la séquence insérée, et le génome de référence (MAC), la séquence de l'IES est déduite.

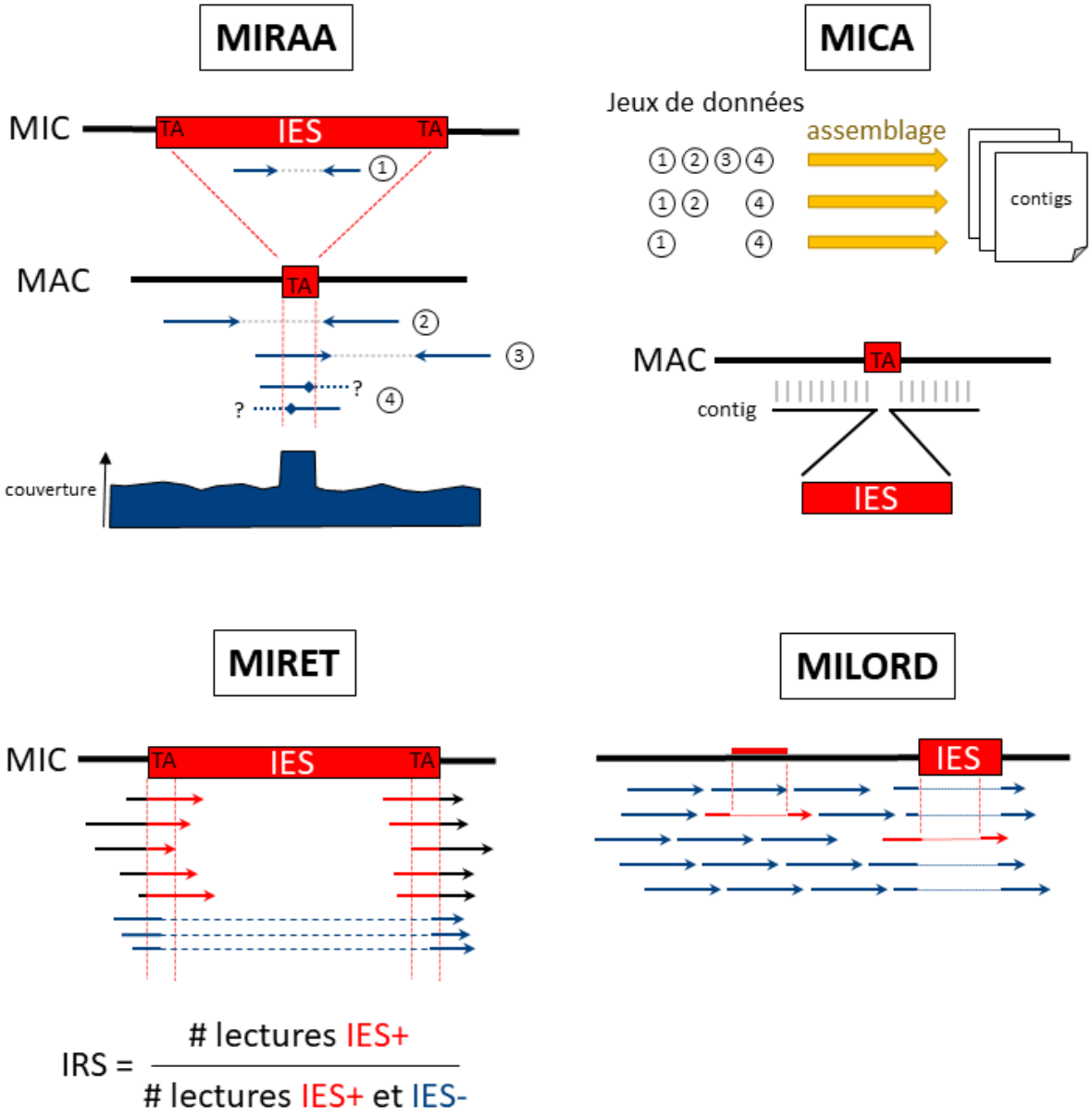


FIGURE V.3 – *Les fonctionnalités de ParTIES*

Les 4 modules de ParTIES : MIRAA, MICA, MIRET et MILORD. Cette figure est commentée dans le texte.

MIRET Dans un contexte sauvage, l'excision des IES est reproductible et efficace mais, comme tout processus biologique, elle reste faillible. Le module MIRET calcule le taux d'excision des IES du génome à partir de données de séquençage. A chaque borne d'IES, l'analyse des alignements permet de compter le nombre de lectures traversant la jonction MAC (lectures IES- ou numéro 3 sur la **Figure V.3**) et le nombre de lectures contenant une partie de l'IES (lectures IES+ ou numéro 4) (voir **Figure V.3** p128). Le score de rétention d'une IES (IRS pour *IES Retention Score*) est calculé en divisant le nombre de lectures IES+ par le nombre de lectures IES+ ou IES-. Un IRS à 0 traduit une excision parfaite et un IRS de 1 une rétention absolue de l'IES. La faible imperfection du système d'excision, dont je parlais précédemment, est visible sur la **Figure V.4A** (p130), où toutes les IES n'ont pas un IRS exactement de 0 dans un contexte sauvage. Comme nous l'avons vu plus haut, l'ADN des ébauches est contaminé par de l'ADN provenant des fragments de l'ancien MAC (voir **section III.2.3.2** p.61 et la **Figure V.1** p106). C'est pourquoi, et même si Pgm est requis pour l'excision de toutes les IES, l'IRS de l'ADN "PGM" n'est pas de 1 (voir **Figure V.4B** p130). Pour chaque expérience, le niveau de contamination peut varier, et plus l'ADN des ébauches sera contaminé par l'ADN des fragments MAC, plus le IRS se rapprochera de 0. Par ailleurs, il s'avère que la déplétion de certains facteurs n'affecte pas l'ensemble des IES (voir **section III.3.2.3** p.75). A titre d'exemple, la **Figure V.4C** (p130) montre la distribution des IRS MIRET dans un contexte où le gène *EZL1* a été inactivé. L'aspect bimodal de l'histogramme révèle la présence d'une classe d'IES insensible ($\sim$1/3 des IES avec un IRS proche de 0) à l'inactivation d'*EZL1* (Lhuillier-Akakpo et al. 2014). MIRET est devenu un outil classique pour analyser l'impact fonctionnel d'un facteur sur les réarrangements programmés des IES.

MILORD La dernière méthode, appelée MILORD, recherche les rares erreurs d'excision (voir **Figure V.3** p128). Elle est basée sur la même idée de recherche de *TA-indels* avec des lectures *Sanger* (Duret et al. 2008) (voir **section III.3.2.1** p.71). MILORD va révéler des événements rares de réarrangements à partir de lectures courtes *Illumina*. MILORD détecte les lectures partiellement alignées sur le génome et tente de repositionner à un autre *locus* génomique la partie non alignée pour inventorier tous les événements de recombinaison chromosomique. La grande profondeur de séquençage des NGS rend l'analyse beaucoup plus sensible qu'avec les lectures *Sanger*.

Ma contribution à cette étude : Pendant sa thèse Cyril Denby Wilkes a développé les logiciels MIRET et MILORD. J'ai encapsulé les 4 modules (MIRAA, MICA, MIRET et MILORD) au sein d'un même programme : ParTIES. J'ai adapté les codes pour rendre les calculs compatibles avec une exécution parallélisée par plusieurs processeurs. J'ai déposé le code sur GitHub. La genèse et l'analyse des données simulées ont été faites par Cyril Denby Wilkes. J'ai participé à la préparation de l'article.

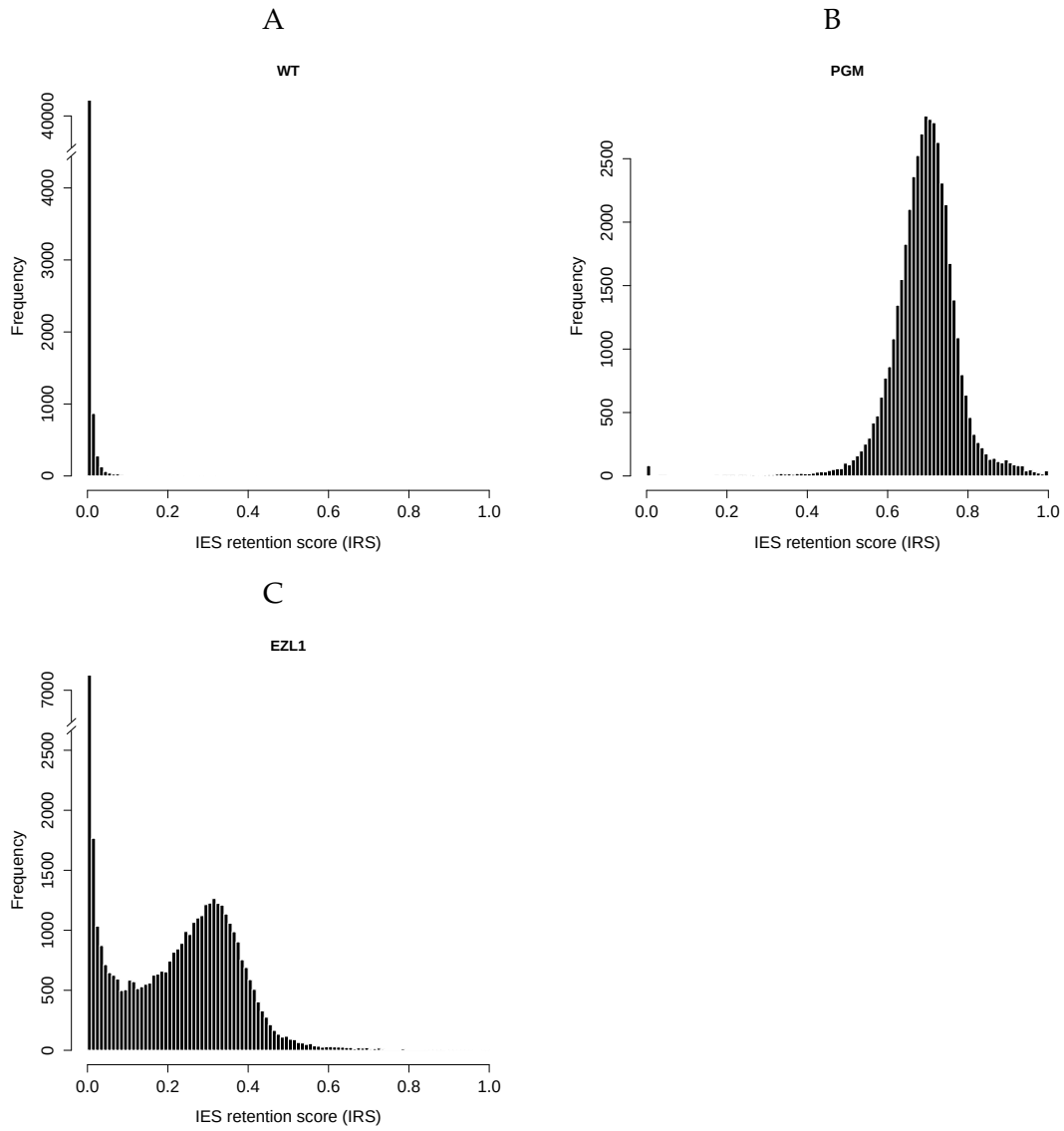


FIGURE V.4 – *Scores de retention d'IES*

Score de retention d'IES (IRS) dans un contexte sauvage (A), dans un contexte PGM déplété (B) et dans une contexte EZL1 déplété (C).

Bioinformatics, 32(4), 2016, 599–601
doi: 10.1093/bioinformatics/btv691
Advance Access Publication Date: 20 November 2015
Applications Note

OXFORD

Genome analysis

ParTIES: a toolbox for *Paramecium* interspersed DNA elimination studies

Cyril Denby Wilkes, Olivier Arnaiz and Linda Sperling*

Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on September 15, 2015; revised on October 28, 2015; accepted on November 14, 2015

Abstract

Motivation: Developmental DNA elimination occurs in a wide variety of multicellular organisms, but ciliates are the only single-celled eukaryotes in which this phenomenon has been reported. Despite considerable interest in ciliates as models for DNA elimination, no standard methods for identification and characterization of the eliminated sequences are currently available.
Results: We present the *Paramecium* Toolbox for Interspersed DNA Elimination Studies (ParTIES), designed for *Paramecium* species, that (i) identifies eliminated sequences, (ii) measures their presence in a sequencing sample and (iii) detects rare elimination polymorphisms.
Availability and implementation: ParTIES is multi-threaded Perl software available at https://github.com/oarnaiz/ParTIES. ParTIES is distributed under the GNU General Public Licence v3.
Contact: linda.sperling@i2bc.paris-saclay.fr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Programmed DNA elimination during somatic development is widely distributed in animal species. First discovered by Boveri in ascaris worms in the 19th century, the phenomenon has to date been characterized in multiple species of nematodes, insects, arachnids, crustaceans, lampreys, fish, birds and mammals, and can involve the programmed reduction of up to 85% of the genome (review: Wang and Davis, 2014). Ciliates, the only unicells that undergo somatic DNA elimination, resemble animals in that they make a germ/soma distinction. One or more diploid germ line nuclei and a polyploid somatic nucleus coexist in a unique cytoplasm. Only the somatic nucleus is transcriptionally active during vegetative growth. As in metazoans, somatic DNA elimination in ciliates can silence transposable elements and cellular genes (Chen *et al.*, 2014), regulate gene dosage (Nowacki *et al.*, 2010) and determine mating type (Cervantes *et al.*, 2013; Singh *et al.*, 2014).

Among ciliates, *Paramecium* is an outstanding model to study DNA elimination. Sexual processes are readily controlled under laboratory conditions and a third of the germ line genome is lost through two types of reproducible, programmed deletions: (i) repeated sequences are heterogeneously eliminated leading to chromosome fragmentation; (ii) single copy elements, called Internal Eliminated Sequences (IESs), are precisely excised. Somatic DNA is also endoreplicated to reach ~ 800 haploid copies. As the 45 000 IESs in the *Paramecium tetraurelia* genome interrupt non-coding and coding sequences (Arnaiz *et al.*, 2012), their elimination is essential to reconstitute open reading frames (ORFs). Both types of DNA elimination depend on a piggyBac domesticated transposase (named PiggyMac) (Baudry *et al.*, 2009), which may be guided by short-RNA driven epigenetic signals (Lepère *et al.*, 2008, 2009).

Cost decrease in High Throughput Sequencing (HTS) is allowing researchers to produce massive genome-wide data to study DNA elimination, but specific bioinformatic methods are still lacking. Here we describe ParTIES: *Paramecium* Toolbox for IES Interspersed DNA Elimination studies. With Illumina DNA-Seq paired-end reads and a somatic reference genome as input, ParTIES performs IES identification, quantitates their presence in the sample and detects rare excision polymorphisms. Benchmarks are provided in Supplementary Materials.

2 Description

Given an input somatic reference genome and any Illumina DNA-Seq sample, be it from germ line or from somatic DNA, ParTIES provides 3 complementary methods which can be run consecutively or independently (Fig. 1).

2.1 Insertion identification

IES identification is important not only for mechanistic studies of DNA elimination, but also to investigate IES origin and evolution across *Paramecium* species. IES identification is a 2-step process. First, an exhaustive list of potential insertion sites is compiled by Method of Identification by Alignment Anomalies (MIRAA). Second, Method of Identification by Comparison of Assemblies (MICA) determines the insertion sequences and their positions. MIRAA uses read mapping to detect an excess of read ends at a given site. Partially mapped reads are presumed to contain additional sequence. Reads that perfectly match the somatic reference at these sites are discarded. The filtered reads are assembled with Velvet (Zerbino and Birney, 2008) to produce contigs potentially containing IESs. MICA carries out global comparison of the contigs and the somatic reference, followed by local realignment to define inserted segments precisely. Optionally for *Paramecium*, the local alignments can be further adjusted to ensure that the ends of the inserted segments conform to the PiggyMac cleavage requirements, namely that insertions be bounded by TA dinucleotides. The IES identification output is a standard GFF3 file.

The above procedure may be simplified if the user already has an assembly containing insertions (such as a germ line reference genome) or if Velvet is not suitable for the available sequencing data, which the user can optionally assemble with a preferred protocol.

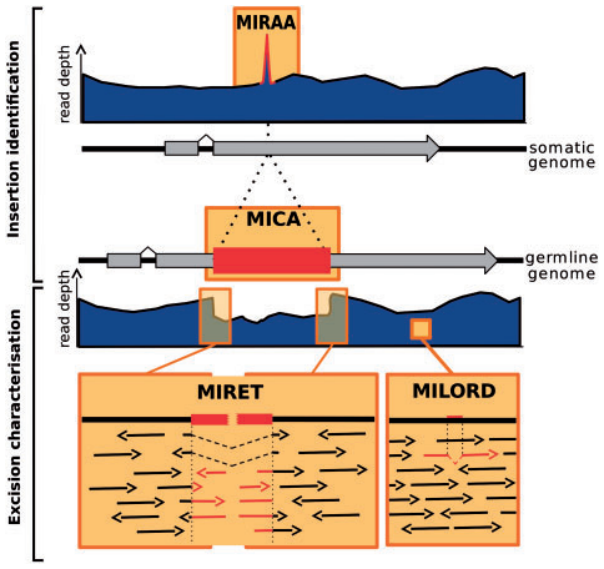


Fig. 1. ParTIES toolbox. Somatic and germ line genomes are evoked by solid black lines with exons (grey boxes) and IES (red box). The dark blue regions represent alignment of Illumina short reads on reference somatic (upper) and germ line (lower) genomes. Arrows in the insets represent mapped reads. MIRAA identifies breakpoints based on excess read coverage and MICA identifies insertions by comparing contigs, assembled from the input reads, to the reference somatic genome; together they output a list of IESs found in a sample. MIRET uses alignments of the short reads to measure IES retention in a sample while MILORD looks for rare deletions in the reads indicative of excision polymorphism

2.2 Excision characterization

2.2.1 IES retention

Many experiments are designed to see whether IESs are excised or retained in the somatic genome after experimental depletion of a factor potentially involved in DNA elimination. The **M**ethod of **IES** **Re**tention (MIRET) module was designed to quantitate the presence of each IES in the genome given a DNA-Seq sample. MIRET uses alignment of the sample reads on the somatic reference to count reads that cross the IES excision junction, designated IES^- reads. MIRET uses alignment of the reads on an IES-containing reference to count reads crossing the junction between an IES and its flanking sequence, designated IES^+ reads. MIRET then calculates a 'boundary score', defined as the ratio of IES^+ reads over the sum of IES^- and IES^+ reads for that boundary. MIRET can also calculate an 'IES retention score' that uses the same counts as the boundary score, with the additional restriction that a read that crosses both ends of an IES is counted only once in the IES retention score calculation.

Determination of the statistical significance of a retention score requires a control sample, provided by sequencing the somatic DNA of untreated cells. Comparison of retention scores of the experimental and control samples is performed to test the null hypothesis that a given IES has the same retention score in both samples. The statistical tests are provided by the R environment and take into account read depth (cf. Sup Mat for details).

2.2.2 Rare deletion events

The IES excision machinery is error-prone and sometimes deletes a segment of somatic DNA or uses an alternative boundary during elimination of a *bona fide* IES (Duret *et al.*, 2008). In order to catalogue these events and evaluate their frequency, we developed the **M**ethod of **I**dentification and **L**ocalization **o**f **R**are **D**eletions (MILORD) module. MILORD looks for a deletion in a read compared to a reference genome. It identifies partially mapped reads and then tries to realign the unmapped part of the read. If a coherent unique alignment is found, a deletion segment is recorded.

3 Discussion

We benchmarked ParTIES using real and simulated data with the *P. tetraurelia* 72 Mb somatic reference genome (Aury *et al.*, 2006) and IES reference set (Arnaiz *et al.*, 2012). The results are presented in Supplementary Materials, and can be used to plan optimal, cost-effective sequencing experiments. We found the minimal requirement for high sensitivity and specificity IES identification and excision quantification is 35× sequencing of a short-insert library of 75 nt paired-end reads, provided the sequencing sample contains at least 25% germ line DNA.

The ParTIES package is expected to set the standard for quantitative analysis of *Paramecium* genomes and DNA elimination.

Acknowledgements

We are grateful to Laurent Duret and Franck Picard for help with statistical tests.

Funding

This work was funded by the CNRS and by the ANR-12-BSV6-0017 'INFERNO' and the ANR-14-CE10-0005-03 'PiggyPack' grants. CDW was supported by a PhD fellowship from the MENRT and by the ANR. This work was carried out in the context of the CNRS-supported European Research Group 'Paramecium Genome Dynamics and Evolution'.

Conflict of Interest: none declared.

References

Arnaiz,O. *et al.* (2012) The Paramecium germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. *PLoS Genet.*, **8**), e1002984.

Aury,J.-M. *et al.* (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature*, **444**, 171–178.

Baudry,C. *et al.* (2009) PiggyMac: a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. *Genes Dev.*, **23**, 2478–2483.

Cervantes,M.D. *et al.* (2013) Selecting one of several mating types through gene segment joining and deletion in *Tetrahymena thermophila*. *PLoS Biol.*, **11**, e1001518.

Chen,X. *et al.* (2014) The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. *Cell*, **158**, 1187–1198.

Duret,L. *et al.* (2008) Analysis of sequence variability in the macronuclear DNA of paramecium tetraurelia: a somatic view of the germline. *Genome Res.*, **18**, 585–596.

Lepère,G. *et al.* (2008) Maternal noncoding transcripts antagonize the targeting of DNA elimination by scanRNAs in *Paramecium tetraurelia*. *Genes Dev.*, **22**, 1501–1512.

Lepère,G. *et al.* (2009) Silencing-associated and meiosis-specific small RNA pathways in *Paramecium tetraurelia*. *Nucleic Acids Res.*, **37**, 903–915.

Nowacki,M. *et al.* (2010) RNA-mediated epigenetic regulation of DNA copy number. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 22140–22144.

Singh,D.P. *et al.* (2014) Genome-defence small RNAs exapted for epigenetic mating-type inheritance. *Nature*, **509**, 447–452.

Wang,J. and Davis,R.E. (2014) Programmed DNA elimination in multicellular organisms. *Curr. Opin. Genet. Dev.*, **27**, 26–34.

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

V.2 Annotation des éléments transposables

Pour mieux comprendre les mécanismes impliqués pendant les réarrangements programmés de la paramécie et plus particulièrement l'élimination spécifique de certaines séquences, il est nécessaire d'avoir la séquence et l'annotation du génome somatique macronucléaire et du génome germinal micronucléaire (voir **section III.3.2** p.68). Le génome MAC et l'annotation des gènes sont disponibles depuis 2006 (Aury et al. 2006, Arnaiz et al. 2017). En revanche, obtenir de l'ADN MIC non contaminé (par de l'ADN MAC ou de l'ADN bactérien) était très difficile, voir impossible avec des techniques classiques de fractionnement cellulaire. En effet, l'ADN germinal ne représente que 0.5% de l'ADN cellulaire. Avec une ploïdie de 800n (contre 2n pour le MIC), la moindre contamination par un noyau MAC est dramatique. En 2012, un artifice génétique consistant à dépléter un gène (PGM) nécessaire aux réarrangements, nous a permis d'obtenir un aperçu du génome non réarrangé. A partir du séquençage de cet ADN PGM, 45 000 IES et 3 types d'ET à ADN (classe 2 de la famille Tc1/Mariner) ont été identifiés (Arnaiz et al. 2012) (voir **section II.2.1.2** p.44).

L'étude qui va suivre tire profit de la technique de tri cellulaire par cytométrie de flux (FACS pour *Fluorescence Activated Cell Sorting*) pour sélectionner des types de noyaux spécifiques (MIC ou ébauche en développement) dans un mélange nucléaire complexe (Guérin et al. 2017). Après le tri des noyaux, l'ADN est séquencé sur une plateforme *Illumina*. Afin de démontrer le bénéfice de la procédure, des ébauches en développement de cellules déplétées en PGM sont triées. Comme attendu, la pureté de l'échantillon est améliorée par rapport à des techniques de purification nucléaire par gradient de densité *Percoll* utilisées dans (Arnaiz et al. 2012). Par ailleurs, Lhuillier-Akakpo et al. (2016) ont montré, qu'un transgène de l'histone H3 centromérique (CenH3 voir **section I.2.1** p.7) fusionné à la GFP marque spécifiquement les micronoyaux. Un tri FACS d'objets basé sur ce signal GFP, la taille des objets, la granularité et le marquage DAPI de l'ADN a permis d'isoler un échantillon contenant environ 500 000 MIC pur à 97%. L'utilisation de ParTIES (Denby Wilkes et al. 2016) sur les données de séquençage de cet échantillon a confirmé la présence de 45 000 IES dans le génome MIC de *Paramecium tetraurelia* et que l'ensemble de ces IES requiert PGM pour leur excision (Arnaiz et al. 2012).

L'assemblage des lectures de l'échantillon MIC a également corroboré l'estimation de la complexité du génome MIC à ∼100Mb (Arnaiz et al. 2012). Malheureusement avec une seule banque de lectures courtes pairées, l'assemblage ne pouvait être que fragmenté par rapport au génome MAC. Pourtant, ce premier assemblage MIC de paramécie a apporté des informations intéressantes. A l'aide d'approches par homologie (voir **section II.2.2.1** p.48), 61 nouveaux ET ont été caractérisés dont les premiers rétrotransposons (RT) de paramécie. Grâce à cette annotation plus exhaustive d'ET, nous avons estimé que près de 10% de la partie du génome MIC éliminée correspondent à des ET et 5% à des satellites (voir **section III.3.2.2** p.74).

L'assemblage MIC a apporté un autre élément inattendu. Par une analyse de couver-

ture de séquençage de l'assemblage MIC, nous avons remarqué que l'ADN MIC n'est pas tout à fait équivalent à l'ADN PGM. En effet, environ 3Mb, constituées principalement de satellites, sont absents de l'ADN d'ébauches de cellules déplétées pour PGM. Cette observation nous conforte dans l'idée que l'ADN d'un échantillon de noyaux MIC triés, plus compliqué à obtenir, est plus complet qu'un ADN PGM.

Ma contribution à cette étude : Mon implication a porté sur des analyses bioinformatiques. J'ai réalisé l'assemblage MIC utilisé dans cette étude. J'ai annoté les IES et les répétitions en tandem. Durant sa thèse, Cyril Denby Wilkes a mis en évidence l'existence de régions du MIC pouvant s'exciser indépendamment de l'activité de Pgm. J'ai formalisé et réalisé l'analyse de couverture différentielle avec des données biologiques de séquençage et avec des données simulées.

Guérin *et al. BMC Genomics* (2017) 18:327
DOI 10.1186/s12864-017-3713-7

BMC Genomics

RESEARCH ARTICLE Open Access

Flow cytometry sorting of nuclei enables the first global characterization of *Paramecium* germline DNA and transposable elements

Frédéric Guérin[1], Olivier Arnaiz[2], Nicole Boggetto[1], Cyril Denby Wilkes[2,3], Eric Meyer[4], Linda Sperling[2] and Sandra Duharcourt[1*]

Abstract

Background: DNA elimination is developmentally programmed in a wide variety of eukaryotes, including unicellular ciliates, and leads to the generation of distinct germline and somatic genomes. The ciliate *Paramecium tetraurelia* harbors two types of nuclei with different functions and genome structures. The transcriptionally inactive micronucleus contains the complete germline genome, while the somatic macronucleus contains a reduced genome streamlined for gene expression. During development of the somatic macronucleus, the germline genome undergoes massive and reproducible DNA elimination events. Availability of both the somatic and germline genomes is essential to examine the genome changes that occur during programmed DNA elimination and ultimately decipher the mechanisms underlying the specific removal of germline-limited sequences.

Results: We developed a novel experimental approach that uses flow cell imaging and flow cytometry to sort subpopulations of nuclei to high purity. We sorted vegetative micronuclei and macronuclei during development of *P. tetraurelia*. We validated the method by flow cell imaging and by high throughput DNA sequencing. Our work establishes the proof of principle that developing somatic macronuclei can be sorted from a complex biological sample to high purity based on their size, shape and DNA content. This method enabled us to sequence, for the first time, the germline DNA from pure micronuclei and to identify novel transposable elements. Sequencing the germline DNA confirms that the Pgm domesticated transposase is required for the excision of all ~45,000 Internal Eliminated Sequences. Comparison of the germline DNA and unrearranged DNA obtained from *PGM*-silenced cells reveals that the latter does not provide a faithful representation of the germline genome.

Conclusions: We developed a flow cytometry-based method to purify *P. tetraurelia* nuclei to high purity and provided quality control with flow cell imaging and high throughput DNA sequencing. We identified 61 germline transposable elements including the first *Paramecium* retrotransposons. This approach paves the way to sequence the germline genomes of *P. aurelia* sibling species for future comparative genomic studies.

Keywords: Flow cytometry, Non-LTR retrotransposons, ITm DNA transposons, Programmed DNA elimination, High throughput sequencing

* Correspondence: sandra.duharcourt@ijm.fr
[1]Institut Jacques Monod, CNRS, UMR 7592, Université Paris Diderot, Sorbonne Paris Cité, Paris F-75205, France
Full list of author information is available at the end of the article

Background

Major genome changes can occur during somatic differentiation. In diverse organisms, programmed DNA elimination leads to the removal of specific-germline DNA sequences during development of somatic cells and thus generates germline and somatic genomes with distinct architectures. This process has been described in a wide variety of animals and in ciliates, suggesting that it has likely arisen independently in different lineages [1]. Ciliates are unicellular eukaryotes with separate germline and somatic nuclei. In the ciliate *Paramecium tetraurelia*, two small, genetically identical diploid micronuclei (MIC, 2n, ~ 3 μm) contain the germline genome that is transmitted to sexual progeny after meiosis. A large, transcriptionally active somatic macronucleus (MAC, 800n, ~ 30 μm) contains a reduced genome streamlined for gene expression. At each sexual cycle, the parental MAC is lost, while new MICs and MACs, destined for the progeny, develop from a copy of the diploid zygotic nucleus. In the new developing MAC, the germline genome is endoreplicated to reach its final ploidy of ~ 800n and undergoes massive programmed DNA elimination (for review [2]) (Fig. 1). Large DNA regions containing transposable elements and other repeated sequences are eliminated, leading to chromosome breakage and *de novo* telomere addition. In addition, ~ 45,000 short, unique, Internal Eliminated Sequences (IESs) are precisely excised. At least 25% of the ~ 100 Mb MIC genome is removed [3]. The distinctive genome architectures of ciliates make them attractive model systems to study the complex mechanisms underlying programmed DNA elimination. Meiosis-specific small RNA and chromatin modification pathways, similar to those found in plants and animals for the formation of heterochromatin and silencing of repeated sequences, are involved in the epigenetic programming of DNA elimination [4, 5].

Comprehensive description of genome changes that occur during programmed DNA elimination requires comparison of the germline and the somatic genomes. While the rearranged somatic MAC genome was sequenced and assembled 10 years ago allowing gene annotation [6], technical difficulties in obtaining pure MIC DNA (0.5% of total genomic DNA) have long been a major obstacle to sequencing the germline genome of *P. tetraurelia*. Pioneering work used Percoll gradient centrifugation to separate MICs from MACs [7]. Despite high bacterial contamination of the resultant MIC DNA [3], this led to the discovery of germline-limited sequences [7, 8]. More recently, DNA enriched in unrearranged germline-like sequences was obtained from cells RNAi-depleted of PiggyMac (Pgm), the domesticated transposase required for developmental genome rearrangements [9]. Deep-sequencing of this DNA (PGM DNA) enabled genome wide-characterization of 45,000 IESs in *P. tetraurelia* [3]. However, how faithfully PGM DNA mimics the true germline genome found in the MIC remains an open question.

We report here a new and reliable method to purify MICs involving a critical step of flow cytometry. The method also allows isolation of developing MACs. Complete separation of nuclei was validated by flow cell imaging and by high throughput DNA sequencing. We show that PGM DNA is in fact not equivalent to MIC DNA. Contigs assembled from the MIC DNA allowed discovery of new *P. tetraurelia* transposable element families.

Results and Discussion

Purification of new developing MACs

Before tackling the purification of the tiny MICs, we decided to purify new developing MACs from cells undergoing the sexual process of autogamy (self-fertilization) (Fig. 1). At each sexual cycle, the parental MAC

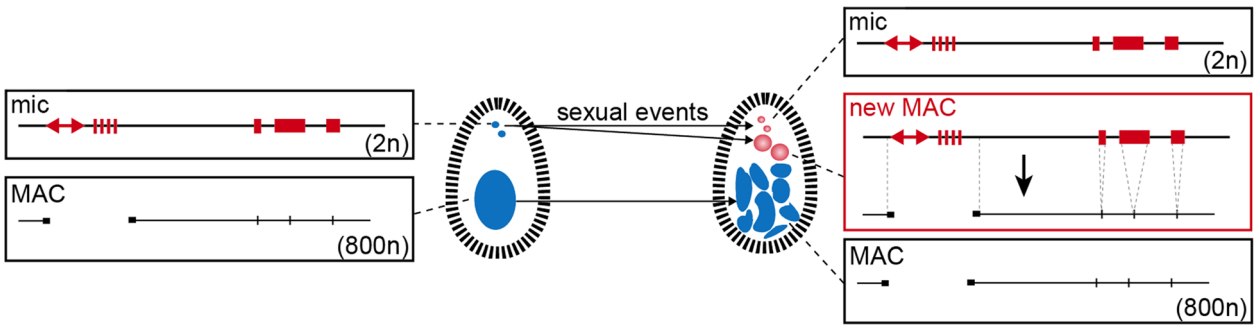


Fig. 1 Nuclear dimorphism and programmed DNA elimination in *Paramecium tetraurelia*. *Left panel*. Each cell contains two distinct types of nuclei: two diploid germline micronuclei (MIC, 2n) and one highly polyploid somatic macronucleus (MAC, 800n). Both nuclei develop from copies of the zygotic nucleus after fertilization. *Right panel*. Massive and reproducible elimination of germline DNA occurs during macronuclear development. Imprecise elimination of germline DNA containing repetitive sequences such as minisatellites (*hatched rectangle*) and transposable elements (*double-headed arrow*) is associated with the fragmentation of germline chromosomes into shorter macronuclear molecules healed by *de novo* telomere addition (*black rectangles*). In addition, 45,000 short, non-coding Internal Eliminated Sequences (IESs) (*red rectangles*) scattered throughout the germline genome are precisely excised from coding and intergenic sequences, restoring open reading frames and allowing gene expression

disintegrates into about 30 small pieces that persist in the cytoplasm, while new MICs and MACs, destined for the progeny, develop from a copy of the diploid zygotic nucleus. Thus, new developing MACs coexist with the two MICs and about 30 small fragments of the maternal MAC (Fig. 2a). We used a published procedure to fractionate the nuclei of Pgm-depleted cells [3] (Fig. 2b). Briefly, nuclei from lysed cells were separated from contaminating organelles and cell debris on a sucrose cushion. The nuclear fraction, containing a mixture of different types of nuclei, was then submitted to flow cytometry (Additional file 1: Figure S1).

A fully developed MAC has a ploidy of 800n [10]. Therefore, new MACs at an advanced developmental stage emit a more intense DAPI (DNA staining) signal than the other nuclei present in the cell at the same stage (MICs and fragments of the maternal MAC). They are also considerably larger than the other nuclei, to accommodate this large amount of chromatin, and are spherical in shape (Fig. 2a). Taking advantage of these characteristics, we FACS-sorted new MACs (~15 μm) according to size (Forward-scattered light, FSC), granularity (Side-scattered light, SSC), pulse width and DAPI signal (Fig. 2c). Purity was measured by flow cell imaging before and after sorting. The developing MAC fraction, that represented 54% of the total nuclear sample before sorting, was enriched to 98% after sorting (Fig. 2d-e). Thus, the sorting procedure conferred considerable improvement over the pre-existing protocol.

To further validate the sorting procedure, we performed high throughput Illumina sequencing of DNA extracted from 266,000 sorted developing MACs ("sorted PGM DNA") (Additional file 2: Table S1). To identify the IESs in a sequencing sample, we used our previously published pipeline [11]. A total of 44,947 IESs was identified in the sorted PGM DNA, compared to 44,928 IESs in unsorted PGM DNA [3]. The fact that 97% ($n = 43{,}839$) of the IESs identified in the sorted PGM DNA correspond to the same IESs identified in unsorted PGM DNA testifies to the reliability of our procedure. The 3% difference lies within the estimated error rate of the method [3, 11].

We then quantified the enrichment of our samples in un-rearranged sequences, by calculating a retention score for each of the 44,928 IES sequences present in the previously published *P. tetraurelia* IES reference set [3]. Retention score values range from 0 for no IES retention to 1 for complete IES retention, when the IES is retained in all sequenced copies of the genomic locus in question. As expected (Fig. 2f), retention score distribution in the rearranged MAC DNA control sample is close to 0 (mean 0.005), whereas a Gaussian distribution is observed for the unsorted non-rearranged PGM DNA, with a mean retention score of 0.69. Even if the Pgm endonuclease is required for all IES excision events, the mean retention score of this sample can never reach 1, because the un-rearranged DNA from the developing new MACs is present in the unsorted sample alongside rearranged DNA from the fragments of the maternal MAC. By contrast, the sorted PGM DNA gave a Gaussian distribution with a mean retention score of 0.82. This higher retention score, obtained from the same starting material, reflects greater enrichment in un-rearranged DNA, and thus in developing nuclei, providing further validation for the superiority of the sorting procedure over the existing protocol. In conclusion, this experiment establishes the proof of principle that nuclei can be sorted from a complex biological sample to high purity based on their size, shape and DNA content.

Purification of MICs from vegetative cells

We used a similar strategy to sort the small germline MICs from vegetative cells (Fig. 3 and Additional file 1: Figure S1). The available MIC isolation method, that relies on Percoll density gradient centrifugation [7], does not provide a MIC fraction sufficiently pure for exclusive MIC genome sequencing, owing to contamination from i) the MAC DNA (800n vs 2n in MIC), and ii) bacteria, on which *Paramecium* cells feed. MIC isolation has been achieved in other ciliates [12–14] but the same methods were not successful in *Paramecium*. We hypothesized that the contamination issues can be solved by the use of a specific fluorophore that is unambiguously and exclusively associated with the MICs. We previously generated transgenic *Paramecium* cells that constitutively express a MIC-localized version of the Green Fluorescent Protein (GFP) fused to centromeric histone H3 (CenH3a) [15]. Transgenic *CENH3a-GFP* cells have green fluorescent MICs, but neither the MAC nor the bacteria are GFP positive (Fig. 3a). We used the same fractionation scheme as the one previously published, with some improvements, to enrich for MICs [7] (Fig. 3b), and submitted the sample to flow cytometry. MICs were sorted based on the SSC, FSC, DAPI (DNA staining) and GFP signals (Fig. 3c-d). The procedure was optimized by flow cell imaging to define the population of interest and refine the sorting parameters (Additional file 1: Figure S1). We obtained 528,000 MICs from 3 million cells.

As previously, purity before and after sorting was measured by flow cell imaging. The MICs represented only 3% of the total sample before sorting and 97% after sorting (Fig. 3e-f). Thus, the sorting procedure is indispensable for effective MIC purification. We performed high throughput Illumina sequencing of the DNA extracted from sorted MICs (528,000 sorted MICs; 60 ng) and from the MIC-enriched sample before sorting. As expected, the bacterial DNA contamination greatly diminished after sorting (8.2% of known contaminants before and 0.2% of

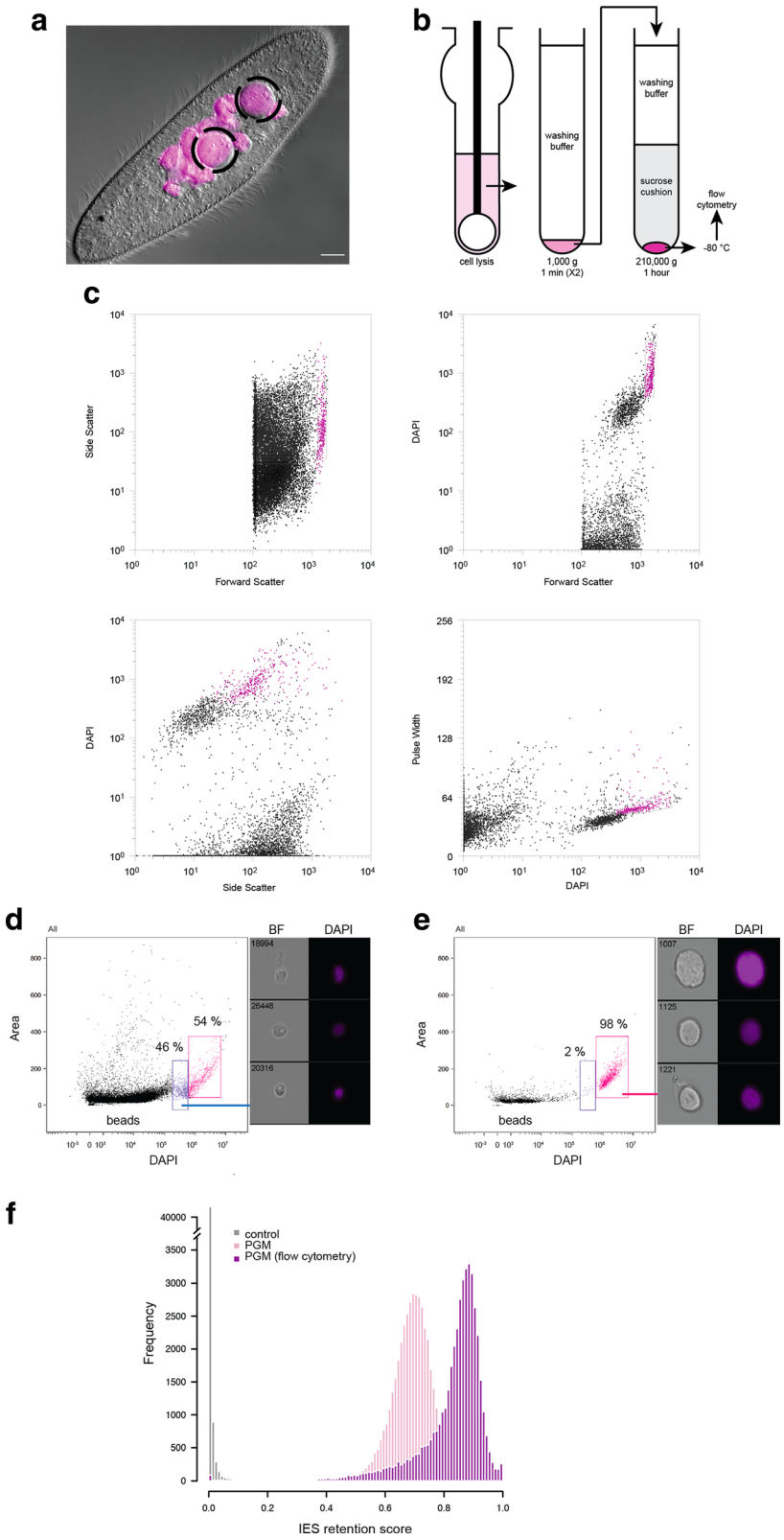


Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 Purification of new developing MACs from *Paramecium tetraurelia* by flow cytometry and validation by flow cell imaging and high throughout DNA sequencing. **a**. DAPI staining of a cell upon *PGM* RNAi at a late developmental stage of the sexual process of autogamy (self-fertilization) is shown on the picture: the two large new developing MACs (*dotted circle*) and the small fragments of the maternal MAC are detected. The scale bar is 10 microns. **b**. Following gentle lysis and cell fractionation, the nuclei preparation is submitted to flow cytometry after staining with DAPI. **c**. Multi-gating flow cytometry strategy used for sorting. Sorting is based on size, granularity and DAPI staining signal of the new developing MACs. An empiric iterative procedure coupled with flow imaging allowed discrimination between developing MACs and fragments, identification of the population of interest, and optimization of the sorting strategy. **d** - **e**. The Amnis ImageStream^X imaging flow cytometer is used for quality control. Distribution of DAPI intensity is shown for each event in the sample before (**d**) and after sorting (**e**), respectively. Representative images are displayed in BF (bright field) and DAPI. Objective ×60. **f.** Validation of the sorting strategy by high-throughput DNA sequencing. Histograms of IES retention scores are shown for control (no RNAi), *PGM* RNAi (no sorting) and *PGM* RNAi after sorting (flow cytometry)

known contaminants after sorting) (Additional file 2: Table S1). We identified 44,851 IESs in the sorted sample, but only 5,192 IESs in the unsorted nuclear fraction. Calculation of mean IES retention scores indicated that enrichment in MIC-limited sequences increased from 0.04 in MIC DNA to 0.38 in sorted MIC DNA (Fig. 3g). The fact that 97% MIC purity only led to approximately 40% MIC DNA is explained by the much higher DNA content of the 3% MAC-derived contaminating fraction. We conclude that flow cytometry sorting is necessary to directly sequence all IESs in unperturbed cells. The fact that 97% (n = 43,741) of the IESs identified in the sorted MIC DNA correspond to the same IESs identified in the sorted PGM DNA confirms that the genome-wide set of IESs in PGM DNA reflects the complete set of MIC IESs (Additional file 1: Figure S2). These data demonstrate that the Pgm domesticated transposase is required for the excision of all IESs.

A first glimpse of the germline genome reveals new transposable elements

The sequence complexity of the MIC assembly is presented in Table 1. Coverage by MAC reads was used to define MAC-destined as opposed to MIC-limited compartments. The 98 Mb assembly consists of 74 Mb (~75%) of MAC-destined sequences and 24 Mb (~25%) of MIC-limited sequences, consistent with the size of the MAC reference genome assembly (72 Mb, [6]). It is important to realize that the MIC assembly we have obtained is highly fragmented (N50 = 37 kb; half of the assembly is in contigs smaller than 37 kb). The most fragmented part of the assembly is the MIC-limited compartment (N50 = 13 kb; half of the MIC-limited sequence is in contigs smaller than 13 kb). With such an assembly, it is possible to annotate germline-limited elements such as IESs and transposable elements (TEs), but not to analyze long-range features such as chromosome structure. For that, additional information, e.g. from mate-pair libraries or third generation long read sequencing, is necessary to handle repeats and build scaffolds.

The MIC assembly consists of all of the contigs assembled using Velvet as launched by ParTIES [11] (Additional file 2: Table S2). The MIC-limited and the MAC-destined parts of the assembly are defined as a function of MAC read depth, using the 3 MAC datasets described in (Additional file 2: Table S1). Any nucleotide with a MAC read depth <20× is considered MIC-limited, else the nucleotide is MAC-destined. N50 means that half an assembly is contained in contigs larger than the N50 value. The MIC-limited part of the assembly is thus much more fragmented than the MAC-destined part. The number of nucleotides covered by Internal Eliminated Sequences (IES), Transposable Elements (TE) and Tandem Repeats (TR) are given. MIC-limited sequences contain almost all IESs and TEs, 95.8% and 92% respectively. The majority (65%) of TR are found in the MIC-limited sequences, however 35%, reflecting WD40, TPR and other repeats, are found in the MAC-destined compartment.

The MIC contigs were used to identify TEs, starting from three previously identified *Paramecium* DNA transposons [3, 16] and a partial reverse transcriptase (RT) consensus (see Methods). tblastn searches using the DDE transposases or RT as queries identified a number of distinct elements, and potentially functional consensus sequences were reconstructed in most cases from the alignment of 10–20 copies (full range 4–48). The majority of TEs (n = 38) are Class I non-LTR retrotransposons, while 13 belong to the IS630-Tc1-mariner (ITm) super-family of Class II DNA transposons. The remaining consensus sequences are putative non-autonomous Class I SINE or solo-ORF1 elements. Characterization of the elements is provided (Additional file 3: Table S3 and Table S4). This analysis significantly augments knowledge of TE in the *Paramecium* germline and presents the first *Paramecium* Class I elements.

The non-LTR retrotransposons all have an ORF2 that contains both apurinic/apyrimidinic endonuclease (APE) and reverse transcriptase (RT) domains, like most known groups of non-LTR retrotransposons [17, 18]. They fall in 5 groups, the first 3 of which also contain an upstream ORF1 (Fig. 4). A phylogeny was built using an alignment of the *Paramecium* RT domains with those of elements belonging to 11 previously characterized major clades [18] (Fig. 4, Additional file 1: Figure S3, Additional file 2: Table S5). The *Paramecium* retrotransposons, along with elements from

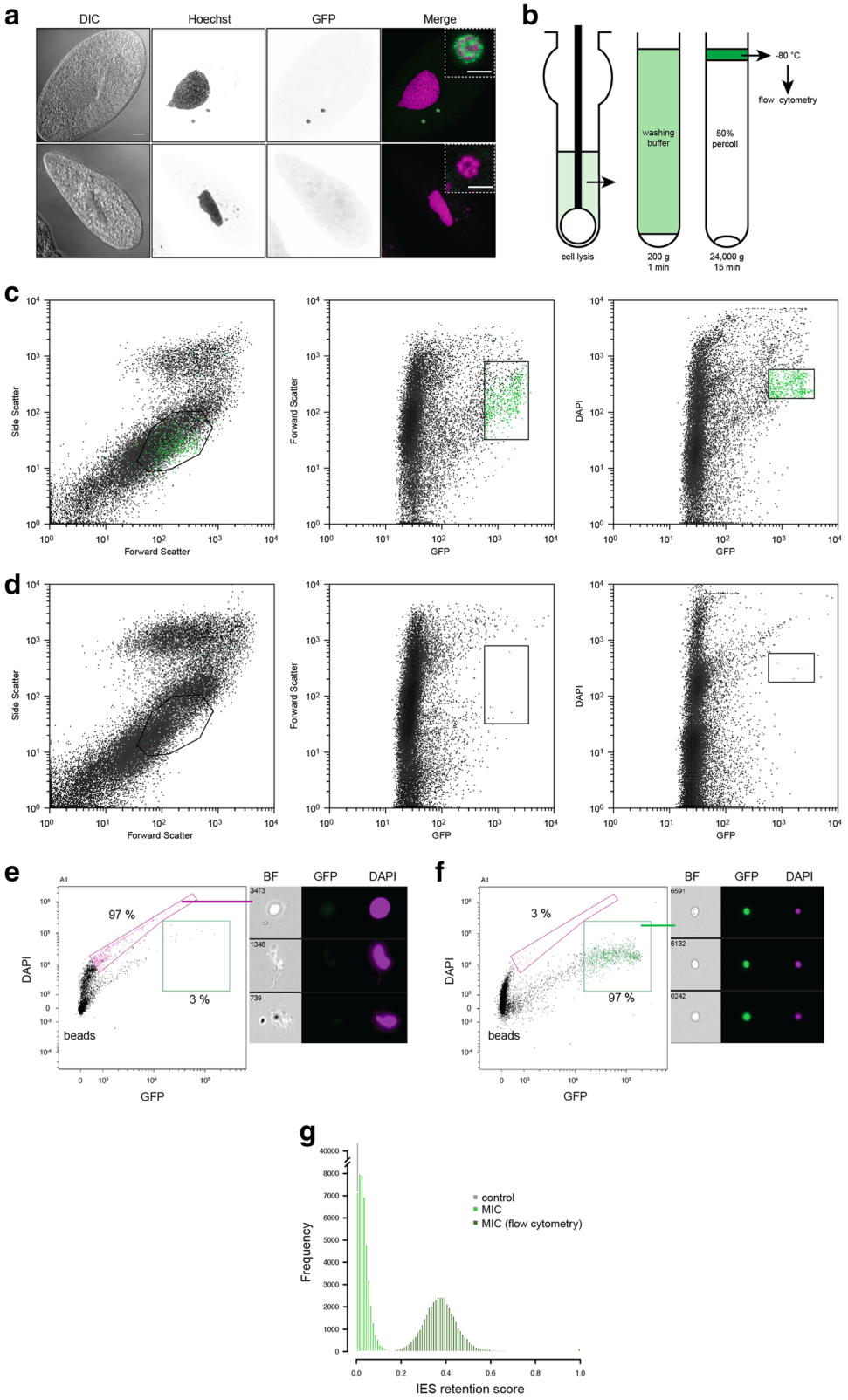


Fig. 3 (See legend on next page.)

(See figure on previous page.)
Fig. 3 Purification of germline MICs from vegetative *Paramecium tetraurelia* by flow cytometry and validation by flow cell imaging and high throughout DNA sequencing. **a**. In *CENH3a-GFP* transgenic *Paramecium* vegetative cells (*upper panels*), but not in control cells (*lower panels*), the MICs are GFP positive. Scale bar is 10 microns. Higher magnification: Scale bar is 3 microns. **b**. Fractionation scheme used to isolate the MIC-enriched fraction. **c** - **d**. Multi-gating strategy used for sorting the MICs. Sorting is based on size, granularity and DAPI staining and GFP signals in **c**) CENH3a-GFP transgenic cells and **d**) control cells. An empiric iterative procedure coupled with flow imaging allowed discrimination between MICs and DAPI containing contaminants, identification of the population of interest, and optimization of the sorting strategy. **e** - **f**. The Amnis ImageStream^X imaging flow cytometer is used for quality control: sample before (**e**) and after sorting (**f**). **g.** Validation of the sorting strategy by high-throughput DNA sequencing. Histograms of IES retention scores are shown for control (no RNAi), MIC (no sorting) and MIC after sorting (flow cytometry)

the ciliate *Tetrahymena thermophila* [19], emerge as a distinct new clade in the tree, with good branch support. The consensus sequences of the first 3 groups, which contain an ORF1, suggest that ORF2 translation depends on +1 ribosomal frameshifting or translation re-initiation (Groups 1 and 2), or on translational read-through of the ORF1 stop codon (Group 3). Like other non-LTR retrotransposons [20, 21], these elements contain short stretches of variable tri-, tetra-, or penta-nucleotide repeats at their 3′ ends (Additional file 3: Table S3). Seven elements (solo ORF1s A-G) appear to contain only an ORF1, ending with a zinc finger similar to that found at the C-terminal end of ORF1 in Groups 1–3, and are likely mobilized in *trans* by proteins encoded by other elements; a (TAAA) n repeat was found at the end of the element in 3 cases.

The 13 DNA transposons, all of the ITm superfamily [22], are unusual in that they contain multiple ORFs (Additional file 3: Table S4). In addition to the DDE ORF common to all ITm elements, an ORF2 of unknown function is found in all *Paramecium* transposons and shares detectable sequence similarity among all of them (Additional file 1: Figure S4). The largest *Paramecium* transposons contain 4 ORFs, ORF4 being a tyrosine recombinase, a property shared with TEC and TBE transposons from distantly related ciliates [23–25]. As seen in the Maximum Likelihood tree built using many ITm DDE domains [22, 26] (Fig. 5, Additional file 1: Figure S5, Additional file 2: Table S6), the composite *Paramecium* elements with a tyrosine recombinase group together, along with TEC1 and TEC2. A distance of 32 aa between the second and third residues of the DDE catalytic triad, characteristic of the 7 tyrosine-recombinase containing *Paramecium* ITm and 3 of the 6 simpler elements, is among the shortest ever reported for ITm.

RepeatMasker was used to identify copies of the TEs in the MIC contigs. Tandem Repeat Finder was used to identify putative satellite sequences (see Methods). As shown in Table 1, 96% of the short unique copy IESs and 92% of the TE copies are in the MIC-limited compartment. However, about one third of tandem repeats were found in the MAC-destined compartment and include WD40, TPR and surface antigen repeats.

MIC and PGM DNA are not equivalent

To compare the sorted MIC DNA with the unrearranged DNA from *PGM*-silenced cells, used until now to represent germline DNA, we calculated the depth of coverage of the MIC assembly by the sorted MIC DNA and the sorted PGM DNA sequencing datasets. The calculation was performed for 90,017 non-overlapping 1-kb windows.

We visualized the comparison between the two datasets by creating dot plots of the depth for each window, and representing the density of the dots using heat map colors. To help interpret the comparison, we simulated PGM and MIC datasets, using enrichments in MIC-limited sequences of 80 and 40% respectively (see Methods). As shown in Fig. 6a left plot, the simulated data present two clouds of points. The larger cloud, with the higher depth of coverage in both samples, corresponds to windows present in both the MIC and the MAC DNA. The smaller cloud, with lower depth of coverage in both samples, represents sequence windows present only in MIC DNA. The real data deviates from this unbiased profile (Fig. 6a, right). The larger clouds representing windows present in both MIC and MAC DNA are comparable (Additional file 1: Figure S6). Surprisingly, the smaller cloud is now vertically elongated, indicating that genome coverage in the PGM

Table 1 Characterization of MIC contigs

	MIC assembly	MAC-destined	MIC-limited
Complexity	98 489 268 bp	74 212 942 bp (75.4%)	24 276 326 bp (24.6%)
N50	37.2 Kb	46.9 Kb	12.7 Kb
GC content	27.40%	27.97%	25.66%
IES	3 517 996 bp	147 387 bp (4.2%)	3 370 609 bp (95.8%)
Transposable Elements	2 973 685 bp	237 838 bp (8%)	2 735 847 bp (92%)
Tandem Repeats	1 393 130 bp	485 112 bp (34.8%)	908 018 bp (65.2%)

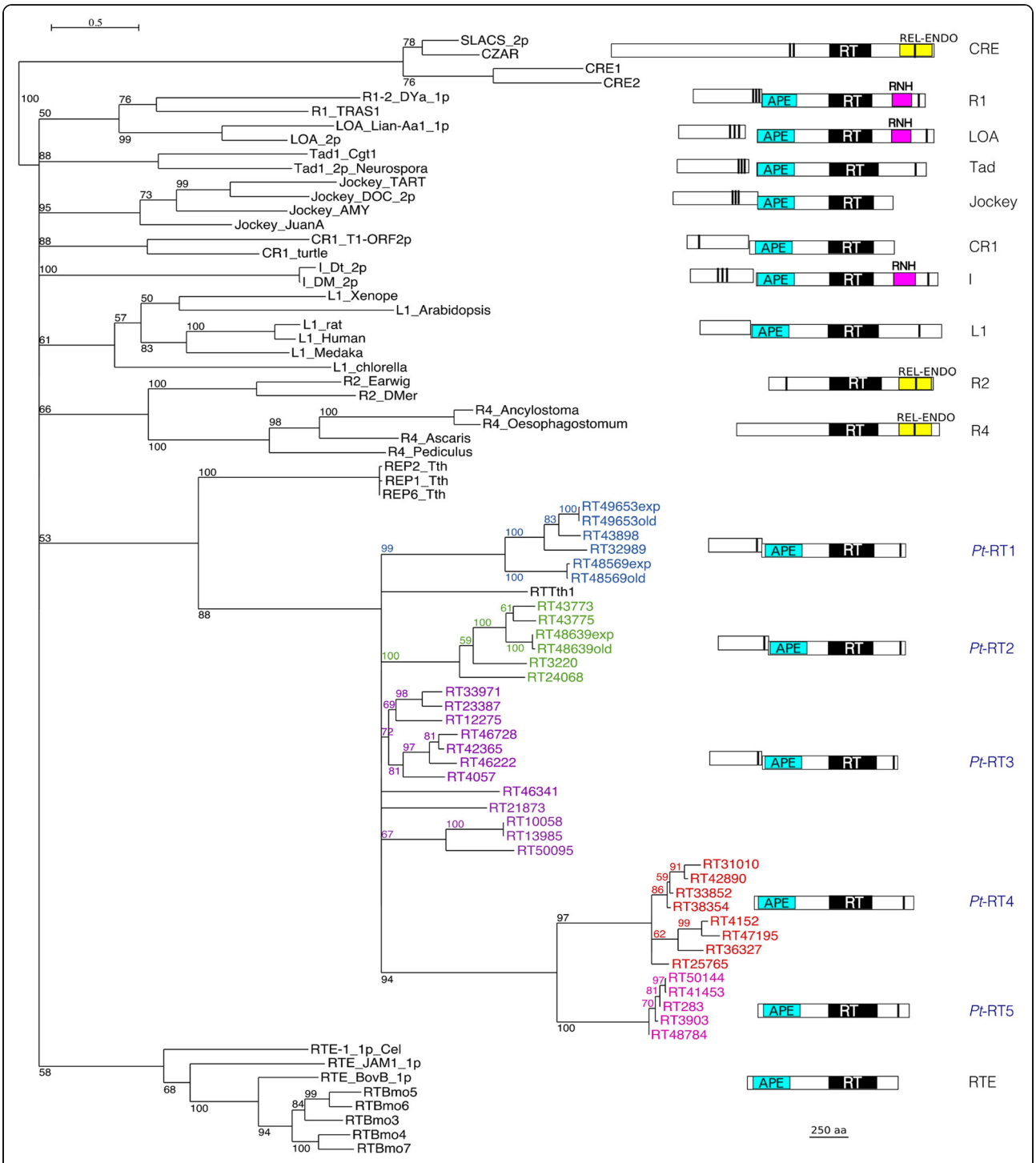


Fig. 4 (See legend on next page.)

(See figure on previous page.)
Fig. 4 Phylogeny of non-LTR elements based on their RT domains. The phylogeny is based on the alignment shown in (Additional file 1: Figure S3) of the ~ 250 aa catalytic RT domains of the elements listed in (Additional file 2: Table S5). The phylogeny is a 50% maximum likelihood tree, rooted with the CRE clade. The numbers at nodes represent the percentage of bootstrap values for 100 replicates. Clade names are prefixed to the element names for the 11 major non-LTR clades. The ciliate non-LTR form a new clade. The names of the elements for the 5 *Paramecium* groups are colored: *blue*, Group 1; *green*, Group 2; *magenta*, Group 3; *red*, Group 4; *pink*, Group 5. The amino acid divergence scale is indicated. Schematic diagrams of ORF structure of representative *Paramecium* elements from each group and representatives of the 11 major clades identified in [18] are shown next to the phylogeny. The representatives are the same as in [18]; however for Tad, Tad1 from *N. crassa* is shown; for R1, TRAS1 from *B. mori* is shown; and for I, the element from *D. melanogaster* is shown. The domains are RT, reverse transcriptase; APE, apurinic/apyrimidinic endonuclease; REL-ENDO, restriction enzyme-like endonuclease; RNH, RNase H domain. Vertical bars represent zinc-finger domains. The two ORFs are shown as offset whether or not they are in the same frame. For Group 1 and Group 2, there is a +1 frameshift. For Group 3, the two ORFs are in the same frame

DNA is variable and mostly less covered than expected (depth between 0 and 7). Both PGM samples behave in the same way. The same windows are found to be under-represented in both PGM and unsorted PGM samples (Additional file 1: Figure S6).

To refine this observation and determine which sequences are missing from the PGM DNA, we used the uniquely-mapped read counts in the 1-kb non-overlapping windows to identify differentially covered windows, in the same way as RNA-Seq counts for genes are used to identify differentially expressed genes (see Methods). The statistical software package we used takes into account the small number of independent samples (2 or 3 biological replicates for most samples, Additional file 2: Table S1).

We looked for windows less covered by MAC or PGM reads with respect to MIC reads (Additional file 1: Figure S7). This allowed us to define three genomic compartments (Fig. 6b, Table 2): 80% of the MIC genome non-overlapping windows are not differentially covered and represent the part of the germline genome that is collinear with MAC chromosomes ("MAC-destined"). The remaining 20% of the windows was significantly less covered by MAC than by MIC reads, corresponding to the germline-limited part of the genome ("MIC-limited"). As anticipated by the previous analysis of read depth, ~ 3% of the windows not covered by MAC reads are not well-covered by the PGM reads. We thus subdivided the MIC-limited compartment into "MIC PGM" and "MIC-only" sub-compartments (Fig. 6b). Figure 6c shows barplots of the normalized read counts of the windows for each of the samples, for the "MIC PGM" and "MIC-only" sub-compartments. As expected, the two sub-compartments are not covered by MAC reads and are well-covered by MIC reads. Interestingly, the "MIC-only" sub-compartment, which is poorly covered by PGM reads, is well-covered by DCL2/3 and EZL1 reads (Additional file 1: Figure S6). These two factors are required for developmental DNA elimination and act respectively in small RNA and histone post-translational modification pathways upstream of the introduction of DNA double-strand breaks by the Pgm endonuclease [27].

Columns from left to right: "MAC-destined" is the genomic compartment covered by MIC, MAC and PGM reads (i.e. windows with no differential coverage according to the DESeq2 analysis, see Methods); "MIC PGM" is the sub-compartment covered by MIC and PGM reads; "MIC only" is the sub-compartment covered only by MIC reads. These compartments are represented schematically in Fig. 6b. The IES reference set was mapped to the MIC assembly and then the IESs were assigned to a window. The total complexity of tandem repeats (micro- and mini-satellite) was calculated using Tandem Repeats Finder. Low complexity sequences identified by Repeat Masker include stretches of poly-purine or poly-pyrimidine and regions of high AT (>87%) or high GC (>89%) content. Repeat Masker was also used to find TE copies, using the TE consensus library reported in this study (See Methods, Additional file 4: Text S1 and Text S2. The difference between the "MIC PGM" and the "MIC only" sub-compartments was judged highly significant for Tandem repeats and for TE (p-value: 9.88e-324 and 9.45e-105, respectively). The MIC only sub-compartment, representing germline-limited sequences not present in either of the PGM samples, is thus enriched in satellites and depleted in TEs.

Different sequence characteristics were calculated for the three genomic sub-compartments (Table 2). GC content and low complexity content did not vary across sub-compartments. Approximately 99% of the IES reference set could be mapped to the MIC assembly. Since 90% of IESs are shorter than 100 bp (median IES size 51 nt) it is not surprising that nearly all IES-containing 1-kb windows are covered by MAC reads and are thus found in the "MAC-destined" compartment.

The TE consensus library was used to find TE copies in the 3 genomic sub-compartments (Table 2). The important difference between the 16 Mb "MIC PGM" and the 3 Mb "MIC-only" sub-compartments is that the latter is significantly depleted in TE copies and enriched in tandem repeats i.e. micro- and mini-satellite (Table 2).

We can suggest two possible, non-exclusive explanations for why 3 Mb of sequence complexity present in

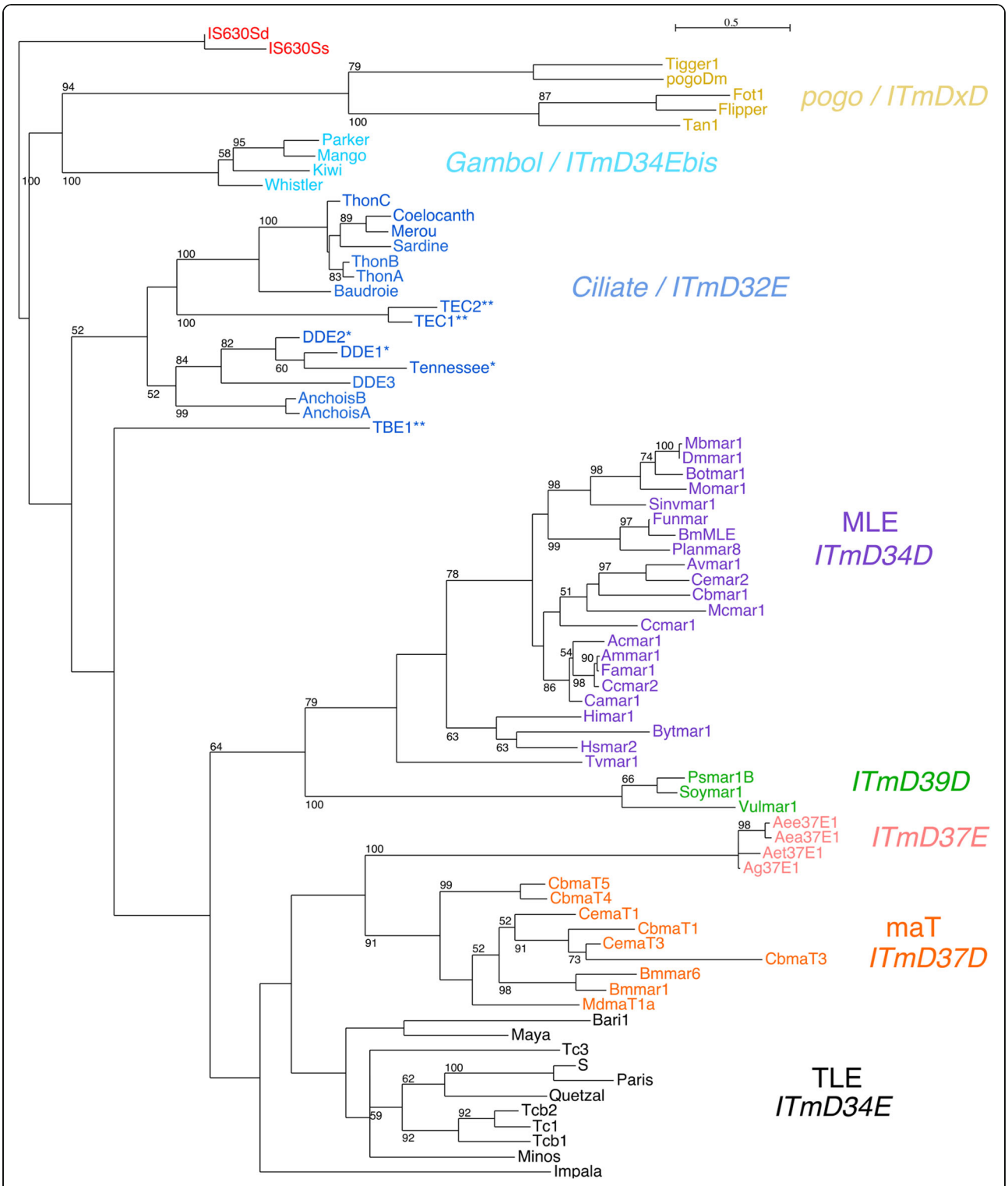


Fig. 5 Phylogeny of ITm elements based on their DDE domains. The phylogeny is based on the alignment shown in (Additional file 1: Figure S5) of the ~ 150 aa catalytic DDE domains of the elements listed in (Additional file 2: Table S6). The phylogeny is a maximum likelihood tree, rooted with bacterial IS630 elements. The numbers at some nodes represent the percentage of bootstrap values for 100 replicates if 50% or greater. Clade names are in bold to the right of the tree. As in [22, 26], the names include the distance between the last two catalytic residues. Most of the elements in the ciliate clade are D32E, however those with one star are D33E and those with two stars are D34E. The amino acid divergence scale is indicated

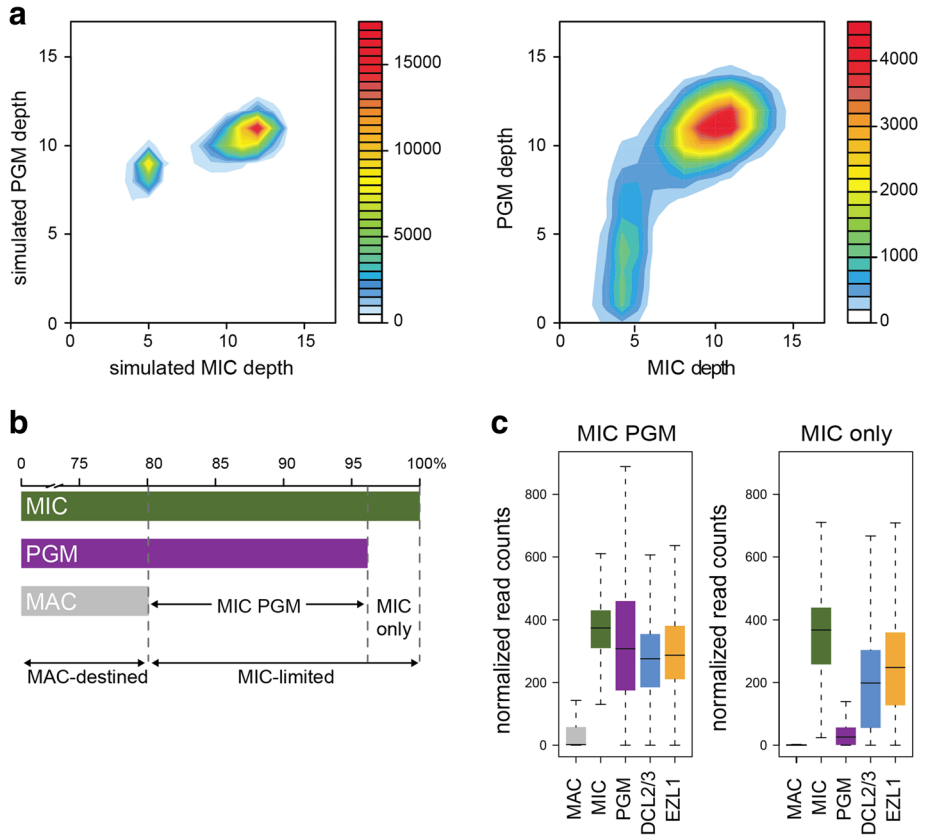


Fig. 6 Coverage of the MIC assembly by different sequencing samples. **a**. Global comparison of the sequences in simulated (*left*) and real (*right*) PGM and MIC sequencing samples. Depth is calculated by mapping reads to the MIC assembly, and counting the reads in 1-kb non-overlapping windows. The graph shows the density of windows as a heat map color, for each combination of MIC and PGM normalized depth values. **b**. Representation of the genomic compartments identified by analysis of differential read coverage of the MIC assembly (cf. Methods, DESeq2 analysis and Table 2). The horizontal bars show the percentage of the MIC assembly covered by each sequencing sample, defining three genomic sub-compartments. "MAC-destined" is the genomic compartment covered by MIC, MAC and PGM reads, i.e. windows with no differential coverage according to the DESeq2 analysis; "MIC PGM" is the compartment covered by MIC and PGM reads; "MIC-only" is the compartment covered only by MIC reads. **c**. Barplots of the normalized DESeq2 read counts, across all windows and all samples (Additional file 2: Table S1) for the "MIC PGM" compartment (*left*) and the "MIC-only" compartment (*right*)

Table 2 Characterization of different sub-compartments of the MIC assembly

	MAC-destined	MIC PGM	MIC only
Complexity	76 130 194 bp	15 983 936 bp	2 905 995 bp
Genome proportion	80.12%	16.82%	3.06%
Longest contiguous region	286 000 bp	58 000 bp	79 000 bp
GC content	27.58%	26.29%	27.47%
IES	97.70%	2.26%	0.04%
Low complexity	5.40%	5.50%	5.43%
Tandem repeats	0.83%	1.39%	5.95%
TE	1.33% (0.97 Mb)	21.47% (3.43 Mb)	7.35% (0.23 Mb)
not TE	98.67% (75.16 Mb)	78.53% (12.56 Mb)	92.65% (2.68 Mb)
TIR	0.45% (0.33 Mb)	4.96% (0.79 Mb)	0.53% (0.02 Mb)
LINE	0.84% (0.62 Mb)	15.98% (2.56 Mb)	6.57% (0.2 Mb)
SINE	0.04% (0.03 Mb)	0.53% (0.08 Mb)	0.25% (0.01 Mb)

the MIC DNA are absent from the PGM DNA: i) the Pgm domesticated transposase is not needed for the elimination of some MIC-limited sequences or ii) *PGM* RNAi is released at the end of development and this release is sufficient for elimination of some MIC-limited sequences. Consistent with the latter explanation, depletion of other factors involved in programmed DNA elimination, and whose function is likely upstream of Pgm, did not lead to underrepresentation of MIC-limited sequences (Fig. 6c and Additional file 1: Figure S6). Whatever the reason of the under-representation of MIC-limited sequences upon *PGM* RNAi, it indicates that PGM DNA, used up until now as a proxy for MIC DNA, does not provide a faithful representation of the MIC genome.

Conclusions

We report the development of an efficient flow cytometry-based method to sort nuclei in *P. tetraurelia*. This method represents a major breakthrough over previously published methods [3, 7], in that it provides (i) improved reliability; (ii) high purity; and (iii) quality control evaluated by flow cell imaging and high throughput sequencing. Our work also provides a clear demonstration that flow cell imaging is a powerful means to detect the population of interest and help refine sorting parameters.

We expect that cytometry-based purification of subpopulations of macronuclei during development may allow kinetic studies of the DNA elimination and endoreplication processes. We have shown that our procedure allows high throughput Illumina sequencing of the *P. tetraurelia* germline genome, paving the way for sequencing the germline genome of other *P. aurelia* sibling species for future comparative genomic studies.

So far, only a few studies have made use of flow cytometry to sort nuclei [28–40], mostly in plants and neurons. Our work highlights the unique potential of flow cytometry to analyze and sort heterogeneous populations of nuclei. It demonstrates that flow cytometry and sorting provide a powerful way to purify minority subpopulations of nuclei, provided that specific nuclear characteristics or a specific fluorophore can be unambiguously and exclusively associated with the subpopulation of interest.

The contigs assembled from the sorted MIC DNA have allowed discovery of 61 germline TEs. The majority are Class I non-LTR retro-transposons (LINE elements), never before characterized in *Paramecium*. This library of manually curated TE consensus sequences constitutes a precious resource for future automated approaches to TE identification and classification in the germline genomes of *Paramecium* species, especially given the relatively large phylogenetic distances to related elements from other taxa.

Methods

Cells and cultivation

All experiments were carried out with the entirely homozygous strain 51 of *Paramecium tetraurelia*. Cells were grown in a wheat grass powder (WGP, Pines International, USA) infusion medium bacterized the day before use with *Klebsiella pneumoniae*, unless otherwise stated, and supplemented with 0.8 mg/L of β-sitosterol (Merck). Cultivation and autogamy were carried out at 27 °C.

Developing MAC purification

We used the feeding method described in [41] to silence the *PGM* gene. *Escherichia coli* HT115 [42] harboring plasmid L4440 [43], with the 567-bp HindIII-NcoI fragment of the *PGM* gene inserted between two convergent T7 promoters [9], was induced for the production of PGM dsRNA in WGP1X medium containing 100 μg/mL ampicillin by overnight growth at 37 °C with shaking. The next day, the culture was diluted into the same medium to OD600 = 0.04. IPTG (Euromedex) was added at a final concentration of 0.4 mM to induce dsRNA synthesis. After 4 h of induction at 37 °C with shaking, the medium was cooled down to 27 °C, and supplemented with 0.8 mg/L of β-sitosterol just before use.

P. tetraurelia cells were first grown in standard *K. pneumoniae* medium for 20–30 vegetative fissions then washed twice in silencing medium. Cells were allowed to grow for 8 to 10 vegetative fissions in a final volume of 3 L of silencing medium (freshly induced medium was added the second day) then starved to trigger autogamy. Progression of autogamy was monitored by Hoechst staining (Sigma). At day 4 of starvation, 30 autogamous cells were picked and transferred individually to 200 μL of *K. pneumoniae* medium to monitor the viability of sexual progeny and evaluate the efficiency of *PGM* silencing. As expected, *PGM* RNAi led to high rates of lethality in the sexual progeny.

At day 4 of starvation, cells were 100% autogamous with about 90% of cells displaying two large developing MACs. Purification of developing new MACs was performed using the protocol described in [3] with minor modifications. Cultures were filtered on 8 layers of sterile gauze. Cells were centrifuged at 600 g for 1 min in an oil-testing centrifuge (Sigma 6–16, rotor 13116) then washed in 100 mL of Tris–HCl 10 mM pH 7.4 and centrifuged again to obtain a compact pellet (~1 mL). After centrifugation, the cell pellet was resuspended in 2 volumes of lysis buffer (~2 mL) (0.25 M sucrose; 10 mM MgCl2; 10 mM Tris pH 6.8; 0.2% NP40) and kept on ice for 5 min. All steps were performed at 4 °C. Cells were then lysed with a Dounce homogenizer until approximately 90% of the cells were broken as observed under a microscope (×20). Developing MACs were collected by centrifugation at 1,000 g for 1 min. The pellet that

contained the developing MACs was washed twice with 9 volumes (~9 mL) of washing buffer. The pellet was then resuspended in 2 mL sucrose solution (2.1 M sucrose; 10 mM MgCl2; 10 mM Tris pH 7.4) and loaded on top of a 3 mL sucrose solution layer in an Ultra-clear centrifuge tube (Beckman Coulter 344059). After gentle addition of washing buffer to fill the tubes, the samples were centrifuged at 210,000 g for 1 h, in a SW41ti swinging rotor (Optima L-80 XP ultracentrifuge, Beckman Coulter). After centrifugation, the sucrose solution was carefully removed. The pellet was gently rinsed with washing buffer, before resuspension into ~ 3 mL of washing buffer containing glycerol (13% final concentration). The samples were aliquoted and frozen at −80 °C.

Micronuclei purification

Transgenic *Paramecium* cells expressing a micronuclear (MIC)-localized version of the Green Fluorescent Protein (GFP) were obtained by microinjection of the vegetative macronucleus with the CenH3a-GFP plasmid, described in [15], in which the centromeric histone variant (CenH3a) gene fused to GFP is expressed under the control of the constitutive promoter of the elongation factor Tu. In the transformed clones, GFP was exclusively found in the MICs. Transformed clones were selected for their GFP signal/noise ratio. Transgene quantification indicated a copy number close to the endogenous CenH3a gene level (transgene/endogenous gene ~ 0.6 to 1). Viability of the sexual progeny after autogamy of the transformed clones was systematically monitored to make sure that the presence of the transgene did not impair the functionality of the MICs.

Transformed and non-injected cells were grown in standard *K. pneumoniae* medium in a final volume of 3 L at a cell density of 1,000 to 1,500 cells/mL. The vegetative state of the cells was assessed by nuclear staining with a 33:1 (vol/vol) mix of carmine red (0.5% in 45% acetic acid) and fast green (1% in ethanol). Detection of GFP signal in the MICs was monitored in the transformed cells. Cultures were filtered on 8 layers of sterile gauze. Cells were centrifuged at 600 g for 1 min in an oil-testing centrifuge (Sigma 6–16, rotor 13116) then washed in 100 mL of Tris–HCl 10 mM pH 7.4 and centrifuged again to obtain a compact pellet.

We used the same fractionation scheme as the one previously published to enrich in MICs [7] with some improvements. After centrifugation, the cell pellet was resuspended in 2 volumes of lysis buffer (0.25 M sucrose; 10 mM MgCl2; 10 mM Tris pH 6.8; 0.2% NP40) and kept on ice for 5 min. All steps were performed at 4 °C. Cells were then lysed with a Dounce homogenizer until approximately 90% of the cells were broken as observed under a microscope (×20). Three volumes of washing buffer (0.25 M sucrose; 10 mM Tris pH 7.4; 5 mM MgCl2; 15 mM NaCl; 60 mM KCl; 0.5 mM EGTA) were added. The sample was dispatched into 2 mL Eppendorf tubes and mixed by inversion 5 times then centrifuged at 200 g for 1 min. The supernatant that contained most MICs was recovered and presence of the MICs was verified under a microscope. The supernatant was then transferred into Ultra-clear centrifuge tubes (Beckman Coulter 344059, 2 mL per tube), and 10 mL of 50% Percoll solution (50% Percoll pH 7.5; 0.25 M Sucrose; 10 mM MgCl2) were added drop by drop with gentle agitation. The supernatant and the Percoll solution were gently mixed by pipetting and centrifuged at 24,000 g for 15 min in a SW41Ti swinging rotor (Optima L-80 XP ultracentrifuge, Beckman Coulter). During centrifugation, the Percoll gradient is formed and MICs accumulated at the top of the gradient and MACs at the bottom. After centrifugation, MICs were carefully recovered in a white-to-brown powderous band with a 200 μL Pipetman into a 1.5 mL Eppendorf tube. The MIC-enriched sample was gently mixed then diluted 1/1/1 with washing buffer and glycerol 40% (13% glycerol final concentration). Usually several hundred MICs per microliter could be counted under a microscope. The samples were aliquoted and frozen at – 80 °C for further flow cytometry analysis and sorting.

Flow cytometry

Samples of MICs and developing MACs were thawed on ice, diluted 1/5 to 1/10 in washing buffer and stained with DAPI (3 μM final, Invitrogen #D3571). All steps were performed at 4 °C. The samples were filtered (30 μm Sysmex filters, 04-004-2326) and sorted on an Influx 500 cell sorter (BD Biosciences) with a 488 nm laser for scatter measurements (Forward Scatter, or FCS, and Side Scatter, or SSC) and GFP excitation, and a 355 nm laser for DAPI excitation. GFP and DAPI staining signals were collected using a 528–38 nm band pass filter and a 460–50 nm band pass filter, respectively. Phosphate Buffered Saline (Isoflow TM Sheath Fluid, Beckman Coulter) was used as sheath and run at a constant pressure of 15 PSI. Frequency of drop formation was 27 kHz. The instrument used a 100 μm nozzle. For the MIC samples, a threshold on the GFP signal was optimized to increase collecting speed (2500 events per second). For developing MACs, an important threshold on FCS was optimized to not consider the crystals present in the sample and increase collecting speed. *Paramecium* cells contain crystals composed of guanine, xanthine and hypoxanthine [44], which are pelleted together with developing MACs during the purification procedure and can represent an important part of the elements detected by the instrument. Since they do not contain DNA, hiding crystals allowed a faster collecting speed without increasing DNA contamination. Sorting rates typically ranged from

10,000 to 100,000 MICs per hour depending on the preparation. Data were collected using Spigot software. Micronuclei were sorted based on their SSC, FSC, GFP and DAPI signals. Events in GFP and DAPI gates were backgated onto FSC vs SSC to optimize the sorting. Developing MACs were sorted based on their SSC, FSC, DAPI, and time-of-flight (pulse width) signals. Events with high DPAI signal were backgated onto FSC vs SSC to optimize the sorting. Nuclei were recovered in washing buffer into a 1.5 mL Eppendorf tube.

Flow cell imaging

Purity of the sorted samples was evaluated by flow cell imaging. Samples before and after sorting were imaged on a 2 camera, 12 channel ImageStream^X (Amnis/MerckMillipore) imaging flow cytometer with a 60× magnification, using 405, 488, and 785 nm lasers, at respectively 125, 100, and 0.05 mW. Phosphate Buffered Saline (137 mM NaCl; 2.7 mM KCl; 6.7 mM Na2HPO4; 1.5 mM KH2PO4) was used as sheath. Acquisitions were performed using Inspire software. Brightfield was collected in channel 1 and 9, SSC in channel 6 (745–800 nm bandwidth), GFP in channel 2 (480–560 nm bandwidth), and DAPI in channel 7 (430–505 nm bandwidth). At least 5,000 elements were analyzed for each sample (before and after sorting) in order to detect enough MICs, given the rarity of MICs in the sample (~0.2–3% of all events detected by the Influx cell sorter before sorting). Cell classifiers were set for channel 1 area lower limit of 10 to allow the instrument to focus despite low concentration of the sample after sorting. Beads were excluded from the analysis based on their low DAPI and GFP fluorescence signals. Analysis was performed using the IDEAS software.

Genomic DNA extraction and sequencing

After sorting, MICs and developing MACs were treated with 3 volumes of proteinase K solution (0.5 M EDTA pH 9; 1% N-lauroylsarcosine; 1% SDS; 1 mg/mL proteinase K) at 55 °C overnight. Genomic DNA was extracted with the addition of one volume of Tris–HCl-phenol pH 8 with gentle agitation at room temperature for 1 h (no vortex). After centrifugation at 300 g for 15 min, the aqueous phase was recovered, dialyzed twice against TE (10 mM Tris–HCl; 1 mM EDTA, pH 8) 25% ethanol for 2 h, against TE overnight, then against Tris 1 mM pH 8 for 2 h. DNA was concentrated with a Concentrator plus (Eppendorf) down to 50 to 100 μL. DNA concentration was quantified using QuBit High sensibility kit (Invitrogen) and stored at 4 °C. DNA was then sequenced by a paired-end strategy using Illumina Hi-Seq next-generation sequencer (Additional file 2: Table S1). DNA-seq datasets have been deposited at the NCBI short read archive (SRA) (Accession numbers: SAMN05323659; SAMN05323660; SAMN05323661).

Transposable element annotation

Putative LINE elements were discovered as follows. Reverse transcriptase coding domains were identified from a small cluster of homologous sequences retained in the MAC genome, after building a consensus from their alignment. These partial peptide sequences were then used to search the MIC contigs (tblastn using default parameters, with no low complexity filter). The matches were culled and used to extend the consensus protein sequences. Then blastn searches (default parameters, no low complexity filter) were used against the MIC contigs to find more copies. The procedure was used recursively to extend and find as many copies as possible. Copies were aligned with MUSCLE [45] and adjusted manually, with a requirement of potentially functional ORF1 and ORF2 sequences. Finally, the best adjusted consensus sequences were used to search for other related elements by a tblastn search for long, poorly scoring matches which might be recent copies of a different element. In this way, 5 distinct groups of LINE elements were found. A similar procedure was used to annotate Class II DNA transposons, starting from published sequences for the *P. primaurelia* Tennessee element ORFs [16] and the *P. tetraurelia* Sardine and Anchois element ORFs [3]. Finally, some sequences inserted in other elements were found to be present in multiple copies in the MIC assembly but yielded consensus sequences with no protein-coding potential; these sequences were annotated as putative SINE elements. Fasta files with the nucleotide and putative peptide sequences are provided (Additional file 4: Text S1-S2), (Additional file 3: Tables S3-S4).

Phylogenetic tree reconstruction

Non-LTR Class I retrotransposon ORF2 (pol) protein sequences representative of different clades [18] and IS630-Tc1-mariner (ITm) superfamily transposase protein sequences [22, 26] were recovered from GenBank or RepBase (Additional file 2: Tables S5-S6). Corresponding *Paramecium* consensus sequences were added to each set of proteins. The proteins were aligned using MSAProbs [46]. The alignments were trimmed manually to correspond to the RT and DDE catalytic domains, respectively (Additional file 1: Figures S3 and S5) and used for phylogenetic tree reconstruction by Maximum Likelihood [47, 48], with PhyML version 3.1 (PhyML -d aa -m LG -v 0.0 -c 4 -a E -f M –no_memory_check -i < phylip_alignment_file > –b 100). The non-LTR retrotransposon RT tree was collapsed if branch support (determined using 100 bootstrap replicates) was less than 50%, using TreeGraph2 [49]. Seaview [50] was used

for preliminary tree-building, to convert alignment formats and to visualize, re-root, swap branches and prepare figures of the trees.

Bioinformatic analyses

IES retention

IES retention scores were calculated with ParTIES v1.0 [11] (MIRET module, –max_mismatch 1 –score –method Boundaries) using the *P. tetraurelia* IES reference set [3] and two reference genome assemblies available from ParameciumDB (http://paramecium.cgm.cnrs-gif.fr/download/fasta/assemblies/): ptetraurelia_mac_51.fa and ptetraurelia_mac_51_with_ies.fa. The score for each IES corresponds to the mean of the two boundary scores.

Assembly of MIC reads

The MIC flow cytometry sequencing reads (acc. no. SAMN05323660; Additional file 2: Table S1) were assembled into contigs using ParTIES v1.0 [11] (default parameters except for the Assembly module, –k 51). ParTIES filters out reads that contain a MAC IES junction using the MAC reference genome prior to a Velvet (version 1.2.10) [51] assembly. Assembly statistics for the resulting MIC contigs are given in (Additional file 2: Table S2).

Analysis of depth

The MIC contigs (Additional file 2: Table S2) were used as reference genome for this analysis. The contigs were divided into 1-kb non-overlapping windows. For each sequencing sample, the mean depth for each window was calculated with Samtools [52] depth (v0.1.18 –q 30 –Q 30) on Bowtie 2 [53] (v2.2.3 –local –× 800) mappings. The mean depth was normalized according to the number of nucleotides sequenced in the sample, after excluding reads which match known contaminants (mitochondrial DNA, rDNA, bacterial genomes).

Sequencing simulation

We simulated sequencing data using ART version 2.3.7 [54] (––noALN –len 100 –seqSys HS10 –qShift 90 –qShift2 90 –mflen 300 –sdev 100). We specifed coverage using the fcov parameter, to obtain final coverage of 100×. Thus, to obtain a dataset with 40% enrichment in MIC sequences, we simulated 40× coverage on the MIC assembly and 60× coverage on the MAC assembly and pooled the simulated reads. The analysis of depth was applied to the simulated read datasets.

Differential coverage analysis

DESeq2 software [55] was designed for differential analysis of NGS count data, and is typically used for gene expression studies i.e. to compare RNA-Seq read counts for genes across experimental conditions. We used DESeq2 (v. 1.14.0) to compare DNA-Seq read counts for non-overlapping MIC windows (1 kb windows and >400 bp windows at contig ends) across samples. For each sample (Additional file 2: Table S1), we provide to DESeq2 the number of uniquely mapping reads in each window. We considered windows with a fold-change >2 between MIC and other samples and an adj.p-value < 0.05 to be differentially covered. Barplots (Fig. 6c) used the normalized counts determined for each sample by DESeq2.

Sequence properties

For selected windows (see text and Table 2), tandem repeats (micro- and mini-satellite) were identified using Tandem Repeats Finder [56] (version 4.07b, TRF parameters: 2 7 7 80 10 50 500) and the corresponding complexity determined using the R Bioconductor package "GenomicRanges_1.26.1" [57]. RepeatMasker [58] (version 3.3.0) was used to identify low complexity sequences (RepeatMasker -noint –no_is –s) and transposable elements (TE; RepeatMasker -nolow –no_is –s –lib < TE consensus library>). The TE consensus library is that reported in this study (Additional file 3: Tables S3-S4). We performed exact binomial tests using the R package binom_1.1–1 [59].

Additional files

Additional file 1: This PDF contains the following supplementary figures: **Figures S1- S7.** Legends for these figures appear at the beginning of Additional file 1 (PDF 12230 kb)

Additional file 2: This word file contains the following supplementary tables: **Tables S1, S2, S5, S6.** (DOCX 110 kb)

Additional file 3: This excel file contains the following supplementary tables: **Tables S3-S4.** (XLSX 45 kb)

Additional file 4: This text file contains the following supplementary text: Text S1-S2. Text S1 is a fasta file of TE consensus nucleotide sequences. Text S2 is a fasta file of putative TE protein sequences. (TXT 280 kb)

Abbreviations

APE: Apurinic/apyrimidinic endonuclease; FSC: Forward-scattered light; GFP: Green fluorescent protein; IES: Internal Eliminated Sequence; ITm: IS630-Tc1-mariner; MAC: Macronucleus; MIC: Micronucleus; RT: Reverse transcriptase; SSC: Side-scattered light; TE: Transposable element

Acknowledgements

We thank Jean Cohen for his advice during the development of the purification procedure. We thank Isadora Cohen for critical reading of the manuscript and the SD lab members for support and discussion. We acknowledge the ImagoSeine facility, member of the France BioImaging infrastructure supported by the ANR-10-INSB-04. The sequencing benefited from the facilities and expertise of the high-throughput sequencing platform of I2BC.

Funding

This research was supported by intramural funding from the CNRS, grant ANR-12-BSV6-0017 "INFERNO" to S. D., L. S. and E.M., program "Investissements d'Avenir" launched by the French government and implemented by ANR with the references ANR-10-LABX-54 MEMOLIFE and ANR-11-IDEX-0001-02 PSL Research University, an 'Equipe FRM DEQ20150331763' grant to EM, grant ANR-14-CE10–0005–04 'PIGGYPACK' to S. D. and L. S., grants from LABEX "Who am I?" grant supported by the ANR-11-LABX-0071_WHOAMI and the

ANR-11-IDEX-0005-02, from 'Comité d'Ile de France de la Ligue Nationale Contre le Cancer', and an 'Equipe FRM DEQ20160334868' grant to S. D.
The funding bodies had no role in the design of the study, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
All data generated or analysed during this study are listed in (Additional file 2: Table S1) (accession numbers to DNAseq datasets) or included in this published article and its supplementary information files.

Authors' contributions
FG performed all the experiments. NG carried out the sorting experiments. OA, CDW and LS performed the bioinformatic analyses. EM identified and annotated transposable elements. SD conceived the study and wrote the paper. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Dedication
We dedicate this work to the late John R. Preer Jr., who, with his wife Bertie, pioneered *Paramecium* nuclear purification and germline DNA characterization more than twenty years ago. John passed away in April 2016.

Publisher's Note

Author details
[1]Institut Jacques Monod, CNRS, UMR 7592, Université Paris Diderot, Sorbonne Paris Cité, Paris F-75205, France. [2]Institute of Integrative Biology of the Cell, UMR9198 CNRS CEA Univ, Paris-Sud Université Paris-Saclay, 91198 Gif-sur-Yvette, France. [3]Current address: Institut de Biologie et de Technologies de Saclay (IBITECS), CEA, F-91191 Gif-sur-Yvette Cedex, France. [4]IBENS, Département de Biologie, Ecole Normale Supérieure, CNRS, Inserm, PSL Research University, F-75005 Paris, France.

Received: 24 December 2016 Accepted: 20 April 2017
Published online: 26 April 2017

References
1. Wang J, Davis RE. Programmed DNA elimination in multicellular organisms. Curr Opin Genet Dev. 2014;27C:26–34.
2. Betermier M, Duharcourt S. Programmed Rearrangement in Ciliates: Paramecium. Microbiol Spectr. 2014;2.
3. Arnaiz O, Mathy N, Baudry C, Malinsky S, Aury J-M, Wilkes CD, et al. The Paramecium germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. PLoS Genet. 2012;8:e1002984.
4. Holoch D, Moazed D. RNA-mediated epigenetic regulation of gene expression. Nat Rev Genet. 2015;16:71–84.
5. Coyne RS, Lhuillier-Akakpo M, Duharcourt S. RNA-guided DNA rearrangements in ciliates: Is the best genome defence a good offence? Biol Cell. 2012;104:1–17.
6. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, et al. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature. 2006;444:171–8.
7. Preer LB, Hamilton G, Preer JR. Micronuclear DNA from Paramecium tetraurelia: serotype 51 A gene has internally eliminated sequences. J Protozool. 1992;39:678–82.
8. Steele CJ, Barkocy-Gallagher GA, Preer LB, Preer JR. Developmentally excised sequences in micronuclear DNA of Paramecium. Proc Natl Acad Sci U S A. 1994;91:2255–9.
9. Baudry C, Malinsky S, Restituito M, Kapusta A, Rosa S, Meyer E, et al. PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. Genes Dev. 2009;23:2478–83.
10. Berger JD. Selective inhibition of DNA synthesis in macronuclear fragments in Paramecium aurelia exconjugants and its reversal during macronuclear regeneration. Chromosoma. 1973;44:33–48.
11. Denby Wilkes C, Arnaiz O, Sperling L. ParTIES: a toolbox for Paramecium interspersed DNA elimination studies. Bioinforma Oxf Engl. 2016;32:599–601.
12. Chen X, Bracht JR, Goldman AD, Dolzhenko E, Clay DM, Swart EC, et al. The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. Cell. 2014;158:1187–98.
13. Fass JN, Joshi NA, Couvillion MT, Bowen J, Gorovsky MA, Hamilton EP, et al. Genome-Scale Analysis of Programmed DNA Elimination Sites in Tetrahymena thermophila. G3 Bethesda Md. 2011;1:515–22.
14. Hamilton EP, Kapusta A, Huvos PE, Bidwell SL, Zafar N, Tang H, et al. Structure of the germline genome of Tetrahymenathermophila and relationship to the massively rearranged somatic genome. elife. 2016;5.
15. Lhuillier-Akakpo M, Guérin F, Frapporti A, Duharcourt S. DNA deletion as a mechanism for developmentally programmed centromere loss. Nucleic Acids Res. 2016;44:1553–65.
16. Le Mouel A, Butler A, Caron F, Meyer E. Developmentally regulated chromosome fragmentation linked to imprecise elimination of repeated sequences in paramecia. Eukaryot Cell. 2003;2:1076–90.
17. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.
18. Malik HS, Burke WD, Eickbush TH. The age and evolution of non-LTR retrotransposable elements. Mol Biol Evol. 1999;16:793–805.
19. Fillingham JS, Thing TA, Vythilingum N, Keuroghlian A, Bruno D, Golding GB, et al. A Non-Long Terminal Repeat Retrotransposon Family Is Restricted to the Germ Line Micronucleus of the Ciliated Protozoan Tetrahymena thermophila. Eukaryot Cell. 2004;3:157–69.
20. Tay WT, Behere GT, Batterham P, Heckel DG. Generation of microsatellite repeat families by RTE retrotransposons in lepidopteran genomes. BMC Evol Biol. 2010;10:144.
21. Grandi FC, An W. Non-LTR retrotransposons and microsatellites. Mob Genet Elem. 2013;3, e25674.
22. Shao H, Tu Z. Expanding the Diversity of the IS630-Tc1-mariner Superfamily: Discovery of a Unique DD37E Transposon and Reclassification of the DD37D and DD39D Transposons. Genetics. 2001;159:1103–15.
23. Herrick G, Cartinhour S, Dawson D, Ang D, Sheets R, Lee A, et al. Mobile elements bounded by C4A4 telomeric repeats in oxytricha fallax. Cell. 1985;43:759–68.
24. Jaraczewski JW, Frels JS, Jahn CL. Developmentally regulated, low abundance Tec element transcripts in Euplotes crassus–implications for DNA elimination and transposition. Nucleic Acids Res. 1994;22:4535.
25. Doak TG, Witherspoon DJ, Jahn CL, Herrick G. Selection on the genes of Euplotes crassus Tec1 and Tec2 transposons: evolutionary appearance of a programmed frameshift in a Tec2 gene encoding a tyrosine family site-specific recombinase. Eukaryot Cell. 2003;2:95–102.
26. Brillet B, Bigot Y, Augé-Gouillou C. Assembly of the Tc1 and mariner transposition initiation complexes depends on the origins of their transposase DNA binding domains. Genetica. 2007;130:105–20.
27. Lhuillier-Akakpo M, Frapporti A, Denby Wilkes C, Matelot M, Vervoort M, Sperling L, et al. Local effect of enhancer of zeste-like reveals cooperation of epigenetic and cis-acting determinants for zygotic genome rearrangements. PLoS Genet. 2014;10, e1004665.
28. Macas J, Lambert GM, Dolezel D, Galbraith DW. Nuclear expressed sequence Tag (NEST) analysis: a novel means to study transcription through amplification of nuclear RNA. Cytometry. 1998;33:460–8.
29. Borges F, Gardner R, Lopes T, Calarco JP, Boavida LC, Slotkin RK, et al. FACS-based purification of Arabidopsis microspores, sperm cells and vegetative nuclei. Plant Methods. 2012;8:44.
30. Samadder P, Weng N, Doetschman T, Heimark RL, Galbraith DW. Flow cytometry and single nucleus sorting for Cre-based analysis of changes in transcriptional states. Cytometry A. 2016;89:430–42.
31. Zhang C, Barthelson RA, Lambert GM, Galbraith DW. Global characterization of cell-specific gene expression through fluorescence-activated sorting of nuclei. Plant Physiol. 2008;147:30–40.
32. Marion-Poll L, Montalban E, Munier A, Hervé D, Girault J-A. Fluorescence-activated sorting of fixed nuclei: a general method for studying nuclei from specific cell populations that preserves post-translational modifications. Eur J Neurosci. 2014;39:1234–44.

33. Haenni S, Ji Z, Hoque M, Rust N, Sharpe H, Eberhard R, et al. Analysis of C. elegans intestinal gene expression and polyadenylation by fluorescence-activated nuclei sorting and 3′-end-seq. Nucleic Acids Res. 2012;40:6304–18.
34. Bushman DM, Kaeser GE, Siddoway B, Westra JW, Rivera RR, Rehen SK, et al. Genomic mosaicism with increased amyloid precursor protein (APP) gene copy number in single neurons from sporadic Alzheimer's disease brains. elife. 2015;4.
35. Schoft VK, Chumak N, Bindics J, Slusarz L, Twell D, Köhler C, et al. SYBR Green-activated sorting of Arabidopsis pollen nuclei based on different DNA/RNA content. Plant Reprod. 2015;28:61–72.
36. Lake BB, Ai R, Kaeser GE, Salathia NS, Yung YC, Liu R, et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science. 2016;352:1586–90.
37. Okada S, Saiwai H, Kumamaru H, Kubota K, Harada A, Yamaguchi M, et al. Flow cytometric sorting of neuronal and glial nuclei from central nervous system tissue. J Cell Physiol. 2011;226:552–8.
38. Lacar B, Linker SB, Jaeger BN, Krishnaswami S, Barron J, Kelder M, et al. Nuclear RNA-seq of single neurons reveals molecular signatures of activation. Nat Commun. 2016;7:11022.
39. Krishnaswami SR, Grindberg RV, Novotny M, Venepally P, Lacar B, Bhutani K, et al. Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. Nat Protoc. 2016;11:499–524.
40. Wiedenheft B, Sternberg SH, Doudna JA. RNA-guided genetic silencing systems in bacteria and archaea. Nature. 2012;482:331–8.
41. Galvani A, Sperling L. RNA interference by feeding in Paramecium. Trends Genet. 2002;18:11–2.
42. Timmons L, Court DL, Fire A. Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in Caenorhabditis elegans. Gene. 2001;263:103–12.
43. Timmons L, Fire A. Specific interference by ingested dsRNA. Nature. 1998;395:854.
44. Creutz CE, Mohanty S, Defalco T, Kretsinger RH. Purine composition of the crystalline cytoplasmic inclusions of Paramecium tetraurelia. Protist. 2002; 153:39–45.
45. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
46. Liu Y, Schmidt B, Maskell DL. MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. Bioinformatics. 2010;26:1958–64.
47. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003;52:696–704.
48. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst. Biol. 2010;59:307–21.
49. Stöver BC, Müller KF. TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics. 2010;11:7.
50. Gouy M, Guindon S, Gascuel O. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Mol Biol Evol. 2010;27:221–4.
51. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–9.
52. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinforma Oxf Engl. 2009; 25:2078–9.
53. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
54. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinforma Oxf Engl. 2012;28:593–4.
55. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
56. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80.
57. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9, e1003118.
58. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013. http://www.repeatmasker.org/.
59. Dorai-Raj S. binom: Binomial Confidence Intervals For Several Parameterizations. R package version 1.0–5. 2009. http://CRAN.R-project.org/package=binom.

Troisième partie

Discussion et perspectives

Cette dernière partie est consacrée à une réflexion sur l'évolution de la génomique et plus particulièrement sur la génomique de la paramécie. Sans avoir l'ambition d'exposer une grande vision à long terme sur des questions biologiques fondamentales, je vais énoncer des perspectives de mon travail de thèse sur l'annotation des génomes de paramécies pour les cinq prochaines années. Pour ce faire, je m'appuierai sur des données encore non publiées. Les idées décrites sont, en partie, le fruit d'une réflexion collective du consortium *Paramécie* francilien et Lyonnais.

Aujourd'hui, re-séquencer un génome de paramécie à partir de quelques µg (voir ng) d'ADN n'est qu'une question d'un millier d'euros. Avec un séquençage *Illumina* de 100 millions de lectures pairées de 75 nt, l'ensemble du génome de 100 Mb est couvert à 150X (en théorie). En 2007, l'apparition des NGS a profondément changé notre approche de la génomique de la paramécie. Le premier génome MAC de paramécie a été publié en 2006, nous faisant entrer dans l'ère post-génomique (Aury et al. 2006). Basées sur l'annotation des gènes, les premières analyses à haut débit par puce à ADN sont apparues en 2010 (Gout et al. 2010, Arnaiz et al. 2010). Pendant 6 ans (entre 2006 et 2012), nous avons joué avec *seulement* quelques Go de données. Puis en 2012, le premier article sur la paramécie et traitant des données NGS sortait (Arnaiz et al. 2012). Depuis cette date, l'afflux de données n'a jamais cessé, d'autant que la communauté a recours de plus en plus souvent au séquençage. Ce déluge d'information a nécessité une refonte de notre architecture de travail. En effet, les besoins en ressource informatique, à la fois au niveau puissance de calcul et stockage, ont considérablement augmenté. Nous sommes passés de quelques Go à plus de 25To (en décembre 2019) de données de séquençage.

Assez logiquement dans la recherche, le chercheur souhaite accéder, et afficher, toutes les données disponibles afin d'avoir une vue la plus complète possible. Cependant, nous sommes vite noyés par cette avalanche d'information... Par ailleurs, toutes les données ne sont pas forcément disponibles au même endroit, et il est parfois difficile d'y accéder. Pour tenter de palier à cette difficulté, j'ai créé ParameciumDB (`https://paramecium.i2bc.paris-saclay.fr/` pour intégrer toutes les informations génomiques sur la paramécie et optimiser leur visualisation en fonction des besoins spécifiques des paraméciologues. Dans ce manuscrit, je n'ai pas beaucoup parlé de ParameciumDB, mais je trouvais plaisant d'en dire quelques mots car ce système d'information me tient à cœur et occupe toujours une partie de mon activité. Bien que d'une taille modeste, face aux bases de données comme EnsEMBL ou GenBank, ParameciumDB a trouvé sa place au sein de la communauté. Outil indispensable, elle rassemble des données qu'aucune autre base de données généraliste n'intègre. Sélectionnée comme faisant partie du nœud français ELIXIR 2019, ParameciumDB est reconnue et apporte une vraie valeur ajoutée à la recherche. Dès la naissance de ParameciumDB, nous avons adhéré à la philosophie des logiciels libres distribués par le projet GMOD (*Generic Model Organism Database* `http://gmod.org`). A notre échelle, nous favorisons l'accessibilité et le partage des données à l'image des *FAIR data* (pour *Findable, Accessible, Interoperable, Reusable* ; Facile à trouver, Accessible, Interopérable et Réutilisable `https://www.datafairport.org`).

Un autre changement majeur est à l'œuvre pour l'étude du génome de la paramécie. Pendant des années nous avons essoré le génome de l'espèce *Paramecium tetraurelia*. Un des messages du récent article, décrivant la nouvelle version de ParameciumDB (Arnaiz et al. 2019), est que l'étude du génome de la paramécie va maintenant s'appuyer de plus en plus sur une vision multi-espèce, mais également multi-souche d'une même espèce. Avec cette idée, un projet *France Génomique* de séquençage à grande échelle a financé le séquençage et l'assemblage de génomes de plusieurs espèces de paramécie, dont les génomes MIC de 7 espèces (voir **Tables VI.1** p. 160). Le but de ce projet était d'étudier en quoi le système de ciblage d'ADN à éliminer pendant les réarrangements programmés de génome de la paramécie par les petits ARN a eu un impact évolutif sur l'émergence de nouvelles espèces de paramécie (voir les **sections III.3.2**, **III.3.2.3** et **III.1.3** aux pages 68, 75 et 56). Au delà de cette question biologique et du point de vue de l'annotation, les génomes séquencés nous ont permis d'avoir une meilleure vision de la composition en gènes, en IES et en ET des *chromosomes* MIC de paramécies.

Chapitre VI

Les gènes et génomes MAC

Le génome somatique de la paramécie est débarrassé de ses séquences répétées ainsi que de ses IES. Contrairement à une procédure d'annotation de gènes classique où l'on commence par masquer le répétome, la paramécie se charge déjà, physiologiquement, de cette tâche dans le MAC. Cette situation favorable a d'ailleurs convaincu le Génoscope d'accepter le premier projet de séquençage, assemblage et annotation du génome MAC de *Paramecium tetraurelia* (Aury et al. 2006).

En concertation avec le laboratoire de M. Lynch (USA), nous nous sommes répartis les espèces à séquencer. Les génomes MAC de *P. biaurelia*, *P. sexaurelia*, *P. tetraurelia* et *P. caudatum* étant déjà assemblés, nous avons choisi 5 espèces pour lesquelles nous avions prévu de purifier les MIC (*P. octaurelia*, *P. pentaurelia*, *P. primaurelia*, *P. sonneborni* et un re-séquençage de notre souche favorite *P. tetraurelia*) ainsi que l'espèce *P. polycaryum*, en raison de son positionnement phylogénétique intéressant. De son côté, le laboratoire de Lynch a séquencé l'ADN MAC de 6 autres espèces du groupe *aurelia* : *P. decaurelia*, *P. dodecaurelia*, *P jenningsi*, *P. novaurelia*, *P. quadecaurelia*, *P. tredecaurelia* et une espèce appartenant à un groupe externe *P. multimicronucleatum* (voir **Figures III.5** p.57 et **VI.3** p.166). Disposant de la chaine de procédures pour annoter des génomes de paramécie (Arnaiz et al. 2017), je me suis chargé de l'annotation des gènes pour tous les génomes MAC disponibles (voir **Tables VI.1** p. 160).

Espèce	Souche	Compart.	Lab.	Complexité (Mb)	GC	N50 (kb)	Plus longue séq.	Gènes	Gènes non-codants	IES (par Mb)
P. bursaria	110224	MAC	Miao	29	28	96	266	19458	1169	
P. polycaryum	Hb20-6	MAC	CFG	46	40	524	1908	28712	6559	639 * (14)
		MAC-cons	CFG	25	44	567	1908	17962	1743	222 * (9)
P. multimicronucleatum	MO3c4	MAC	Lynch	36	25	442	141	17834	NA	
P. caudatum	43c3d	MAC	Lynch	30	28	314	794	18934	261	8762 (288)
P. biaurelia	V1-4	MAC	Lynch	75	25	159	1048	40823	562	45384 (606)
		MIC	CFG	110	24	689	1577			
P. novaurelia	TE	MAC	Lynch	65	23	79	625	35831	297	
P. primaurelia	AZ9-3	MAC	CFG	86	23	473	1053	43392	773	48479 (562)
		MAC-cons	CFG	74	23	484	1035	40231	471	43766 (595)
		MIC	CFG	102	23	499	1090			
	Ir4-2	MAC	Lynch	71	23	470	994	34474	NA	
P. pentaurelia	87	MAC	CFG	87	23	515	1263	42817	1141	48292 (553)
		MAC-cons	CFG	73	23	483	1263	38422	496	42686 (587)
		MIC	CFG	99	23	564	1187			
P. quadecaurelia	NiA	MAC	Lynch	59	23	224	744	34107	314	
P. tredecaurelia	209 AP38	MAC	Lynch	66	22	497	1224	36670	491	42275 (641)
P. decaurelia	223	MAC	Lynch	72	27	189	697	41808	998	
P. dodecaurelia	274	MAC	Lynch	72	27	176	653	41776	691	
P. octaurelia	138	MAC	CFG	86	28	493	997	46712	2314	50330 (582)
		MAC-cons	CFG	73	28	487	991	41985	1047	44509 (613)
		MIC	CFG	99	27	521	1016			
P. tetraurelia	32	MAC	CFG	81	28	491	1184	45320	1528	60198 (746)
		MAC-cons	CFG	72	28	495	1184	42238	864	
	51	MAC	Sanger	72	28	413	981	41533	1073	44928 (623)
		MAC	CFG	81	28	484	1029	46502	1872	49260 (611)
		MAC-cons	CFG	70	28	476	978	41997	832	44128 (630)
		MIC	CFG	102	27	571	1145			
	d4-2	MAC	Sanger	72	28	413	981	39642	NA	
P. sexaurelia	AZ8-4	MAC	Lynch	68	24	422	1100	36518	424	47002 (691)
		MIC	CFG	91	23	406	1055			
P. jenningsi	M	MAC	Lynch	65	23	213	808	37524	426	
P. sonneborni	30995	MAC	CFG	98	23	500	1142	50697	746	66952 (683)
		MAC-cons	CFG	83	23	511	1142	45689	474	60198 (729)
		MIC	CFG	217	21	34	332			

TABLE VI.1 – *Statistiques sur les génomes MIC et MAC, gènes et IES de paramécies*

Données non publiées ou tirées de AURY ET AL. (2006), ARNAIZ ET AL. (2012; 2017), MCGRATH ET AL. (2014b;a), GOUT ET AL. (2019), HE ET AL. (2019) . L'ordre des espèces est basé sur le positionnement phylogénétique de la **Figure III.5** (p.57).
* : nombre de *TA-indels* déduit d'un séquençage d'ADN MAC

VI.1 La contrepartie d'une grande profondeur de séquençage

Dans la **Table VI.1** (p. 160), il est remarquable que les espèces, dont les génomes MAC sont séquencés par le CFG (Consortium France Génomique), ont toujours une complexité supérieure aux génomes réalisés aux États-Unis. Par exemple, la souche AZ9-3 de *P. primaurelia* séquencée par le CFG a une complexité de génome MAC d'environ 86 Mb alors que la souche Ir4-2 de *P. primaurelia* séquencée par le laboratoire de M. Lynch est de 71 Mb. J'estime que la divergence entre ces deux souches est de l'ordre de 0.4% par des analyses de recherche de SNP. De manière analogue, le séquençage *Sanger* de la souche d4-2 de *P. tetraurelia* produit un génome de 72 Mb alors qu'un séquençage *Illumina* de la souche 51 conduit à un génome de 81 Mb. En sachant que le polymorphisme entre ces souches est minime (voir **section V.1.1** p. 105), une différence de 9 Mb (ou 15 Mb pour *P. primaurelia* correspondant à 17% de complexité supplémentaire) est très surprenante. La **Figure VI.1** (p.163) présente un *scaffold* MAC avec une représentation de sa densité en gènes non-codants (en orange) ainsi que les couvertures en lectures de séquençage d'ADN MAC (utilisées pour l'assemblage) et d'ARNm de cellules au stade végétatif (histogrammes violet et rouge respectivement sur la figure). Grâce aux lectures de séquençage d'ADN contenant, en partie, des répétitions télomériques (voir **section III.3.2.2** p. 74), il est possible de déduire la localisation des sites de télomérisation (en vert sur la figure). Les sites de télomérisation doivent correspondre, en théorie, aux extrémités des chromosomes MAC. Sur la figure, on constate qu'une portion du *scaffold* (à partir de 640 kb jusqu'à la fin) n'est que faiblement couverte par un séquençage d'ADN MAC, par rapport au reste du *scaffold*. De plus, de nombreux sites de télomérisation semblent être détectés dans cette région, ainsi qu'une densité en gènes non-codants anormalement élevée. Dans la **section III.3.2.2** (p. 74) nous avons vu que la paramécie réarrange son génome MAC de manière imprécise et surtout alternative. Ces régions faiblement couvertes dans un ADN MAC pourraient être dues à cette variabilité de réarrangement et pourraient correspondre à des séquences MIC présentes en faible nombre de copies dans le MAC des cellules. Autres possibilités, moins probables à mon avis, est que seulement certaines cellules au sein d'une population gardent ces régions dans le MAC, ou que cet ADN proviendrait en réalité du MIC.

Alors pourquoi les trouve-t-on plus dans nos assemblages *Illumina* ? La réponse viendrait simplement de la profondeur de séquençage. En effet, le CFG a séquencé beaucoup plus profondément que le laboratoire de Lynch, et cette couverture a été suffisamment importante pour assembler ces régions et surtout les lier aux chromosomes MAC. En réalité, certains segments de ces régions faiblement couvertes sont présents dans les assemblages *Sanger* ou ceux de Lynch. Cependant, elles sont souvent isolés dans de petits *scaffolds* et représentent un consensus de plusieurs copies. Dans le premier assemblage de *P. tetraurelia* nous avons toujours considéré que l'ensemble de la complexité du génome MAC était rassemblé dans les 188 plus grands *scaffolds*. Toutefois, j'ai toujours eu des scrupules à enlever ces petits *scaffolds* car certains d'entre eux contiennent des gènes fonctionnels (par

exemple CenH3a est sur le *scaffold* 466 (Lhuillier-Akakpo et al. 2016)). De manière générale, je pense que l'ensemble de ces régions faiblement couvertes dans les assemblages n'existent pas réellement dans le génome, et sont très probablement chimériques ou tout au moins mal assemblées.

Pour ces nouveaux génomes, et en s'appuyant sur des représentations comme la **Figure VI.1** (p.163), il était tentant de délimiter ces régions. A l'aide d'une procédure de prédiction automatique complétée avec un ajustement manuel, j'ai défini deux catégories de séquences : le "MAC constitutif", les régions présentes de manière homogène dans les MAC contenant les gènes codants, et le "MAC alternatif" défini par une faible couverture en séquençage d'ADN MAC et une haute densité en gènes non-codants peu ou pas exprimés (arc bleu extérieur sur la figure). La **Table VI.1** (p. 160) indique les complexités des génomes MAC constitutifs. Avec des complexités comparables, tous les résultats sont maintenant plus cohérents. La **Figure VI.2** (p.164) montre une bien meilleure corrélation entre la taille des génomes MAC et le nombre de gènes ou le nombre d'IES. Sans plus m'étendre sur le sujet, je souhaite noter la particularité de l'espèce *Paramecium polycaryum* avec un taux de G+C d'environ 40%, très loin des 23 à 28% habituels des autres paramécies ou des espèces plus distantes comme *Tetrahymena* ($\sim$22%) ou *Oxytricha* ($\sim$31%). Il serait intéressant de creuser plus avant les raisons de cette originalité.

Il est vraiment amusant de constater qu'une grande profondeur de séquençage, tant désirée, apporte une dose de variabilité inattendue. Avec nos séquençages de populations de cellules contenant des MAC alternativement réarrangés, nous sommes probablement confrontés à la même problématique de variabilité, retrouvée lors du séquençage de cellules uniques (Eberwine et al. 2014).

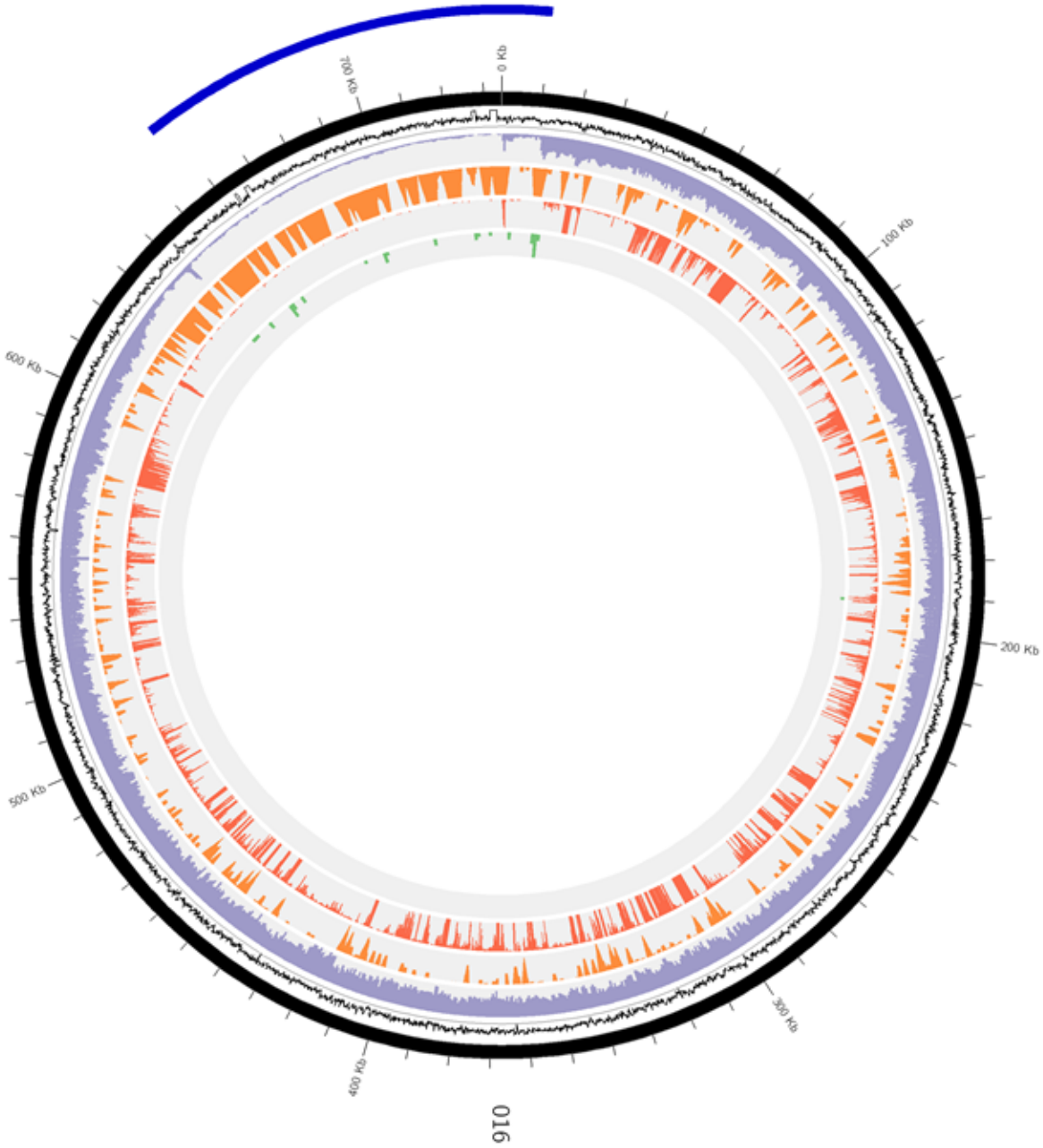


FIGURE VI.1 – *Représentation circulaire du scaffold 016*

Représentation circos du *scaffold* 016 MAC de *P. tetraurelia*. En violet, l'histogramme de couverture en lectures Illumina de ADN MAC utilisées pour l'assemblage. En orange, la densité en gènes non-codants. En rouge, la couverture en lectures ARN-seq d'un échantillon de cellules au stade végétatif. En vert, la densité en sites de télomérisation détectés par les lectures de séquençage contenant en partie des répétitions télomériques. L'arc bleu extérieur représente la portion de séquence masquée, et donc n'appartenant pas à ce qu'on appelle le MAC constitutif.

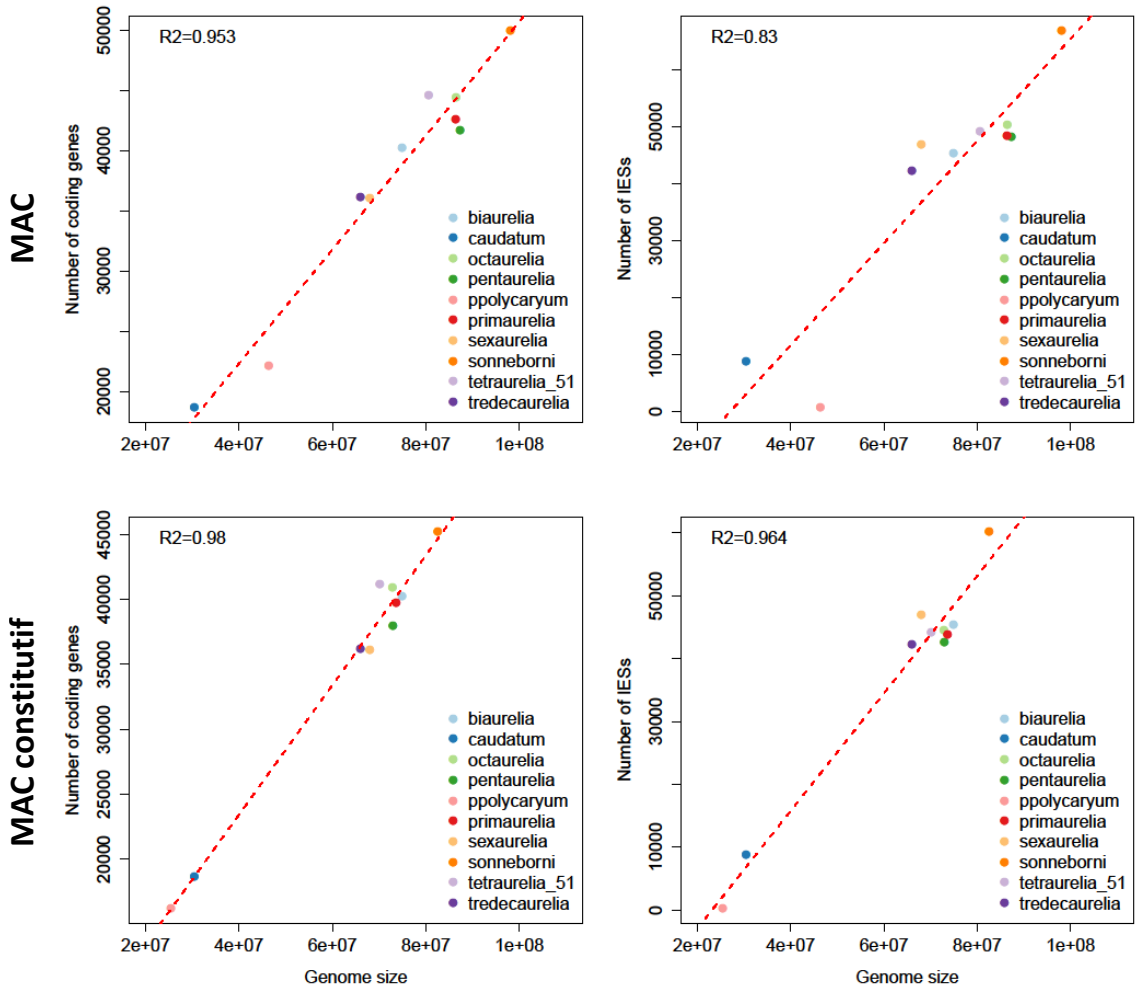


FIGURE VI.2 – *Corrélation entre taille de génome et nombre de gènes ou d'IES*

Un point représente le génome d'une espèce de paramécie. L'abscisse donne la complexité du génome et l'ordonnée le nombre de gènes codants (colonne de gauche) ou le nombre d'IES (colonne de droite). Les génomes MAC assemblés sont sur la première ligne et les génomes MAC constitutifs sur la deuxième (voir texte principal).

VI.2 Positionnement phylogénétique des WGD

Dans les **sections III.3.4** et **IV** (p. 85 et p. 89), j'ai exposé pourquoi la paramécie est un bon modèle pour étudier l'évolution et la dynamique des génomes après un événement aussi dramatique qu'une WGD. En effet, le génome de *Paramecium tetraurelia* montre les stigmates d'au moins trois WGD (récente, intermédiaire et vieille), dont la plus récente est encore très visible au niveau nucléotidique (Aury et al. 2006). Une WGD entraine généralement une vague de perte massive de gènes. Cette perte ne se fait pas au hasard, elle est soumise à différentes pressions sélectives. Le pré-article de Gout et al. (2019) utilise les nouveaux génomes de paramécie disponibles afin de documenter la dynamique de perte de gènes après une WGD. Cette étude propose notamment que la perte de gènes après une WGD serait plus lente chez les paramécies que chez d'autres espèces, suggérant une pression sélective plus forte.

McGrath et al. (2014b) ont montré que les deux dernières WGD (WGD récente et intermédiaire) subies par *Paramecium bi- sex- et tetr- aurelia* ont eu lieu après la divergence entre *P. caudatum* et le groupe *aurelia*. Dans Gout et al. (2019), nous confirmons que toutes les *aurelia* partagent effectivement ces deux WGD, et qu'elles ont eu lieu après la divergence de *P. caudatum* et *P. multimicronucleatum* (voir **Figures VI.3** p.166). Malheureusement, les places phylogénétiques des nouveaux génomes de paramécie ne permettent pas d'affiner plus précisément la datation de ces deux WGD. Le positionnement phylogénétique de la WGD qualifiée de vieille reste largement incertain. Ruehle et al. (2016) ne voient aucune évidence d'une WGD chez *Tetrahymena thermophila*, et McGrath et al. (2014a) ont montré que cette WGD s'est produite avant la divergence de *P. caudatum* (et *P. multimicronucleatum* par Gout et al. (2019)). Le séquençage de génome d'espèces avec une place phylogénétique entre *Tetrahymena* et *P. caudatum* ainsi qu'une analyse dédiée des génomes de *P. polycaryum* et *P. bursaria* pourraient nous apporter des précisions sur la datation de la "vieille" WGD (voir **Figures VI.3** p.166).

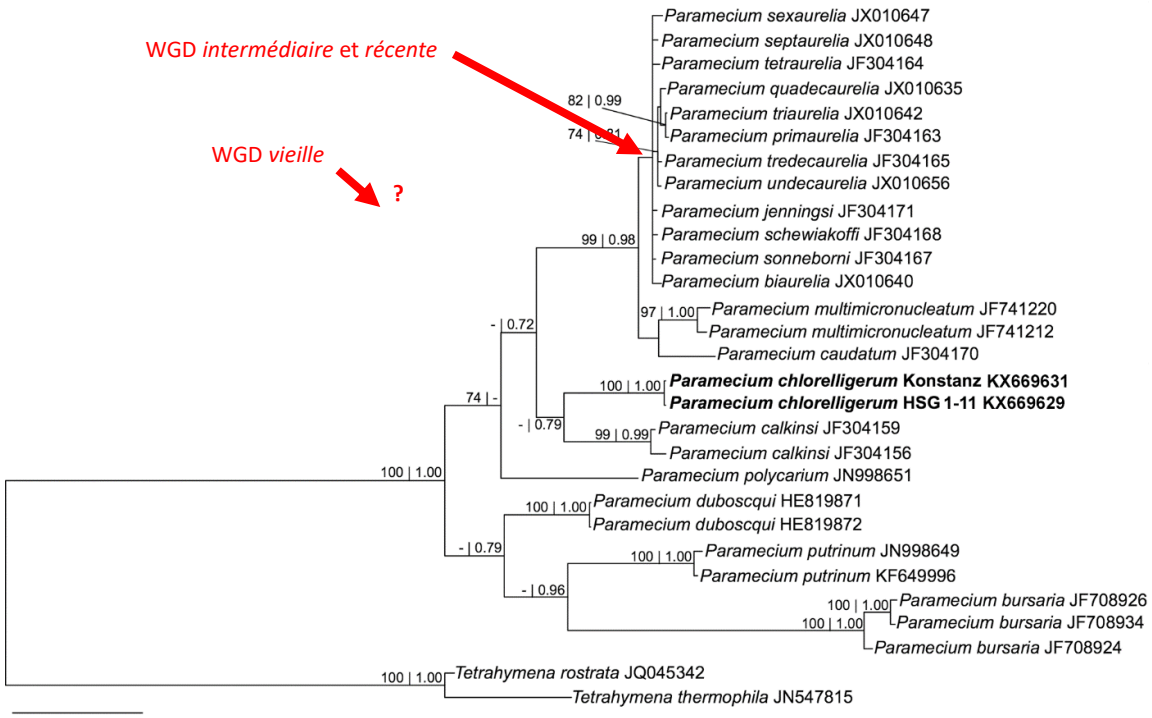


FIGURE VI.3 – *Positionnement phylogénétique des WGD*

Phylogénie des paramécies, inférée à partir des séquences ITS1-5.8S-ITS2. Les valeurs de bootstrap (1000 pseudo-replicats) et les probabilités bayésienne sont indiquées sur les nœuds. Les flèches rouges indiquent le positionnement phylogénétique des WGD dans l'histoire évolutive des paramécies. La WGD qualifiée de vieille n'est pas encore positionnée précisément. Figure modifiée de LANZONI ET AL. (2016).

VI.3 Les pseudogènes et gènes non-codants

Les génomes de paramécies sont très denses en gènes. La taille extrêmement basse des introns conduit à un génome codant à ∼80%. De plus, la détection d'ORF est facilitée par le code génétique des ciliés, n'utilisant qu'un seul codon terminateur de traduction (voir **section I.2.2.2** p. 12). Entrainé à ce genre de caractéristiques génomiques, le prédicteur de gènes EuGene a du mal à *supporter* une grande portion de séquence sans gène (Arnaiz et al. 2017). Dans le cas où l'ORF n'est pas évidente, de faibles signaux intrinsèques (voir **section II.1.2** p. 34) vont conduire à l'annotation de nombreux gènes non-codants. Au delà des classiques ARNt, ARN ribosomaux, snoARN, snARN et SRP ARN (voir **section I.2.2.2** (p. 12) Kumazaki et al. (1982), Chen et al. (2009)), seuls quelques exemples de gènes non-codants sont connus chez la paramécie (ARN pour la télomérisation *de novo* (McCormick-Graham and Romero 1996) ; MS2a et b, ARN spécifiques de l'autogamie (Tanabe 2006)). Plusieurs centaines de gènes non-codants putatifs ont été annotés dans chacune des espèces (voir **Table VI.1** p160). Une conservation inter-espèce pourrait être un bon indicateur de la réelle existence du gène et d'un potentiel rôle fonctionnel.

La procédure d'annotation des gènes n'intègre pas l'existence de pseudogènes (voir **section I.2.2.2** p. 15). Un *locus* portant un pseudogène dont la séquence codante est abimée va avoir tendance à être annoté en plusieurs petits gènes codants ou non. Une partie des gènes codants annotés sont donc probablement des pseudogènes plus ou moins détériorés. Il y a plusieurs années, dans le cadre d'un projet de ré-annotation communautaire du génome de *P. tetraurelia* avec Apollo (voir **section II.1.5** p. 40), l'alignement des paralogues de WGD permettait tout d'abord de détecter les erreurs d'assemblage et donc d'annotation, mais décelait également le ou les membres de la famille multi-génique en cours de pseudogénisation. Avec la même idée, il serait intéressant d'utiliser les autres génomes de paramécies pour annoter les pseudogènes et ainsi corriger et enrichir l'annotation. Pour réaliser cette analyse, des liens d'orthologie sont nécessaires. Même si pour cette étude des pseudogènes le problème ne se pose pas forcément, le calcul de l'orthologie entre les espèces du groupe *aurelia* est assez difficile car les WGD compliquent la tâche. En effet, pour certains couples d'espèces (*P. tetraurelia* et *P. sexaurelia*), la divergence entre paralogues WGD est comparable à la divergence entre orthologues. Donc la possibilité d'établir un mauvais lien d'orthologie est grande. Toujours est-il ce travail doit être réalisé le plus proprement possible pour profiter au maximum de cette nouvelle ressource génomique.

Chapitre VII

Les génomes MIC

Le génome MAC commence à être bien caractérisé. En revanche, la structure et la composition du génome MIC, de par son accessibilité difficile, restent encore peu connus. Pour caractériser ses éléments spécifiques, nous comparons (*bioinformatiquement*), à la manière de la paramécie, des séquences germinales non réarrangées au génome somatique réarrangé.

VII.1 Les chromosomes MIC

L'article de Guérin et al. (2017) a posé le principe de la méthode de purification de noyaux MIC végétatifs. L'échantillon contenait suffisamment de matériel d'ADN purifié pour être séquencé et assemblé. Bien que fragmenté, l'assemblage a apporté des éléments intéressants comme la taille du génome MIC de *P. tetraurelia* (∼100 Mb) ou la découverte de nouvelles familles d'ET. Il a également révélé que certaines séquences du génome MIC éliminées pendant les réarrangements étaient absentes dans un séquençage d'ébauches en développement de cellules déplétées en *PGM* (ADN PGM) (voir **section V.2** des résultats p. 135). Cet essai concluant a prouvé l'utilité et le bénéfice d'avoir un échantillon d'ADN micronucléaire afin d'accéder au génome MIC dans son ensemble. Frédéric Guérin, du CFG, a purifié les noyaux MIC de 7 nouvelles espèces de paramécie pour le séquençage (voir **Table VI.1** p. 160).

VII.1.1 Assemblage des génomes MIC

Pour réaliser le meilleur assemblage possible, le CFG a adopté la stratégie de séquençage suivante : (1) Séquençage très profond des échantillons d'ADN MIC en *Illumina* avec des lectures pairées de 250 nt de fragments de 400pb (2) Annotation des IES grâce au logiciel ParTIES (Denby Wilkes et al. 2016) (3) afin d'aider l'assembleur, élimination de l'ambiguïté apportée par les lectures provenant de l'ancien MAC au niveau des IES (voir **section V.1.1** des résultats p. 105). Au fur et à mesure, F. Guérin a amélioré la pureté des échantillons et ce nettoyage n'a eu possiblement que peu d'impact sur les derniers

échantillons. (4) Fusion des lectures chevauchantes pour former des séquences de 400 pb de haute qualité. (5) Assemblage des lectures MIC à l'aide d'un assembleur OLC (voir **section I.3.3** p. 26) (6) Ne pouvant obtenir des quantités d'ADN MIC suffisantes pour des banques de fragments longues distances, nous avons opté pour un séquençage *mate-pair Illumina* d'un ADN PGM (au moins 3 tailles de fragments). Nous étions conscients des limites de ces échantillons d'ADN PGM mais, au moment du séquençage, c'était la meilleure option.

Toute cette procédure est assez complexe et assez coûteuse. Dans l'avenir, il est probable que nous essaierons d'avoir recours à des stratégies plus simples et à l'utilisation de séquenceurs de troisième génération (voir **section I.3.2.3** p. 23). Malgré le taux d'erreur, les lectures longues pourront résoudre un certain nombre de problèmes que nous allons aborder dans le paragraphe suivant. Toutefois, il est vrai que la faible quantité de matériel génétique purifiable reste un problème majeur à résoudre. Aujourd'hui, ces technologies requièrent des concentrations d'ADN MIC que nous ne sommes pas encore en mesure de fournir (quelques µg). Pourtant des données préliminaires montrent des résultats encourageants d'assemblage *de novo* de lectures longues ONT sur des cellules déplétées en Pgm ou en Ezl1. Toutes ces méthodes sont en constante évolution. A l'image des projets de séquençage de cellules uniques, le rêve serait de mettre un MIC dans un tube et de séquencer l'ADN qu'il contient...

VII.1.2 Quelle est la structure des chromosomes ?

La **Table VI.1** (p. 160) donne des statistiques sur 7 assemblages MIC obtenus (*P. biaurelia*, *P. octaurelia*, *P. pentaurelia*, *P. primaurelia*, *P. sexaurelia*, *P. tetraurelia* et pour *P. sonneborni* nous ne disposons que de *contigs*). Malheureusement, pour les espèces *P. caudatum* et *P. tredecaurelia* nous n'avons pas pu obtenir de *scaffolds*. Les séquençages des échantillons MIC ont permis, néanmoins, d'annoter les IES de ces génomes (voir **section VII.2.2.1** suivante p. 174).

Tous les génomes MIC d'espèces du groupe *aurelia* semblent avoir une complexité d'environ 100 Mb (voir **Table VI.1** p. 160). Il faut noter néanmoins la particularité du génome de *P. sonneborni* dont les ∼200 Mb seraient liées, d'après des analyses préliminaires (de L. Duret, E. Meyer et L. Sperling), à des phénomènes d'introgression de matériel génétique provenant d'espèces du groupe *aurelia*. En revanche, tous les *scaffolds*, de l'ensemble des génomes MIC, n'excèdent pas une taille maximum de ∼1.1 Mb, une longueur comparable aux *scaffolds* MAC. Même si les statistiques globales de qualité d'assemblage sont loin d'être honteuses, nous avons été déçu par ces résultats. Nous espérions assembler les chromosomes MIC. Comme discuté par Duret et al. (2008), chaque *scaffold* MIC, borné par des séquences télomériques, aurait rejoint deux (ou plus) chromosomes MAC. Or nous observons que les *scaffolds* MIC sont plus ou moins comparables aux *scaffolds* MAC. Il est vrai que les *scaffolds* MIC portent plusieurs milliers de Kb supplémentaires, notamment aux extrémités. Dans la **section VII.2** suivante (p. 172), nous verrons ce que

contiennent ces séquences MIC. Cependant, la question de la structure des chromosomes MIC demeure. Est ce que les chromosomes MIC sont, en effet, équivalents aux chromosomes MAC ? Ou bien est ce que malgré nos efforts, nous nous heurtons toujours aux mêmes problèmes d'assemblage au niveau probablement de répétitions complexes.

Seul marqueur de la fin des chromosomes, nous n'avons aucune certitude sur ce qu'est un télomère MIC (voir **section I.2.1** p. 7). Ils pourraient être équivalents aux télomères MAC (voir **sections III.3.2.2** p. 74), ou composés de répétitions complètement différentes. Pourtant des évidences préliminaires semblent nous apporter quelques éléments de réponse. Des lectures longues ONT portant des répétitions télomériques (semblables à celles des MAC), alignées sur l'assemblage MIC et l'utilisation d'assemblages *de novo* de lectures ONT suggéreraient que les *scaffolds* MIC s'interrompraient juste avant les extrémités des chromosomes. Des cartes de recombinaison réalisées par Laurent Duret iraient également dans ce sens. Je garde bon espoir que nous parviendrons prochainement à estimer le nombre de chromosomes MIC de la paramécie.

Autre élément structurant les chromosomes : les centromères. Sur ce sujet, le mystère est encore plus grand que pour les télomères. Nous n'avons aucune idée de ce qu'est un centromère de paramécie (voir **sections I.2.1** p. 7 et **III.3.1** p. 64). Les chromosomes MIC, sont ils monocentriques ou holocentriques ? Si les centromères sont composés de séquences de basse complexité, il est possible que ces séquences ne soient tout simplement pas dans l'assemblage et que les scaffolds s'interrompent à ces *loci*. Lhuillier-Akakpo et al. (2016) ont démontré que le marquage du variant de l'histone H3 centromérique (CenH3) disparaissait au cours du développement macronucléaire. L'inactivation des gènes *PGM* ou *EZL1* (voir **section III.3.2.3** p. 75) empêche la disparition du marquage CenH3, suggérant que la perte des séquences centromériques est causée par l'élimination d'ADN pendant les réarrangements programmés. Les séquences centromériques ont donc un comportement similaire aux IES. Il est donc envisageable que toutes ou une partie des IES jouent le rôle de centromères. Les procédures de ChIP-seq sur des modifications d'histones ont été mises au point chez la paramécie (Frapporti et al. 2019). Il serait intéressant de tenter de sélectionner les séquences centromériques associées à CenH3. L'exploitation des données pourrait être assez acrobatique en raison de la nature vraisemblablement répétée des centromères. Encore une fois, l'apport d'un séquençage de lectures longues pourrait lever un certain nombre d'ambiguïtés.

VII.2 Que contiennent les génomes MIC ?

VII.2.1 Toutes les paramécies ont elles le même contenu en ET ?

À l'aide d'approches par homologie, le premier assemblage MIC de *P. tetraurelia* a permis de caractériser les premiers retrotransposons (RT) de paramécie (Guérin et al. 2017) (voir **section V.2** des résultats p. 135). Ces consensus de LINE viennent s'ajouter aux quelques ET à ADN de la super-famille des TIR (Arnaiz et al. 2012) (voir **section V.1.1** des résultats p. 105). Uniquement 10% du génome MIC éliminé pendant la maturation du MAC de *P. tetraurelia* est annoté comme dérivé d'ET.

Afin d'avoir une vision plus exhaustive du paysage des ET de paramécie, le CFG a engagé un grand projet d'annotation des ET par des approches *de novo* (voir **section II.2.2.1** p. 48). Le logiciel d'annotation REPET (Flutre et al. 2011) a été utilisé sur les génomes MIC (voir **Table VI.1** p. 160). La **Figure VII.1** (p.173) présente les résultats préliminaires de cette analyse REPET. Les ET annotés sont majoritairement des LINE (RIX) et des TIR (DTX). Déjà détecté chez *Oxytricha* et *Tetrahymena* , nous observons, pour la première fois chez la paramécie, des évidences de la présence d'ET de la super-famille des Hélitrons (DHX) (voir **section II.2** p. 42) (Chen et al. 2014, Hamilton et al. 2016). Comme pour *Tetrahymena*, les ET de la super-famille LTR semblent ne pas avoir colonisé les génomes de paramécies (Hamilton et al. 2016). Pourtant, avec cette image plus complète des familles d'ET de paramécie, les copies n'occupent toujours que 12% à 25% du génome MIC réarrangé (entre 3 et ∼8 Mb ; voir **Figure VII.1** p.173). Les proportions de complexité occupée par chaque classe d'ET semblent assez similaires au sein des espèces de paramécie. Même si nous ne sommes pas encore certains de la raison, on peut néanmoins noter que les Hélitrons paraissent occuper une plus grande proportion du génome de *P. octaurelia*. Une explication pourrait venir de l'état de dégradation des copies, variable selon les espèces, et donc de leur capacité a être assemblées. En effet, toutes ces méthodes d'annotation dépendent de la qualité de l'assemblage. Il est vrai que l'ensemble des copies d'ET de paramécie publiées jusqu'à présent semblent être fortement dégradées dans le génome MIC, favorisant l'assemblage mais pénalisant la sensibilité de détection. Pourtant, dans le cas de copies très récentes (ou peu altérées), l'assemblage de ces régions sera confronté au problème de *collapse*, entrainant une sous détection de l'ET (car artéfactuellement non répétées) et/ou une sous estimation de l'abondance de l'ET dans le génome. Quoi qu'il en soit, il reste encore du travail pour décrypter le contenu de ces génomes.

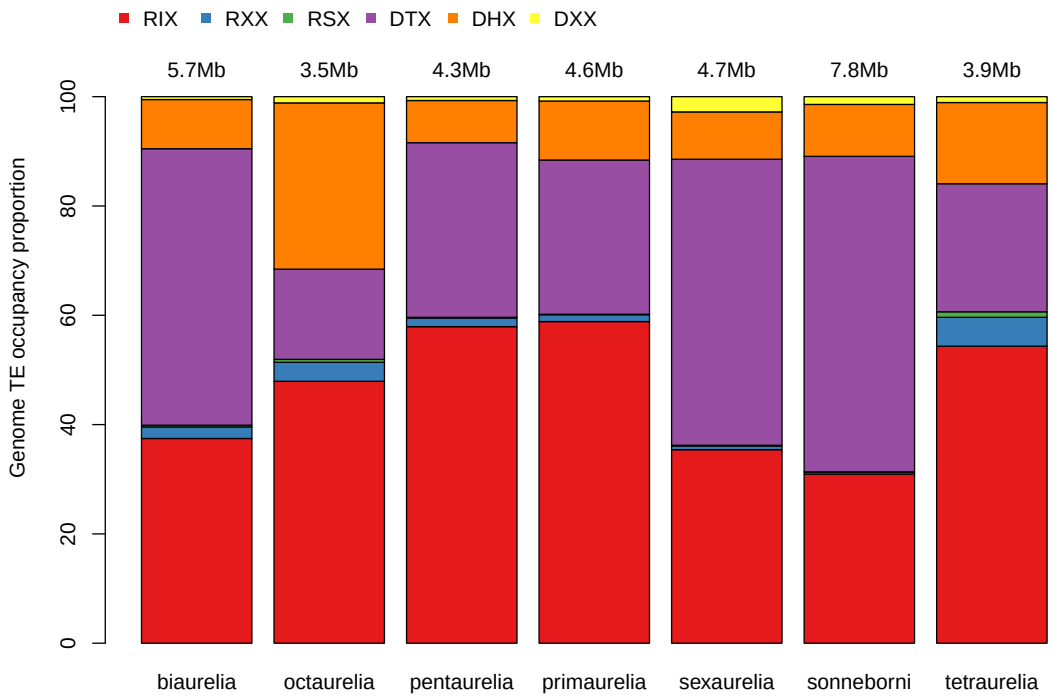


FIGURE VII.1 – *Complexité de génome occupée par les ET*

Proportion de la complexité génome occupée par chaque classe d'ET chez *P. biaurelia*, *P.octaurelia*, *P. pentaurelia*, *P. primaurelia*, *P. sexaurelia*, *P. sonneborni* et *P. tetraurelia*. Les couleurs correspondent aux classes d'ET définies par WICKER ET AL. (2007) (voir **section II.2** p. 42). Les ET de classe 1 sont en rouge (les éléments LINE), bleu (rétroélément dont l'ordre n'a pas été clairement identifié) et vert (les éléments SINE). Les ET de classe 2 sont en violet (les éléments TIR), orange (les éléments Helitron) et jaune (transposon à ADN dont l'ordre n'a pas été clairement identifié). La complexité de génome occupée par l'ensemble des copies d'ET est indiquée au dessus de chacune des barres. Ces résultats ont été tirés de données d'analyses REPET non publiées et donc préliminaires.

VII.2.2 Les IES

VII.2.2.1 Toutes les paramécies ont elles des IES ?

En 2012, les 45 000 IES du génome de *P. tetraurelia* ont été cataloguées (Arnaiz et al. 2012). Avec C. Denby Wilkes, nous avons standardisé les procédures d'identification des IES et créé la suite logicielle ParTIES (Denby Wilkes et al. 2016). Avec un séquençage adapté et un génome MAC de référence, annoter les IES d'un génome de paramécie n'est, *a priori*, plus un problème. En suivant la stratégie de purification de noyaux MIC (Guérin et al. 2017), des échantillons d'ADN MIC ont été séquencés permettant l'annotation des IES pour 8 nouvelles espèces de paramécie (*P. biaurelia*, *P. caudatum*, *P.octaurelia*, *P. pentaurelia*, *P. primaurelia*, *P. sexaurelia*, *P. sonneborni* et *P. tredecaurelia*). La **Table VI.1** (p. 160) donne le nombre d'IES détectées.

Contrairement à *Tetrahymena* qui ne compte que 12 IES excisées précisément dans son génome (Hamilton et al. 2016, Cheng et al. 2016, Feng et al. 2017), tous les génomes de paramécies, jusqu'à présent, semblent contenir un grand nombre d'IES, avec des caractéristiques similaires aux IES de *P. tetraurelia*. Ce résultat n'est pas surprenant, car tous les facteurs de la machinerie d'excision des IES sont relativement bien conservés dans ces espèces (Bischerour et al. 2018). La **Figure VI.2** (p.164) montre une bonne corrélation entre la complexité du génome MAC (constitutif) et le nombre d'IES. Ne disposant pas d'ADN MIC pour l'espèce *P. polycaryum*, la **Table VI.1** (p. 160) présente les résultats de la méthode MILORD (voir **section V.1.2** des résultats p. 127) de détection de *TA-indels* (voir **section III.3.2.1** p. 71) à partir d'un séquençage d'ADN MAC. Cependant, cette sous estimation de 222 IES, ou plus exactement *TA-Indels*, ne doit pas être loin de la vérité si l'on en croit la bonne corrélation entre nombre d'IES et taille de génome MAC (voir **Figure VI.2** p.164). Le séquençage d'un ADN MIC (ou un ADN PGM) de *P. polycaryum* pourrait lever cette incertitude. Plus généralement, l'origine évolutive des IES excisées précisément au sein du phylum des paramécies est une question en suspens. Et la recherche d'IES dans des paramécies avec une place phylogénétique à mi chemin entre *Tetrahymena* et *P. caudatum* (*P. bursaria*, *P. duboscqui*, *P. germanicum* ou *P. goertzi* FT8) pourrait nous apporter des éléments de réponse (voir **Figures VI.3** p.166).

Dans Arnaiz et al. (2012), nous avons documenté la dynamique d'apparition/disparition des IES au cours de l'histoire évolutive de *P. tetraurelia*, en nous servant des marqueurs temporels que sont les WGD. A l'aide des données sur les IES de toutes ces espèces de paramécies, nous pouvons maintenant préciser ces informations. Un article en préparation (Sellis et al.) présente l'étude de l'histoire évolutive des IES et la dynamique de perte et de gain d'IES au sein de la clade *Paramecium*. L'étude confirme, en effet, la plupart des hypothèses avancées dans Arnaiz et al. (2012) : la majorité des IES se sont insérées avant la divergence des espèces du groupe *aurelia* et le taux de perte est globalement assez uniforme durant l'histoire évolutive de la paramécie. Pourtant, l'espèce *P. sonneborni* sort du lot. Elle semble avoir subi une vague d'insertions d'IES plus récente. De plus, son génome contient des IES en plusieurs milliers de copies identiques, également

trouvées dans l'espèce *P. tredecaurelia*. Cette observation pourrait être la première évidence de l'existence d'IES "mobile", non-autonome.

VII.2.2.2 Avons nous catalogué toutes les IES ?

Le module MICA de ParTIES annote les IES par comparaison entre une séquence de référence sans IES et un séquençage ou assemblage contenant les insertions. En pratique, un séquençage d'ADN MIC (ou d'ADN de cellules déplétées pour un facteur impliqué dans les réarrangements ; voir **section III.3.2.3** p. 75) est comparé au génome MAC de référence. Les IES annotées sont donc, par définition, localisées dans les régions génomiques MIC colinéaires au génome MAC. Or au moins 25 Mb de grandes régions génomiques MIC sont éliminées du génome MAC. Par ailleurs, d'abord mis en évidence par Duharcourt et al. (1998), Bétermier et al. (2000) ont caractérisé l'excision d'une petite IES, insérée dans une grande IES, transitoirement détectable pendant les processus sexuels. Il est donc parfaitement envisageable que des segments soient excisés transitoirement et précisément entre deux TA (comme des IES) au milieu des grandes régions génomiques MIC qui seront finalement éliminées du génome MAC.

La déplétion des facteurs indiqués dans la **Table III.2** (p. 80) induit la rétention de certaines catégories d'IES. Par exemple, la dépletion de Dcl2/3 provoque la rétention significative de seulement 6% des IES (N=3000) (Sandoval et al. 2014, Lhuillier-Akakpo et al. 2014) (voir **section III.3.2.3** p. 75). En revanche, toutes les régions imprécisément éliminées du MIC sont couvertes par un séquençage d'un ADN DCL2/3. Autre exemple, dans un contexte où le gène *EZL1* est inactivé, 30% des IES sont parfaitement excisées alors que l'ensemble du génome MIC est séquencé (Lhuillier-Akakpo et al. 2014, Guérin et al. 2017). Avec une analyse préliminaire MILORD (voir **section V.1.2** des résultats p. 127) de détection d'événements d'excision cryptique (voir **section III.3.2.1** p. 71) dans le génome MIC à l'aide des lectures de séquençage d'un ADN DCL2/3, on s'aperçoit que certains *TA-indels* sont très fréquents. La **Figure VII.2** (p.177) montre une région MIC éliminée dans le MAC, où l'on voit deux segments excisés (entre deux TA) dans un contexte DCL2/3 inactivé. Cet exemple indique l'existence d'*IES* dans une région MIC spécifique. Même, si cela pose la question de l'existence réelle de ces nouvelles *IES* dans un contexte sauvage, une analyse dédiée doit être menée pour toutes les annoter.

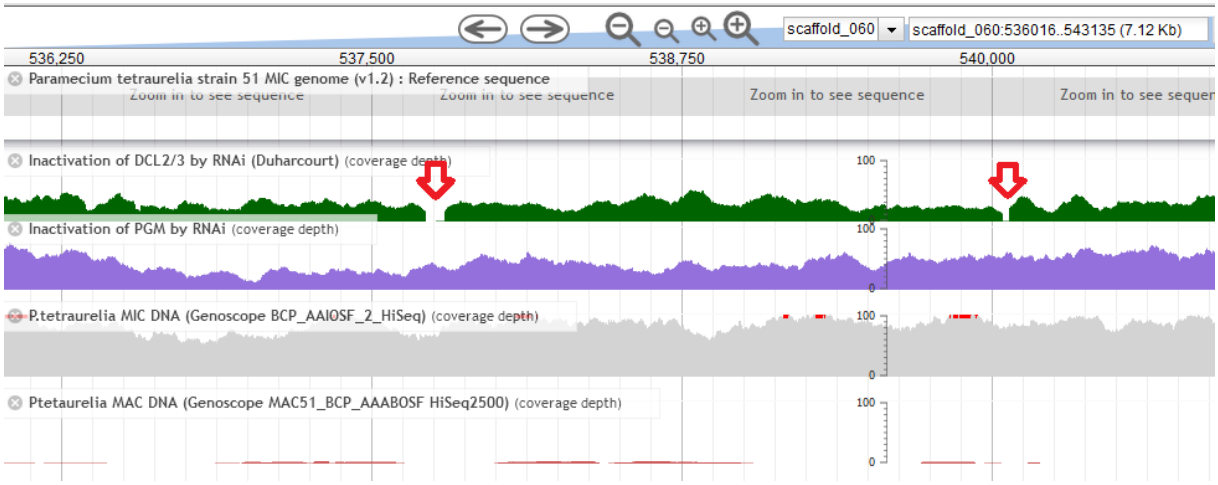


FIGURE VII.2 – *IES dans une région MIC spécifique*

Région génomique éliminée pendant les réarrangements, mis en en évidence par l'absence de couverture en lectures de séquençage MAC (piste rouge en bas). En gris la couverture de séquençage de l'échantillon d'ADN MIC utilisé pour l'assemblage du génome MIC. La couverture en lectures de séquençage de l'ADN d'ébauches en développement de cellules déplétées pour Pgm (en violet) ou pour Dcl2/3 (en vert). L'inactivation de *DCL2/3* entraine la rétention de la majorité des séquences éliminées imprécisement mais ne retient que quelques milliers d'IES sur les 45000. Les flèches rouges montrent deux *TA-indels* visiblement excisés dans un contexte déplété en Dcl2/3.

VII.2.2.3 Un rôle fonctionnel pour les IES ?

Plus de 90% des sites d'insertion d'IES sont conservés au sein des espèces du groupe *aurelia* (Sellis et al. en préparation). Grâce au système d'élimination des IES de la paramécie, il est manifestement difficile ou inutile de se défaire d'une IES dans le génome MIC (Arnaiz et al. 2012). En revanche, en comparant les IES paralogues de la WGD la plus récente, nous observons que leurs séquences divergent rapidement, traduisant une absence de pression de sélection. Dans Sellis et al. (en préparation) sont rapportés des résultats équivalents en comparant les IES orthologues. Pourtant, dans cette même étude, nous identifions ∼70 familles d'IES très conservées au sein des espèces du groupe *aurelia*. Ces IES conservées peuvent se localiser dans les séquences promotrices, les UTR ou les exons des gènes. Ces rares cas de contrainte de sélection forte sur les séquences laissent suspecter un rôle fonctionnel de ces IES. Le seul exemple publié d'IES avec un rôle fonctionnel, est l'IES du gène mtA (voir **section III.2.3.1** p. 61). L'idée d'un rôle fonctionnel pour ces IES conservées est d'autant plus séduisante que certaines IES sont transcrites (basé sur des données mARN-seq) transitoirement pendant l'autogamie.

Comme nous l'avons vu précédemment, pendant les processus sexuels, la maturation du nouveau MAC nécessite notamment l'excision de 45 000 IES pouvant interrompre les ORF des gènes. Avant que les ébauches en développement soient pleinement matures, les fragments de l'ancien MAC assurent la majorité de la transcription génique (80%) (Berger 1973). Progressivement les ébauches prennent le relais transcriptionnel des fragments. Donc dans une fenêtre de temps précoce et relativement courte, il est possible que des copies de gènes contenant encore leurs IES soient transcrites. Dans l'écrasante majorité des cas, ces transcrits aberrants seront non fonctionnels et donc dégradés (voir **section III.3.3** p. 82). Pourtant, l'idée que des IES pourraient être impliquées dans le développement du MAC, a été émise il y a quelques années. En effet, une IES, qui est par nature transitoirement présente dans le génome, pourrait être un élément régulateur parfait. Sa présence ou non modifierait l'activité d'un gène, soit en permettant son expression transitoire (à l'image de l'IES de *mtA* contenant une partie de la séquence régulatrice et des premiers acides aminés), soit en ajoutant des acides aminés nécessaires à la fonction de la protéine. Alimentant cette hypothèse, nous avons montré que le blocage des réarrangements provoque la non répression de certains gènes (exemple *PGM*) suggérant que la non excision des IES induirait un dérèglement transcriptionnel (Marmignon et al. 2014, Frapporti et al. 2019).

Je suis convaincu que l'IES du gène *mtA* n'est pas la seule IES avec un rôle fonctionnel, et la comparaison multi-espèces des séquences sera une source d'information considérable pour dénicher d'autres cas.

VII.2.3 Y a-t-il des gènes cachés dans le MIC ?

Chez les ciliés, le génome macronucléaire somatique est transcriptionnellement actif et contient les gènes fonctionnels, alors que le génome germinal micronucléaire n'est pas exprimé pendant la vie végétative. Comme chez la paramécie, le génome MIC de *Tetrahymena* subit des réarrangements de génome pendant le développement du MAC : une fragmentation des chromosomes et une élimination imprécise de grandes régions avec réligation des jonctions (voir **section III.3.1** p. 64). Lin et al. (2016) ont montré que certains mini-chromosomes sont perdus peu de temps après la division caryonidale. Un de ces chromosomes non maintenus contient le gène *TPB6* (*Tetrahymena* PiggyBac 6, une transposase domestiquée piggyBac) requis pour l'excision précise des quelques IES intragéniques de *Tetrahymena* (Feng et al. 2017). Chez *Euplotes crassus*, un autre cilié, Karamysheva et al. (2003) ont montré qu'un des gènes *TERT* (*EcTERT-2*), codant pour la sous-unité catalytique de la télomérase, est exprimé transitoirement pendant le développement macronucléaire. Cette régulation transcriptionnelle s'exerce par une disparition du gène dans les MAC végétatifs. Ces deux exemples révèlent que des séquences spécifiques de la lignée germinale peuvent porter des gènes fonctionnels. Comme pour les IES, abordées dans le paragraphe précédent, l'élimination de l'ADN conduit naturellement à une régulation transcriptionnelle efficace et originale (Singh et al. 2014).

La composition de la partie du génome MIC éliminée pendant les réarrangements reste encore largement inconnue. Entre 12% et 25% des séquences seraient occupées par des copies d'ET. Sous condition que les séquences répétées soient correctement assemblées (peu probable), les satellites ne couvriraient que 5% du génome MIC spécifique (Guérin et al. 2017). Les 70% restant sont encore un mystère. Toutes ces séquences, ne seraient-elles que de la *matière noire* (séquences inclassifiables pouvant provenir d'anciens ET ou de séquences à basse complexité) (Maumus and Quesneville 2014) ? En relation avec le paragraphe précédent, il est tentant de penser que ce génome contiendrait des gènes fonctionnels comme *TPB6* ou *EcTERT-2*. N'étant sous aucune pression de sélection, il est également possible que cette portion du génome MIC soit un réservoir à pseudogènes. Dans un futur proche, je souhaite m'atteler à percer les secrets, encore cachés, du génome MIC ...

Bibliographie

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. *A map of human genome variation from population-scale sequencing*. Nature, 467(7319) :1061–1073 (2010). doi :10.1038/nature09534. (Cité page 18.)

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Abril, J. F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M. H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D., Scheeler, F., Shen, H., Shue, B. C., Sidén-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z. Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., WoodageT, n., Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R. F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M., and Venter, J. C. *The genome sequence of Drosophila melanogaster*. Science (New York, N.Y.), 287(5461) :2185–2195 (2000). (Cité page 31.)

Adl, S. M., Bass, D., Lane, C. E., Lukeš, J., Schoch, C. L., Smirnov, A., Agatha, S., Berney, C., Brown, M. W., Burki, F., Cárdenas, P., Čepička, I., Chistyakova, L., Del Campo, J., Dunthorn, M., Edvardsen, B., Eglit, Y., Guillou, L., Hampl, V., Heiss, A. A., Hoppenrath, M., James, T. Y., Karnkowska, A., Karpov, S., Kim, E., Kolisko, M., Kudryavtsev, A.,

Lahr, D. J. G., Lara, E., Le Gall, L., Lynn, D. H., Mann, D. G., Massana, R., Mitchell, E. A. D., Morrow, C., Park, J. S., Pawlowski, J. W., Powell, M. J., Richter, D. J., Rueckert, S., Shadwick, L., Shimano, S., Spiegel, F. W., Torruella, G., Youssef, N., Zlatogursky, V., and Zhang, Q. *Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes*. The Journal of Eukaryotic Microbiology, 66(1) :4–119 (2019). doi :10.1111/jeu.12691. (Cité page 4.)

Adl, S. M., Simpson, A. G. B., Lane, C. E., Lukeš, J., Bass, D., Bowser, S. S., Brown, M. W., Burki, F., Dunthorn, M., Hampl, V., Heiss, A., Hoppenrath, M., Lara, E., Le Gall, L., Lynn, D. H., McManus, H., Mitchell, E. A. D., Mozley-Stanridge, S. E., Parfrey, L. W., Pawlowski, J., Rueckert, S., Shadwick, L., Shadwick, L., Schoch, C. L., Smirnov, A., and Spiegel, F. W. *The revised classification of eukaryotes*. The Journal of Eukaryotic Microbiology, 59(5) :429–493 (2012). doi :10.1111/j.1550-7408.2012.00644.x. (Cité pages 3, 4 et 51.)

Allen, J. E. and Salzberg, S. L. *JIGSAW : integration of multiple sources of evidence for gene prediction*. Bioinformatics (Oxford, England), 21(18) :3596–3603 (2005). doi :10.1093/bioinformatics/bti609. (Cité page 39.)

Allen, S. E., Hug, I., Pabian, S., Rzeszutek, I., Hoehener, C., and Nowacki, M. *Circular Concatemers of Ultra-Short DNA Segments Produce Regulatory RNAs*. Cell, 168(6) :990–999.e7 (2017). doi :10.1016/j.cell.2017.02.020. (Cité page 71.)

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. *Basic local alignment search tool*. Journal of Molecular Biology, 215(3) :403–410 (1990). doi :10.1016/S0022-2836(05)80360-2. (Cité page 48.)

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. *Gapped BLAST and PSI-BLAST : a new generation of protein database search programs*. Nucleic Acids Research, 25(17) :3389–3402 (1997). doi :10.1093/nar/25.17.3389. (Cité pages 34 et 36.)

Amar, L. and Dubrana, K. *Epigenetic control of chromosome breakage at the 5' end of Paramecium tetraurelia gene A*. Eukaryotic cell, 3(5) :1136–1146 (2004). doi :10.1128/EC.3.5.1136-1146.2004. (Cité page 74.)

Ambros, V. *microRNAs : tiny regulators with great potential*. Cell, 107(7) :823–826 (2001). (Cité page 13.)

Andersson, R., Sandelin, A., and Danko, C. G. *A unified architecture of transcriptional regulatory elements*. Trends in genetics : TIG, 31(8) :426–433 (2015). doi :10.1016/j.tig.2015.05.007. (Cité page 13.)

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L.-S. L. *UniProt : the Universal Protein knowledgebase*. Nucleic Acids Research, 32(Database issue) :D115–119 (2004). doi :10.1093/nar/gkh131. (Cité page 36.)

Arnaiz, O., Cain, S., Cohen, J., and Sperling, L. *ParameciumDB : a community resource that integrates the Paramecium tetraurelia genome sequence with genetic data*. Nucleic Acids Research, 35(Database issue) :D439–444 (2007). doi :10.1093/nar/gkl777. (Cité pages viii, ix et 40.)

Arnaiz, O., Cohen, J., Tassin, A.-M., and Koll, F. *Remodeling Cildb, a popular database for cilia and links for ciliopathies*. Cilia, 3 :9 (2014). doi :10.1186/2046-2530-3-9. (Cité pages x, 40 et 53.)

Arnaiz, O., Goût, J.-F., Bétermier, M., Bouhouche, K., Cohen, J., Duret, L., Kapusta, A., Meyer, E., and Sperling, L. *Gene expression in a paleopolyploid : a transcriptome resource for the ciliate Paramecium tetraurelia*. BMC genomics, 11 :547 (2010). doi : 10.1186/1471-2164-11-547. (Cité pages x, 71, 89, 91 et 157.)

Arnaiz, O., Malinowska, A., Klotz, C., Sperling, L., Dadlez, M., Koll, F., and Cohen, J. *Cildb : a knowledgebase for centrosomes and cilia*. Database : The Journal of Biological Databases and Curation, 2009 :bap022 (2009). doi :10.1093/database/bap022. (Cité pages x, 40 et 53.)

Arnaiz, O., Mathy, N., Baudry, C., Malinsky, S., Aury, J.-M., Denby Wilkes, C., Garnier, O., Labadie, K., Lauderdale, B. E., Le Mouël, A., Marmignon, A., Nowacki, M., Poulain, J., Prajer, M., Wincker, P., Meyer, E., Duharcourt, S., Duret, L., Bétermier, M., and Sperling, L. *The Paramecium germline genome provides a niche for intragenic parasitic DNA : evolutionary dynamics of internal eliminated sequences*. PLoS genetics, 8(10) :e1002984 (2012). doi :10.1371/journal.pgen.1002984. (Cité pages xii, xiii, xiv, 48, 67, 68, 69, 70, 71, 75, 79, 80, 105, 108, 127, 135, 157, 160, 172, 174 et 178.)

Arnaiz, O., Meyer, E., and Sperling, L. *ParameciumDB 2019 : integrating genomic data across the genus for functional and evolutionary biology*. Nucleic Acids Research, (Database issue) (2019). (Cité pages ix, 40, 56, 57 et 158.)

Arnaiz, O. and Sperling, L. *ParameciumDB in 2011 : new tools and new data for functional and comparative genomics of the model ciliate Paramecium tetraurelia*. Nucleic Acids Research, 39(Database issue) :D632–636 (2011). doi :10.1093/nar/gkq918. (Cité pages ix et 40.)

Arnaiz, O., Van Dijk, E., Bétermier, M., Lhuillier-Akakpo, M., de Vanssay, A., Duharcourt, S., Sallet, E., Gouzy, J., and Sperling, L. *Improved methods and resources for paramecium genomics : transcription units, gene annotation and gene expression*. BMC genomics, 18(1) :483 (2017). doi :10.1186/s12864-017-3887-z. (Cité pages x, xiii, xiv, 38, 71, 75, 82, 83, 89, 107, 135, 159, 160 et 167.)

Audoux, J., Philippe, N., Chikhi, R., Salson, M., Gallopin, M., Gabriel, M., Le Coz, J., Drouineau, E., Commes, T., and Gautheret, D. *DE-kupl : exhaustive capture of biological variation in RNA-seq data through k-mer decomposition*. Genome Biology, 18(1) :243 (2017). doi :10.1186/s13059-017-1372-2. (Cité page 39.)

Aury, J.-M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B. M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., Arnaiz, O., Billaut, A., Beisson, J., Blanc, I., Bouhouche, K., Câmara, F., Duharcourt, S., Guigo, R., Gogendeau, D., Katinka, M., Keller, A.-M., Kissmehl, R., Klotz, C., Koll, F., Le Mouël, A., Lepère, G., Malinsky, S., Nowacki, M., Nowak, J. K., Plattner, H., Poulain, J., Ruiz, F., Serrano, V., Zagulski, M., Dessen, P., Bétermier, M., Weissenbach, J., Scarpelli, C., Schächter, V., Sperling, L., Meyer, E., Cohen, J., and Wincker, P. *Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia*. Nature, 444(7116) :171–178 (2006). doi :10.1038/nature05230. (Cité pages ix, x, 56, 67, 68, 82, 83, 85, 86, 89, 106, 107, 135, 157, 159, 160 et 165.)

Bachmann-Gagescu, R. *Genetic complexity of ciliopathies and novel genes identification*. Medecine Sciences : M/S, 30(11) :1011–1023 (2014). doi :10.1051/medsci/20143011016. (Cité page 53.)

Balakirev, E. S. and Ayala, F. J. *Pseudogenes : are they "junk" or functional DNA ?* Annual Review of Genetics, 37 :123–151 (2003). doi :10.1146/annurev.genet.37.040103.103949. (Cité page 15.)

Baldauf, S. L. *The Deep Roots of Eukaryotes*. Science, 300(5626) :1703–1706 (2003). doi : 10.1126/science.1085544. (Cité page 54.)

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. *SPAdes : a new genome assembly algorithm and its applications to single-cell sequencing*. Journal of Computational Biology : A Journal of Computational Molecular Cell Biology, 19(5) :455–477 (2012). doi :10.1089/cmb.2012.0021. (Cité page 27.)

Bao, W., Kojima, K. K., and Kohany, O. *Repbase Update, a database of repetitive elements in eukaryotic genomes*. Mobile DNA, 6 :11 (2015). doi :10.1186/s13100-015-0041-9. (Cité pages 33 et 48.)

Bao, Z. and Eddy, S. R. *Automated de novo identification of repeat sequence families in sequenced genomes*. Genome Research, 12(8) :1269–1276 (2002). doi :10.1101/gr.88502. (Cité pages 33 et 49.)

Baroin, A., Prat, A., and Caron, F. *Telomeric site position heterogeneity in macronuclear DNA of Paramecium primaurelia.* Nucleic Acids Research, 15(4) :1717 (1987). (Cité page 74.)

Baroin-Tourancheau, A., Delgado, P., Perasso, R., and Adoutte, A. *A broad molecular phylogeny of ciliates : identification of major evolutionary trends and radiations within the phylum*. Proceedings of the National Academy of Sciences of the United States of America, 89(20) :9764–9768 (1992). doi :10.1073/pnas.89.20.9764. (Cité page 51.)

Barrell, B. G., Bankier, A. T., and Drouin, J. *A different genetic code in human mitochondria*. Nature, 282(5735) :189–194 (1979). (Cité page 15.)

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. *High-resolution profiling of histone methylations in the human genome*. Cell, 129(4) :823–837 (2007). doi :10.1016/j.cell.2007.05.009. (Cité page 8.)

Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J. P., and Lander, E. S. *ARACHNE : a whole-genome shotgun assembler*. Genome Research, 12(1) :177–189 (2002). doi :10.1101/gr.208902. (Cité pages 26 et 82.)

Baudry, C., Malinsky, S., Restituito, M., Kapusta, A., Rosa, S., Meyer, E., and Bétermier, M. *PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia*. Genes & Development, 23(21) :2478–2483 (2009). doi :10.1101/gad.547309. (Cité pages 68, 71, 79, 80 et 106.)

Beisson, J. *Preformed cell structure and cell heredity*. Prion, 2(1) :1–8 (2008). doi :10.4161/pri.2.1.5063. (Cité page 56.)

Beisson, J., Bétermier, M., Bré, M.-H., Cohen, J., Duharcourt, S., Duret, L., Kung, C., Malinsky, S., Meyer, E., Preer, J. R., and Sperling, L. *DNA microinjection into the macronucleus of paramecium*. Cold Spring Harbor Protocols, 2010(1) :pdb.prot5364 (2010). doi : 10.1101/pdb.prot5364. (Cité page 61.)

Benoit, V., Mucchielli-Giorgi, M.-H., Dumont, B., Durosay, P., Reymond, N., and Delacroix, H. *PPIDD : an extraction and visualisation method of biological protein-protein interfaces*. Biochimie, 90(4) :640–647 (2008). doi :10.1016/j.biochi.2007.11.008. (Cité page viii.)

Benson, G. *Tandem repeats finder : a program to analyze DNA sequences*. Nucleic Acids Research, 27(2) :573–580 (1999). doi :10.1093/nar/27.2.573. (Cité page 33.)

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. *Accurate whole human genome sequencing using reversible terminator chemistry*. Nature, 456(7218) :53–59 (2008). doi :10.1038/nature07517. (Cité page 21.)

Berger, D. J. D. *Nuclear differentiation and nucleic acid synthesis in well-fed exconjugants of Paramecium aurelia*. Chromosoma, 42(3) :247–268 (1973). doi :10.1007/BF00284774. (Cité pages 64, 66, 68 et 178.)

Berger, J. D. *Selective Autolysis of Nuclei as a Source of DNA Precursors for Other Nuclei within the Same Cell*. Science (New York, N.Y.), 158(3800) :524 (1967). doi :10.1126/science.158.3800.524-a. (Cité page 58.)

Bergman, C. M. and Quesneville, H. *Discovering and detecting transposable elements in ge-*

nome sequences. Briefings in Bioinformatics, 8(6) :382–392 (2007). doi :10.1093/bib/bbm048. (Cité pages 33 et 46.)

Betermier, M. and Duharcourt, S. *Programmed Rearrangement in Ciliates : Paramecium*. Microbiology Spectrum, 2(6) (2014). doi :10.1128/microbiolspec.MDNA3-0035-2014. (Cité pages 59, 64, 66, 71, 72, 77 et 78.)

Bhullar, S., Denby Wilkes, C., Arnaiz, O., Nowacki, M., Sperling, L., and Meyer, E. *A mating-type mutagenesis screen identifies a zinc-finger protein required for specific DNA excision events in Paramecium*. Nucleic Acids Research, 46(18) :9550–9562 (2018). doi : 10.1093/nar/gky772. (Cité page xi.)

Birney, E., Clamp, M., and Durbin, R. *GeneWise and Genomewise*. Genome Research, 14(5) :988–995 (2004). doi :10.1101/gr.1865504. (Cité page 36.)

Bischerour, J., Bhullar, S., Denby Wilkes, C., Régnier, V., Mathy, N., Dubois, E., Singh, A., Swart, E., Arnaiz, O., Sperling, L., Nowacki, M., and Bétermier, M. *Six domesticated PiggyBac transposases together carry out programmed DNA elimination in Paramecium*. eLife, 7 (2018). doi :10.7554/eLife.37927. (Cité pages xiii, 69, 71, 79, 80 et 174.)

Black, B. E. and Bassett, E. A. *The histone variant CENP-A and centromere specification*. Current Opinion in Cell Biology, 20(1) :91–100 (2008). doi :10.1016/j.ceb.2007.11.007. (Cité page 7.)

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. *Scaffolding preassembled contigs using SSPACE*. Bioinformatics (Oxford, England), 27(4) :578–579 (2011). doi :10.1093/bioinformatics/btq683. (Cité page 26.)

Bouhouche, K., Gout, J.-F., Kapusta, A., Bétermier, M., and Meyer, E. *Functional specialization of Piwi proteins in Paramecium tetraurelia from post-transcriptional gene silencing to genome remodelling*. Nucleic Acids Res, 39(10) :4249–64 (2011). doi :10.1093/nar/gkq1283. (Cité page 77.)

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A. J., Poux, S., Bougueleret, L., and Xenarios, I. *UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase : How to Use the Entry View*. Methods in Molecular Biology (Clifton, N.J.), 1374 :23–54 (2016). doi :10.1007/978-1-4939-3167-5_2. (Cité page 36.)

Bracht, J. R., Perlman, D. H., and Landweber, L. F. *Cytosine methylation and hydroxymethylation mark DNA for elimination in Oxytricha trifallax*. Genome Biology, 13(10) :R99 (2012). doi :10.1186/gb-2012-13-10-r99. (Cité page 82.)

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G. J. *Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila*. Cell, 128(6) :1089–1103 (2007). doi :10.1016/j.cell.2007.01.043. (Cité page 38.)

Brent, M. R. *Genome annotation past, present, and future : how to define an ORF at each locus*. Genome Research, 15(12) :1777–1786 (2005). doi :10.1101/gr.3866105. (Cité page 33.)

Brent, M. R. *How does eukaryotic gene prediction work ?* Nature Biotechnology, 25(8) :883–885 (2007). doi :10.1038/nbt0807-883. (Cité page 33.)

Brent, M. R. *Steady progress and recent breakthroughs in the accuracy of automated genome annotation*. Nature Reviews. Genetics, 9(1) :62–73 (2008). doi :10.1038/nrg2220. (Cité page 33.)

Bryant, W. B., Mills, M. K., Olson, B. J., and Michel, K. *Small RNA-Seq Analysis Reveals miRNA Expression Dynamics Across Tissues in the Malaria Vector, Anopheles gambiae*. G3 (Bethesda, Md.), 9(5) :1507–1517 (2019). doi :10.1534/g3.119.400104. (Cité page 35.)

Bétermier, M. *Large-scale genome remodelling by the developmentally programmed elimination of germ line sequences in the ciliate Paramecium*. Research in Microbiology, 155(5) :399–408 (2004). doi :10.1016/j.resmic.2004.01.017. (Cité page 68.)

Bétermier, M., Duharcourt, S., Seitz, H., and Meyer, E. *Timing of developmentally programmed excision and circularization of Paramecium internal eliminated sequences*. Molecular and cellular biology, 20(5) :1553–1561 (2000). (Cité pages 67, 71 et 176.)

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., Goodstein, D. M., Elsik, C. G., Lewis, S. E., Stein, L., and Holmes, I. H. *JBrowse : a dynamic web platform for genome visualization and analysis*. Genome Biology, 17 :66 (2016). doi :10.1186/s13059-016-0924-1. (Cité page 40.)

Buisine, N., Quesneville, H., and Colot, V. *Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets*. Genomics, 91(5) :467–475 (2008). doi :10.1016/j.ygeno.2008.01.005. (Cité page 33.)

Burge, C. and Karlin, S. *Prediction of complete gene structures in human genomic DNA*. Journal of Molecular Biology, 268(1) :78–94 (1997). doi :10.1006/jmbi.1997.0951. (Cité page 34.)

Burki, F. *The eukaryotic tree of life from a global phylogenomic perspective*. Cold Spring Harbor Perspectives in Biology, 6(5) :a016147 (2014). doi :10.1101/cshperspect.a016147. (Cité page 51.)

Burset, M. and Guigó, R. *Evaluation of gene structure prediction programs*. Genomics, 34(3) :353–367 (1996). doi :10.1006/geno.1996.0298. (Cité pages 39 et 40.)

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., Nusbaum, C., and Jaffe, D. B. *ALLPATHS : de novo assembly of whole-genome shotgun microreads*. Genome Research, 18(5) :810–820 (2008). doi :10.1101/gr.7337908. (Cité page 27.)

Caron, F. *A high degree of macronuclear chromosome polymorphism is generated by variable DNA rearrangements in Paramecium primaurelia during macronuclear differentiation*. Journal of Molecular Biology, 225(3) :661–678 (1992). doi :10.1016/0022-2836(92)90393-X. (Cité page 74.)

Caron, F. and Meyer, E. *Does Paramecium primaurelia use a different genetic code in its macronucleus ?* Nature, 314(6007) :185–188 (1985). doi :10.1038/314185a0. (Cité page 15.)

Carradec, Q., Götz, U., Arnaiz, O., Pouch, J., Simon, M., Meyer, E., and Marker, S. *Primary and secondary siRNA synthesis triggered by RNAs from food bacteria in the ciliate Paramecium tetraurelia*. Nucleic Acids Research, 43(3) :1818–1833 (2015). doi :10.1093/nar/gku1331. (Cité page xi.)

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M.-A., Méheust, R., Poulain, J., Romac, S., Richter, D. J., Yoshikawa, G., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Tara Oceans Coordinators, Jaillon, O., Aury, J.-M., Karsenti, E., Sullivan, M. B., Sunagawa, S., Bork, P., Not, F., Hingamp, P., Raes, J., Guidi, L., Ogata, H., de Vargas, C., Iudicone, D., Bowler, C., and

Wincker, P. *A global ocean atlas of eukaryotic genes*. Nature Communications, 9(1) :373 (2018). doi :10.1038/s41467-017-02342-1. (Cité pages 4 et 53.)

Casper, J., Zweig, A. S., Villarreal, C., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., Lee, C. M., Lee, B. T., Karolchik, D., Hinrichs, A. S., Haeussler, M., Guruvadoo, L., Navarro Gonzalez, J., Gibson, D., Fiddes, I. T., Eisenhart, C., Diekhans, M., Clawson, H., Barber, G. P., Armstrong, J., Haussler, D., Kuhn, R. M., and Kent, W. J. *The UCSC Genome Browser database : 2018 update*. Nucleic Acids Research, 46(D1) :D762–D769 (2018). doi : 10.1093/nar/gkx1020. (Cité page 40.)

Caspi, A. and Pachter, L. *Identification of transposable elements using multiple alignments of related genomes*. Genome Research, 16(2) :260–270 (2006). doi :10.1101/gr.4361206. (Cité page 50.)

Cassidy-Hanley, D., Bowen, J., Lee, J. H., Cole, E., VerPlank, L. A., Gaertig, J., Gorovsky, M. A., and Bruns, P. J. *Germline and somatic transformation of mating Tetrahymena thermophila by particle bombardment*. Genetics, 146(1) :135–147 (1997). (Cité page 61.)

Celniker, S. E., Wheeler, D. A., Kronmiller, B., Carlson, J. W., Halpern, A., Patel, S., Adams, M., Champe, M., Dugan, S. P., Frise, E., Hodgson, A., George, R. A., Hoskins, R. A., Laverty, T., Muzny, D. M., Nelson, C. R., Pacleb, J. M., Park, S., Pfeiffer, B. D., Richards, S., Sodergren, E. J., Svirskas, R., Tabor, P. E., Wan, K., Stapleton, M., Sutton, G. G., Venter, C., Weinstock, G., Scherer, S. E., Myers, E. W., Gibbs, R. A., and Rubin, G. M. *Finishing a whole-genome shotgun : release 3 of the Drosophila melanogaster euchromatic genome sequence*. Genome Biology, 3(12) :RESEARCH0079 (2002). (Cité pages 31 et 32.)

Cervantes, M. D., Hamilton, E. P., Xiong, J., Lawson, M. J., Yuan, D., Hadjithomas, M., Miao, W., and Orias, E. *Selecting one of several mating types through gene segment joining and deletion in Tetrahymena thermophila*. PLoS biology, 11(3) :e1001518 (2013). doi :10.1371/journal.pbio.1001518. (Cité page 61.)

Cervantes, M. D., Xi, X., Vermaak, D., Yao, M.-C., and Malik, H. S. *The CNA1 histone of the ciliate Tetrahymena thermophila is essential for chromosome segregation in the germline micronucleus*. Molecular Biology of the Cell, 17(1) :485–497 (2006). doi :10.1091/mbc.e05-07-0698. (Cité page 64.)

Chen, C.-L., Zhou, H., Liao, J.-Y., Qu, L.-H., and Amar, L. *Genome-wide evolutionary analysis of the noncoding RNA genes and noncoding DNA of Paramecium tetraurelia*. RNA, 15(4) :503–514 (2009). doi :10.1261/rna.1306009. (Cité pages 83 et 167.)

Chen, X., Bracht, J. R., Goldman, A. D., Dolzhenko, E., Clay, D. M., Swart, E. C., Perlman, D. H., Doak, T. G., Stuart, A., Amemiya, C. T., Sebra, R. P., and Landweber, L. F. *The Architecture of a Scrambled Genome Reveals Massive Levels of Genomic Rearrangement during Development*. Cell, 158(5) :1187–98 (2014). doi :10.1016/j.cell.2014.07.034. (Cité pages 65, 66 et 172.)

Cheng, C.-Y., Young, J. M., Lin, C.-Y. G., Chao, J.-L., Malik, H. S., and Yao, M.-C. *The piggyBac transposon-derived genes TPB1 and TPB6 mediate essential transposon-like excision during the developmental rearrangement of key genes in Tetrahymena thermophila*. Genes & Development, 30(24) :2724–2736 (2016). doi :10.1101/gad.290460.116. (Cité pages 65 et 174.)

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. *Continuous base identification for single-molecule nanopore DNA sequencing*. Nature Nanotechnology, 4(4) :265–270 (2009). doi :10.1038/nnano.2009.12. (Cité page 23.)

Coleman, A. W. *Paramecium aurelia revisited*. The Journal of Eukaryotic Microbiology, 52(1) :68–77 (2005). doi :10.1111/j.1550-7408.2005.3327r.x. (Cité page 56.)

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., and Mortazavi, A. *A survey of best practices for RNA-seq data analysis*. Genome Biology, 17 :13 (2016). doi :10.1186/s13059-016-0881-8. (Cité page 38.)

Consortium, T. E. P. *The ENCODE (ENCyclopedia Of DNA Elements) Project*. Science, 306(5696) :636–640 (2004). doi :10.1126/science.1105136. (Cité pages 17 et 32.)

Coyne, R. S., Lhuillier-Akakpo, M., and Duharcourt, S. *RNA-guided DNA rearrangements in ciliates : is the best genome defence a good offence ?* Biology of the Cell, 104(6) :309–325 (2012). doi :10.1111/boc.201100057. (Cité pages 64, 65, 74 et 77.)

Coyne, R. S., Thiagarajan, M., Jones, K. M., Wortman, J. R., Tallon, L. J., Haas, B. J., Cassidy-Hanley, D. M., Wiley, E. A., Smith, J. J., Collins, K., Lee, S. R., Couvillion, M. T., Liu, Y., Garg, J., Pearlman, R. E., Hamilton, E. P., Orias, E., Eisen, J. A., and Methé, B. A. *Refined annotation and assembly of the Tetrahymena thermophila genome sequence through EST analysis, comparative genomic hybridization, and targeted gap closure*. BMC genomics, 9 :562 (2008). doi :10.1186/1471-2164-9-562. (Cité pages 65 et 82.)

Cui, B. and Gorovsky, M. A. *Centromeric histone H3 is essential for vegetative cell division and for DNA elimination during conjugation in Tetrahymena thermophila*. Molecular and Cellular Biology, 26(12) :4499–4510 (2006). doi :10.1128/MCB.00079-06. (Cité page 64.)

Cummings, D. J., Tait, A., and Goddard, J. M. *Methylated bases in DNA from Paramecium aurelia*. Biochimica Et Biophysica Acta, 374(1) :1–11 (1974). doi :10.1016/0005-2787(74)90194-4. (Cité page 82.)

Curwen, V., Eyras, E., Andrews, T. D., Clarke, L., Mongin, E., Searle, S. M. J., and Clamp, M. *The Ensembl automatic gene annotation system*. Genome Research, 14(5) :942–950 (2004). doi :10.1101/gr.1858004. (Cité page 40.)

Denby Wilkes, C., Arnaiz, O., and Sperling, L. *ParTIES : a toolbox for Paramecium interspersed DNA elimination studies*. Bioinformatics (Oxford, England), 32(4) :599–601 (2016). doi :10.1093/bioinformatics/btv691. (Cité pages xii, xiii, xiv, 79, 135, 169 et 174.)

Denoeud, F., Aury, J.-M., Da Silva, C., Noel, B., Rogier, O., Delledonne, M., Morgante, M., Valle, G., Wincker, P., Scarpelli, C., Jaillon, O., and Artiguenave, F. *Annotating genomes with massive-scale RNA sequencing*. Genome Biology, 9(12) :R175 (2008). doi :10.1186/gb-2008-9-12-r175. (Cité page 38.)

Derocles, S., Bohan, D., J. Dumbrell, A., Kitson, J., Massol, F., Pauvert, C., Plantegenest, M., Vacher, C., and Evans, D. *Biomonitoring for the 21st Century : Integrating Next-Generation Sequencing Into Ecological Network Analysis*. Advances in Ecological Research, 58 :1–62 (2018). doi :10.1016/bs.aecr.2017.12.001. (Cité page 25.)

Derti, A., Garrett-Engele, P., Macisaac, K. D., Stevens, R. C., Sriram, S., Chen, R., Rohl, C. A., Johnson, J. M., and Babak, T. *A quantitative atlas of polyadenylation in five mammals*. Genome Research, 22(6) :1173–1183 (2012). doi :10.1101/gr.132563.111. (Cité page 36.)

Du, C., Caronna, J., He, L., and Dooner, H. K. *Computational prediction and molecular confirmation of Helitron transposons in the maize genome*. BMC genomics, 9 :51 (2008). doi :10.1186/1471-2164-9-51. (Cité page 50.)

Dubois, E., Bischerour, J., Marmignon, A., Mathy, N., Régnier, V., and Bétermier, M. *Transposon Invasion of the Paramecium Germline Genome Countered by a Domesticated PiggyBac Transposase and the NHEJ Pathway*. International Journal of Evolutionary Biology, 2012 :1–13 (2012). doi :10.1155/2012/436196. (Cité page 69.)

Duharcourt, S., Butler, A., and Meyer, E. *Epigenetic self-regulation of developmental excision of an internal eliminated sequence on Paramecium tetraurelia*. Genes & Development, 9(16) :2065–2077 (1995). (Cité page 76.)

Duharcourt, S., Keller, A. M., and Meyer, E. *Homology-dependent maternal inhibition of developmental excision of internal eliminated sequences in Paramecium tetraurelia*. Molecular and Cellular Biology, 18(12) :7075–7085 (1998). (Cité pages 76 et 176.)

Duharcourt, S. and Sperling, L. *The Challenges of Genome-Wide Studies in a Unicellular Eukaryote With Two Nuclear Genomes*. Methods in Enzymology, 612 :101–126 (2018). doi : 10.1016/bs.mie.2018.08.012. (Cité pages 54, 74 et 75.)

Dunn, N. A., Unni, D. R., Diesh, C., Munoz-Torres, M., Harris, N. L., Yao, E., Rasche, H., Holmes, I. H., Elsik, C. G., and Lewis, S. E. *Apollo : Democratizing genome annotation*. PLoS computational biology, 15(2) :e1006790 (2019). doi :10.1371/journal.pcbi.1006790. (Cité pages 35 et 40.)

Duret, L., Cohen, J., Jubin, C., Dessen, P., Goût, J.-F., Mousset, S., Aury, J.-M., Jaillon, O., Noël, B., Arnaiz, O., Bétermier, M., Wincker, P., Meyer, E., and Sperling, L. *Analysis of sequence variability in the macronuclear DNA of Paramecium tetraurelia : a somatic view of the germline*. Genome Research, 18(4) :585–596 (2008). doi :10.1101/gr.074534.107. (Cité pages xi, 71, 73, 74, 106, 129 et 170.)

Eberwine, J., Sul, J.-Y., Bartfai, T., and Kim, J. *The promise of single-cell sequencing*. Nature Methods, 11(1) :25–27 (2014). doi :10.1038/nmeth.2769. (Cité page 162.)

Echols, N., Harrison, P., Balasubramanian, S., Luscombe, N. M., Bertone, P., Zhang, Z., and Gerstein, M. *Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes*. Nucleic Acids Research, 30(11) :2515–2523 (2002). doi :10.1093/nar/30.11.2515. (Cité page 15.)

Eddy, S. R. *Non-coding RNA genes and the modern RNA world*. Nature Reviews. Genetics, 2(12) :919–929 (2001). doi :10.1038/35103511. (Cité page 13.)

Eddy, S. R. *A new generation of homology search tools based on probabilistic inference*. Genome Informatics. International Conference on Genome Informatics, 23(1) :205–211 (2009). (Cité page 48.)

Edgar, R. C. and Myers, E. W. *PILER : identification and classification of genomic repeats*. Bioinformatics (Oxford, England), 21 Suppl 1 :i152–158 (2005). doi :10.1093/bioinformatics/bti1003. (Cité page 49.)

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner,

S. *Real-time DNA sequencing from single polymerase molecules*. Science (New York, N.Y.), 323(5910) :133–138 (2009). doi :10.1126/science.1162986. (Cité page 23.)

Eisen, J. A., Coyne, R. S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J. R., Badger, J. H., Ren, Q., Amedeo, P., Jones, K. M., Tallon, L. J., Delcher, A. L., Salzberg, S. L., Silva, J. C., Haas, B. J., Majoros, W. H., Farzad, M., Carlton, J. M., Smith, R. K., Garg, J., Pearlman, R. E., Karrer, K. M., Sun, L., Manning, G., Elde, N. C., Turkewitz, A. P., Asai, D. J., Wilkes, D. E., Wang, Y., Cai, H., Collins, K., Stewart, B. A., Lee, S. R., Wilamowska, K., Weinberg, Z., Ruzzo, W. L., Wloga, D., Gaertig, J., Frankel, J., Tsao, C.-C., Gorovsky, M. A., Keeling, P. J., Waller, R. F., Patron, N. J., Cherry, J. M., Stover, N. A., Krieger, C. J., del Toro, C., Ryder, H. F., Williamson, S. C., Barbeau, R. A., Hamilton, E. P., and Orias, E. *Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote*. PLoS biology, 4(9) :e286 (2006). doi :10.1371/journal.pbio.0040286. (Cité pages 65 et 82.)

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., and Finn, R. D. *The Pfam protein families database in 2019*. Nucleic Acids Research, 47(D1) :D427–D432 (2019). doi :10.1093/nar/gky995. (Cité page 48.)

Ellinghaus, D., Kurtz, S., and Willhoeft, U. *LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons*. BMC Bioinformatics, 9(1) :18 (2008). doi : 10.1186/1471-2105-9-18. (Cité page 50.)

Elsik, C. G., Worley, K. C., Bennett, A. K., Beye, M., Camara, F., Childers, C. P., de Graaf, D. C., Debyser, G., Deng, J., Devreese, B., Elhaik, E., Evans, J. D., Foster, L. J., Graur, D., Guigo, R., HGSC production teams, Hoff, K. J., Holder, M. E., Hudson, M. E., Hunt, G. J., Jiang, H., Joshi, V., Khetani, R. S., Kosarev, P., Kovar, C. L., Ma, J., Maleszka, R., Moritz, R. F. A., Munoz-Torres, M. C., Murphy, T. D., Muzny, D. M., Newsham, I. F., Reese, J. T., Robertson, H. M., Robinson, G. E., Rueppell, O., Solovyev, V., Stanke, M., Stolle, E., Tsuruda, J. M., Vaerenbergh, M. V., Waterhouse, R. M., Weaver, D. B., Whitfield, C. W., Wu, Y., Zdobnov, E. M., Zhang, L., Zhu, D., Gibbs, R. A., and Honey Bee Genome Sequencing Consortium. *Finding the missing honey bee genes : lessons learned from a genome upgrade.* BMC genomics, 15 :86 (2014). doi :10.1186/1471-2164-15-86. (Cité page 31.)

Epstein, L. M. and Forney, J. D. *Mendelian and non-mendelian mutations affecting surface antigen expression in Paramecium tetraurelia*. Mol Cell Biol, 4(8) :1583–90 (1984). (Cité pages 61 et 75.)

Fan, Q. and Yao, M. *New telomere formation coupled with site-specific chromosome breakage in Tetrahymena thermophila*. Molecular and Cellular Biology, 16(3) :1267–1274 (1996). doi : 10.1128/mcb.16.3.1267. (Cité page 65.)

Fan, Q. and Yao, M. C. *A long stringent sequence signal for programmed chromosome breakage in Tetrahymena thermophila*. Nucleic Acids Research, 28(4) :895–900 (2000). doi :10.1093/nar/28.4.895. (Cité page 65.)

Fang, W., Wang, X., Bracht, J. R., Nowacki, M., and Landweber, L. F. *Piwi-interacting RNAs protect DNA against loss during Oxytricha genome rearrangement*. Cell, 151(6) :1243–55 (2012). doi :10.1016/j.cell.2012.10.045. (Cité page 66.)

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J. L., Haberle, V., Lassmann, T., Kulakovskiy, I. V.,

Lizio, M., Itoh, M., Andersson, R., Mungall, C. J., Meehan, T. F., Schmeier, S., Bertin, N., Jørgensen, M., Dimont, E., Arner, E., Schmidl, C., Schaefer, U., Medvedeva, Y. A., Plessy, C., Vitezic, M., Severin, J., Semple, C. A., Ishizu, Y., Young, R. S., Francescatto, M., Alam, I., Albanese, D., Altschuler, G. M., Arakawa, T., Archer, J. A. C., Arner, P., Babina, M., Rennie, S., Balwierz, P. J., Beckhouse, A. G., Pradhan-Bhatt, S., Blake, J. A., Blumenthal, A., Bodega, B., Bonetti, A., Briggs, J., Brombacher, F., Burroughs, A. M., Califano, A., Cannistraci, C. V., Carbajo, D., Chen, Y., Chierici, M., Ciani, Y., Clevers, H. C., Dalla, E., Davis, C. A., Detmar, M., Diehl, A. D., Dohi, T., Drabløs, F., Edge, A. S. B., Edinger, M., Ekwall, K., Endoh, M., Enomoto, H., Fagiolini, M., Fairbairn, L., Fang, H., Farach-Carson, M. C., Faulkner, G. J., Favorov, A. V., Fisher, M. E., Frith, M. C., Fujita, R., Fukuda, S., Furlanello, C., Furino, M., Furusawa, J.-i., Geijtenbeek, T. B., Gibson, A. P., Gingeras, T., Goldowitz, D., Gough, J., Guhl, S., Guler, R., Gustincich, S., Ha, T. J., Hamaguchi, M., Hara, M., Harbers, M., Harshbarger, J., Hasegawa, A., Hasegawa, Y., Hashimoto, T., Herlyn, M., Hitchens, K. J., Ho Sui, S. J., Hofmann, O. M., Hoof, I., Hori, F., Huminiecki, L., Iida, K., Ikawa, T., Jankovic, B. R., Jia, H., Joshi, A., Jurman, G., Kaczkowski, B., Kai, C., Kaida, K., Kaiho, A., Kajiyama, K., Kanamori-Katayama, M., Kasianov, A. S., Kasukawa, T., Katayama, S., Kato, S., Kawaguchi, S., Kawamoto, H., Kawamura, Y. I., Kawashima, T., Kempfle, J. S., Kenna, T. J., Kere, J., Khachigian, L. M., Kitamura, T., Klinken, S. P., Knox, A. J., Kojima, M., Kojima, S., Kondo, N., Koseki, H., Koyasu, S., Krampitz, S., Kubosaki, A., Kwon, A. T., Laros, J. F. J., Lee, W., Lennartsson, A., Li, K., Lilje, B., Lipovich, L., Mackay-Sim, A., Manabe, R.-i., Mar, J. C., Marchand, B., Mathelier, A., Mejhert, N., Meynert, A., Mizuno, Y., de Lima Morais, D. A., Morikawa, H., Morimoto, M., Moro, K., Motakis, E., Motohashi, H., Mummery, C. L., Murata, M., Nagao-Sato, S., Nakachi, Y., Nakahara, F., Nakamura, T., Nakamura, Y., Nakazato, K., van Nimwegen, E., Ninomiya, N., Nishiyori, H., Noma, S., Noma, S., Noazaki, T., Ogishima, S., Ohkura, N., Ohimiya, H., Ohno, H., Ohshima, M., Okada-Hatakeyama, M., Okazaki, Y., Orlando, V., Ovchinnikov, D. A., Pain, A., Passier, R., Patrikakis, M., Persson, H., Piazza, S., Prendergast, J. G. D., Rackham, O. J. L., Ramilowski, J. A., Rashid, M., Ravasi, T., Rizzu, P., Roncador, M., Roy, S., Rye, M. B., Saijyo, E., Sajantila, A., Saka, A., Sakaguchi, S., Sakai, M., Sato, H., Savvi, S., Saxena, A., Schneider, C., Schultes, E. A., Schulze-Tanzil, G. G., Schwegmann, A., Sengstag, T., Sheng, G., Shimoji, H., Shimoni, Y., Shin, J. W., Simon, C., Sugiyama, D., Sugiyama, T., Suzuki, M., Suzuki, N., Swoboda, R. K., 't Hoen, P. A. C., Tagami, M., Takahashi, N., Takai, J., Tanaka, H., Tatsukawa, H., Tatum, Z., Thompson, M., Toyodo, H., Toyoda, T., Valen, E., van de Wetering, M., van den Berg, L. M., Verado, R., Vijayan, D., Vorontsov, I. E., Wasserman, W. W., Watanabe, S., Wells, C. A., Winteringham, L. N., Wolvetang, E., Wood, E. J., Yamaguchi, Y., Yamamoto, M., Yoneda, M., Yonekura, Y., Yoshida, S., Zabierowski, S. E., Zhang, P. G., Zhao, X., Zucchelli, S., Summers, K. M., Suzuki, H., Daub, C. O., Kawai, J., Heutink, P., Hide, W., Freeman, T. C., Lenhard, B., Bajic, V. B., Taylor, M. S., Makeev, V. J., Sandelin, A., Hume, D. A., Carninci, P., and Hayashizaki, Y. *A promoter-level mammalian expression atlas*. Nature, 507(7493) :462–470 (2014). doi :10.1038/nature13182. (Cité page 38.)

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., and Turcatti, G. *BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies*. Nucleic Acids Research, 34(3) :e22 (2006). doi :10.1093/nar/gnj023. (Cité page 21.)

Feng, L., Wang, G., Hamilton, E. P., Xiong, J., Yan, G., Chen, K., Chen, X., Dui, W., Plemens, A., Khadr, L., Dhanekula, A., Juma, M., Dang, H. Q., Kapler, G. M., Orias, E., Miao, W., and Liu, Y. *A germline-limited piggyBac transposase gene is required for precise excision in Tetrahymena genome rearrangement*. Nucleic Acids Research, 45(16) :9481–9502 (2017). doi :10.1093/nar/gkx652. (Cité pages 65, 174 et 179.)

Finnegan, D. J. *Eukaryotic transposable elements and genome evolution*. Trends in genetics : TIG, 5(4) :103–107 (1989). (Cité pages 12 et 42.)

Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. *Considering transposable element diversification in de novo annotation approaches*. PloS One, 6(1) :e16526 (2011). doi :10.1371/journal.pone.0016526. (Cité pages 46 et 172.)

Foissac, S. and Schiex, T. *Integrating alternative splicing detection into gene prediction*. BMC bioinformatics, 6 :25 (2005). doi :10.1186/1471-2105-6-25. (Cité pages 34 et 39.)

Forney, J. D. and Blackburn, E. H. *Developmentally controlled telomere addition in wild-type and mutant paramecia.* Molecular and Cellular Biology, 8(1) :251–258 (1988). doi :10.1128/MCB.8.1.251. (Cité pages 74 et 75.)

Frapporti, A., Miró Pina, C., Arnaiz, O., Holoch, D., Kawaguchi, T., Humbert, A., Eleftheriou, E., Lombard, B., Loew, D., Sperling, L., Guitot, K., Margueron, R., and Duharcourt, S. *The Polycomb protein Ezl1 mediates H3k9 and H3k27 methylation to repress transposable elements in Paramecium*. Nature Communications, 10(1) :2710 (2019). doi :10.1038/s41467-019-10648-5. (Cité pages xii, 77, 79, 171 et 178.)

Furrer, D. I., Swart, E. C., Kraft, M. F., Sandoval, P. Y., and Nowacki, M. *Two Sets of Piwi Proteins Are Involved in Distinct sRNA Pathways Leading to Elimination of Germline-Specific DNA*. Cell Reports, 20(2) :505–520 (2017). doi :10.1016/j.celrep.2017.06.050. (Cité page 79.)

Gaertig, J. and Kapler, G. *Transient and stable DNA transformation of Tetrahymena thermophila by electroporation*. Methods in Cell Biology, 62 :485–500 (2000). doi :10.1016/s0091-679x(08)61552-6. (Cité page 61.)

Galvani, A. and Sperling, L. *Transgene-mediated post-transcriptional gene silencing is inhibited by 3′ non-coding sequences in Paramecium*. Nucleic acids research, 29(21) :4387–4394 (2001). (Cité page 63.)

Galvani, A. and Sperling, L. *RNA interference by feeding in Paramecium*. Trends in genetics : TIG, 18(1) :11–12 (2002). (Cité page 63.)

Gentekaki, E., Kolisko, M., Boscaro, V., Bright, K. J., Dini, F., Di Giuseppe, G., Gong, Y., Miceli, C., Modeo, L., Molestina, R. E., Petroni, G., Pucciarelli, S., Roger, A. J., Strom, S. L., and Lynn, D. H. *Large-scale phylogenomic analysis reveals the phylogenetic position of the problematic taxon Protocruzia and unravels the deep phylogenetic affinities of the ciliate lineages*. Molecular Phylogenetics and Evolution, 78 :36–42 (2014). doi :10.1016/j.ympev.2014.04.020. (Cité page 55.)

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., and Snyder, M. *What is a gene, post-ENCODE ? History and updated definition*. Genome Research, 17(6) :669–681 (2007). doi :10.1101/gr.6339607. (Cité page 12.)

Gilley, D. and Blackburn, E. H. *Lack of telomere shortening during senescence in Paramecium*. Proceedings of the National Academy of Sciences of the United States of America, 91(5) :1955–1958 (1994). doi :10.1073/pnas.91.5.1955. (Cité page 58.)

Goerner-Potvin, P. and Bourque, G. *Computational tools to unmask transposable elements*. Nature Reviews. Genetics, 19(11) :688–704 (2018). doi :10.1038/s41576-018-0050-x. (Cité page 46.)

Gogendeau, D., Klotz, C., Arnaiz, O., Malinowska, A., Dadlez, M., de Loubresse, N. G., Ruiz, F., Koll, F., and Beisson, J. *Functional diversification of centrins and cell morphological complexity*. Journal of Cell Science, 121(Pt 1) :65–74 (2008). doi :10.1242/jcs.019414. (Cité page ix.)

Gogendeau, D., Lemullois, M., Aubusson-Fleury, A., Arnaiz, O., Cohen, J., Vesque, C., Schneider-Maunoury, S., Koll, F., and Tassin, A.-M. *MKS-NPHP module proteins regulate ciliary shedding in Paramecium*. bioRxiv, page 676395 (2019). doi :10.1101/676395. (Cité page ix.)

Goodwin, S., McPherson, J. D., and McCombie, W. R. *Coming of age : ten years of next-generation sequencing technologies*. Nature Reviews. Genetics, 17(6) :333–351 (2016). doi : 10.1038/nrg.2016.49. (Cité pages 24 et 29.)

Goubert, C., Modolo, L., Vieira, C., ValienteMoro, C., Mavingui, P., and Boulesteix, M. *De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti)*. Genome Biology and Evolution, 7(4) :1192–1205 (2015). doi :10.1093/gbe/evv050. (Cité page 49.)

Gout, J.-F., Johri, P., Arnaiz, O., Doak, T. G., Bhullar, S., Couloux, A., Guérin, F., Malinsky, S., Sperling, L., Labadie, K., Meyer, E., Duharcourt, S., and Lynch, M. *Universal trends of post-duplication evolution revealed by the genomes of 13 Paramecium species sharing an ancestral whole-genome duplication*. bioRxiv, page 573576 (2019). doi :10.1101/573576. (Cité pages xii, 160 et 165.)

Gout, J.-F., Kahn, D., Duret, L., and Paramecium Post-Genomics Consortium. *The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution*. PLoS genetics, 6(5) :e1000944 (2010). doi :10.1371/journal.pgen.1000944. (Cité pages 85, 89 et 157.)

Gout, J.-F. and Lynch, M. *Maintenance and Loss of Duplicated Genes by Dosage Subfunctionalization*. Molecular Biology and Evolution, 32(8) :2141–2148 (2015). doi : 10.1093/molbev/msv095. (Cité page 89.)

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. *Full-length transcriptome assembly from RNA-Seq data without a reference genome*. Nature Biotechnology, 29(7) :644–652 (2011). doi :10.1038/nbt.1883. (Cité page 38.)

Gratias, A. and Bétermier, M. *Developmentally programmed excision of internal DNA sequences in Paramecium aurelia*. Biochimie, 83(11-12) :1009–1022 (2001). (Cité pages 68 et 71.)

Gratias, A. and Bétermier, M. *Processing of Double-Strand Breaks Is Involved in the Precise Excision of Paramecium Internal Eliminated Sequences*. Molecular and Cellular Biology, 23(20) :7152–7162 (2003). doi :10.1128/MCB.23.20.7152-7162.2003. (Cité page 71.)

Gratias, A., Lepere, G., Garnier, O., Rosa, S., Duharcourt, S., Malinsky, S., Meyer, E., and Betermier, M. *Developmentally programmed DNA splicing in Paramecium reveals short-distance crosstalk between DNA cleavage sites*. Nucleic Acids Research, 36(10) :3244–3251 (2008). doi :10.1093/nar/gkn154. (Cité page 68.)

Green, E. D. *Strategies for the systematic sequencing of complex genomes*. Nature Reviews. Genetics, 2(8) :573–583 (2001). doi :10.1038/35084503. (Cité pages 26 et 27.)

Greider, C. W. and Blackburn, E. H. *Identification of a specific telomere terminal transferase activity in Tetrahymena extracts*. Cell, 43(2 Pt 1) :405–413 (1985). (Cité page 7.)

Gruchota, J., Denby Wilkes, C., Arnaiz, O., Sperling, L., and Nowak, J. K. *A meiosis-specific Spt5 homolog involved in non-coding transcription*. Nucleic Acids Research, 45(8) :4722–4732 (2017). doi :10.1093/nar/gkw1318. (Cité pages xii, 77 et 80.)

Guberman, J. M., Ai, J., Arnaiz, O., Baran, J., Blake, A., Baldock, R., Chelala, C., Croft, D., Cros, A., Cutts, R. J., Di Génova, A., Forbes, S., Fujisawa, T., Gadaleta, E., Goodstein, D. M., Gundem, G., Haggarty, B., Haider, S., Hall, M., Harris, T., Haw, R., Hu, S., Hubbard, S., Hsu, J., Iyer, V., Jones, P., Katayama, T., Kinsella, R., Kong, L., Lawson, D., Liang, Y., Lopez-Bigas, N., Luo, J., Lush, M., Mason, J., Moreews, F., Ndegwa, N., Oakley, D., Perez-Llamas, C., Primig, M., Rivkin, E., Rosanoff, S., Shepherd, R., Simon, R., Skarnes, B., Smedley, D., Sperling, L., Spooner, W., Stevenson, P., Stone, K., Teague, J., Wang, J., Wang, J., Whitty, B., Wong, D. T., Wong-Erasmus, M., Yao, L., Youens-Clark, K., Yung, C., Zhang, J., and Kasprzyk, A. *BioMart Central Portal : an open database network for the biological community*. Database : The Journal of Biological Databases and Curation, 2011 :bar041 (2011). doi :10.1093/database/bar041. (Cité page x.)

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. *QUAST : quality assessment tool for genome assemblies*. Bioinformatics (Oxford, England), 29(8) :1072–1075 (2013). doi :10.1093/bioinformatics/btt086. (Cité page 28.)

Guérin, F., Arnaiz, O., Boggetto, N., Denby Wilkes, C., Meyer, E., Sperling, L., and Duharcourt, S. *Flow cytometry sorting of nuclei enables the first global characterization of Paramecium germline DNA and transposable elements*. BMC genomics, 18(1) :327 (2017). doi :10.1186/s12864-017-3713-7. (Cité pages xii, xiii, xiv, 48, 61, 66, 69, 74, 75, 80, 107, 135, 169, 172, 174, 176 et 179.)

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., and Wortman, J. R. *Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments*. Genome Biology, 9(1) :R7 (2008). doi :10.1186/gb-2008-9-1-r7. (Cité page 39.)

Haas, B. J., Zeng, Q., Pearson, M. D., Cuomo, C. A., and Wortman, J. R. *Approaches to Fungal Genome Annotation*. Mycology, 2(3) :118–141 (2011). doi :10.1080/21501203.2011.606851. (Cité page 40.)

Haberle, V. and Stark, A. *Eukaryotic core promoters and the functional basis of transcription initiation*. Nature Reviews. Molecular Cell Biology, 19(10) :621–637 (2018). doi :10.1038/s41580-018-0028-8. (Cité page 13.)

Hamilton, E. P., Kapusta, A., Huvos, P. E., Bidwell, S. L., Zafar, N., Tang, H., Hadjithomas, M., Krishnakumar, V., Badger, J. H., Caler, E. V., Russ, C., Zeng, Q., Fan, L., Levin, J. Z., Shea, T., Young, S. K., Hegarty, R., Daza, R., Gujja, S., Wortman, J. R., Birren, B. W., Nusbaum, C., Thomas, J., Carey, C. M., Pritham, E. J., Feschotte, C., Noto, T., Mochizuki, K., Papazyan, R., Taverna, S. D., Dear, P. H., Cassidy-Hanley, D. M., Xiong, J., Miao, W., Orias, E., and Coyne, R. S. *Structure of the germline genome ofTetrahymena thermophilaand relationship to the massively rearranged somatic genome*. eLife, 5 (2016). doi :10.7554/eLife.19090. (Cité pages 65, 172 et 174.)

Han, Y. and Wessler, S. R. *MITE-Hunter : a program for discovering miniature inverted-repeat transposable elements from genomic sequences*. Nucleic Acids Research, 38(22) :e199 (2010). doi :10.1093/nar/gkq862. (Cité page 33.)

Hardwick, S. A., Joglekar, A., Flicek, P., Frankish, A., and Tilgner, H. U. *Getting the Entire Message : Progress in Isoform Sequencing*. Frontiers in Genetics, 10 :709 (2019). doi : 10.3389/fgene.2019.00709. (Cité page 23.)

He, D., Fiz-Palacios, O., Fu, C.-J., Fehling, J., Tsai, C.-C., and Baldauf, S. L. *An alternative root for the eukaryote tree of life*. Current biology : CB, 24(4) :465–470 (2014). doi :10.1016/j.cub.2014.01.036. (Cité page 54.)

He, M., Wang, J., Fan, X., Liu, X., Shi, W., Huang, N., Zhao, F., and Miao, M. *Genetic basis for the establishment of endosymbiosis in Paramecium*. The ISME Journal, 13(5) :1360–1369 (2019). doi :10.1038/s41396-018-0341-4. (Cité page 160.)

Heather, J. M. and Chain, B. *The sequence of sequencers : The history of sequencing DNA*. Genomics, 107(1) :1–8 (2016). doi :10.1016/j.ygeno.2015.11.003. (Cité page 18.)

Henson, J., Tischler, G., and Ning, Z. *Next-generation sequencing and large genome assemblies*. Pharmacogenomics, 13(8) :901–915 (2012). doi :10.2217/pgs.12.72. (Cité page 29.)

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., and Quesneville, H. *PASTEC : an automatic transposable element classification tool*. PloS One, 9(5) :e91929 (2014). doi :10.1371/journal.pone.0091929. (Cité page 42.)

Hoehener, C., Hug, I., and Nowacki, M. *Dicer-like Enzymes with Sequence Cleavage Preferences*. Cell, 173(1) :234–247.e7 (2018). doi :10.1016/j.cell.2018.02.029. (Cité page 77.)

Holt, C. and Yandell, M. *MAKER2 : an annotation pipeline and genome-database management tool for second-generation genome projects*. BMC bioinformatics, 12 :491 (2011). doi :10.1186/1471-2105-12-491. (Cité page 40.)

Howe, K. L., Chothia, T., and Durbin, R. *GAZE : a generic framework for the integration of gene-prediction data by dynamic programming*. Genome Research, 12(9) :1418–1427 (2002). doi :10.1101/gr.149502. (Cité pages 34 et 39.)

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., Smit, A. F. A., and Wheeler, T. J. *The Dfam database of repetitive DNA families*. Nucleic Acids Research, 44(D1) :D81–89 (2016). doi :10.1093/nar/gkv1272. (Cité page 48.)

Hugenholtz, P., Skarshewski, A., and Parks, D. H. *Genome-Based Microbial Taxonomy Coming of Age*. Cold Spring Harbor Perspectives in Biology, 8(6) (2016). doi : 10.1101/cshperspect.a018085. (Cité page 4.)

Ignarski, M., Singh, A., Swart, E. C., Arambasic, M., Sandoval, P. Y., and Nowacki, M. *Paramecium tetraurelia chromatin assembly factor-1-like protein PtCAF-1 is involved in RNA-mediated control of DNA elimination*. Nucleic Acids Research, 42(19) :11952–11964 (2014). doi :10.1093/nar/gku874. (Cité page 80.)

International Human Genome Sequencing Consortium. *Finishing the euchromatic sequence of the human genome*. Nature, 431(7011) :931–945 (2004). doi :10.1038/nature03001. (Cité pages 17 et 31.)

International Wheat Genome Sequencing Consortium (IWGSC), IWGSC RefSeq principal investigators :, Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., IWGSC whole-genome assembly principal investigators :, Pozniak, C. J., Stein, N., Choulet, F., Distelfeld, A., Eversole, K., Poland, J., Rogers, J., Ronen, G., Sharpe, A. G., Whole-genome sequencing and assembly :, Pozniak, C., Ronen, G., Stein, N., Barad, O.,

Baruch, K., Choulet, F., Keeble-Gagnère, G., Mascher, M., Sharpe, A. G., Ben-Zvi, G., Josselin, A.-A., Hi-C data-based scaffolding :, Stein, N., Mascher, M., Himmelbach, A., Whole-genome assembly quality control and analyses :, Choulet, F., Keeble-Gagnère, G., Mascher, M., Rogers, J., Balfourier, F., Gutierrez-Gonzalez, J., Hayden, M., Josselin, A.-A., Koh, C., Muehlbauer, G., Pasam, R. K., Paux, E., Pozniak, C. J., Rigault, P., Sharpe, A. G., Tibbits, J., Tiwari, V., Pseudomolecule assembly :, Choulet, F., Keeble-Gagnère, G., Mascher, M., Josselin, A.-A., Rogers, J., RefSeq genome structure and gene analyses :, Spannagl, M., Choulet, F., Lang, D., Gundlach, H., Haberer, G., Keeble-Gagnère, G., Mayer, K. F. X., Ormanbekova, D., Paux, E., Prade, V., Šimková, H., Wicker, T., Automated annotation :, Choulet, F., Spannagl, M., Swarbreck, D., Rimbert, H., Felder, M., Guilhot, N., Gundlach, H., Haberer, G., Kaithakottil, G., Keilwagen, J., Lang, D., Leroy, P., Lux, T., Mayer, K. F. X., Twardziok, S., Venturini, L., Manual gene curation :, Appels, R., Rimbert, H., Choulet, F., Juhász, A., Keeble-Gagnère, G., Subgenome comparative analyses :, Choulet, F., Spannagl, M., Lang, D., Abrouk, M., Haberer, G., Keeble-Gagnère, G., Mayer, K. F. X., Wicker, T., Transposable elements :, Choulet, F., Wicker, T., Gundlach, H., Lang, D., Spannagl, M., Phylogenomic analyses :, Lang, D., Spannagl, M., Appels, R., Fischer, I., Transcriptome analyses and RNA-seq data :, Uauy, C., Borrill, P., Ramirez-Gonzalez, R. H., Appels, R., Arnaud, D., Chalabi, S., Chalhoub, B., Choulet, F., Cory, A., Datla, R., Davey, M. W., Hayden, M., Jacobs, J., Lang, D., Robinson, S. J., Spannagl, M., Steuernagel, B., Tibbits, J., Tiwari, V., van Ex, F., Wulff, B. B. H., Whole-genome methylome :, Pozniak, C. J., Robinson, S. J., Sharpe, A. G., Cory, A., Histone mark analyses :, Benhamed, M., Paux, E., Bendahmane, A., Concia, L., Latrasse, D., BAC chromosome MTP IWGSC–Bayer Whole-Genome Profiling (WGP) tags :, Rogers, J., Jacobs, J., Alaux, M., Appels, R., Bartoš, J., Bellec, A., Berges, H., Doležel, J., Feuillet, C., Frenkel, Z., Gill, B., Korol, A., Letellier, T., Olsen, O.-A., Šimková, H., Singh, K., Valárik, M., van der Vossen, E., Vautrin, S., Weining, S., Chromosome LTC mapping and physical mapping quality control :, Korol, A., Frenkel, Z., Fahima, T., Glikson, V., Raats, D., Rogers, J., RH mapping :, Tiwari, V., Gill, B., Paux, E., Poland, J., Optical mapping :, Doležel, J., Číhalíková, J., Šimková, H., Toegelová, H., Vrána, J., Recombination analyses :, Sourdille, P., Darrier, B., Gene family analyses :, Appels, R., Spannagl, M., Lang, D., Fischer, I., Ormanbekova, D., Prade, V., CBF gene family :, Barabaschi, D., Cattivelli, L., Dehydrin gene family :, Hernandez, P., Galvez, S., Budak, H., NLR gene family :, Steuernagel, B., Jones, J. D. G., Witek, K., Wulff, B. B. H., Yu, G., PPR gene family :, Small, I., Melonek, J., Zhou, R., Prolamin gene family :, Juhász, A., Belova, T., Appels, R., Olsen, O.-A., WAK gene family :, Kanyuka, K., King, R., Stem solidness (SSt1) QTL team :, Nilsen, K., Walkowiak, S., Pozniak, C. J., Cuthbert, R., Datla, R., Knox, R., Wiebe, K., Xiang, D., Flowering locus C (FLC) gene team :, Rohde, A., Golds, T., Genome size analysis :, Doležel, J., Čížková, J., Tibbits, J., MicroRNA and tRNA annotation :, Budak, H., Akpinar, B. A., Biyiklioglu, S., Genetic maps and mapping :, Muehlbauer, G., Poland, J., Gao, L., Gutierrez-Gonzalez, J., N'Daiye, A., BAC libraries and chromosome sorting :, Doležel, J., Šimková, H., Číhalíková, J., Kubaláková, M., Šafář, J., Vrána, J., BAC pooling, BAC library repository, and access :, Berges, H., Bellec, A., Vautrin, S., IWGSC sequence and data repository and access :, Alaux, M., Alfama, F., Adam-Blondon, A.-F., Flores, R., Guerche, C., Letellier, T., Loaec, M., Quesneville, H., Physical maps and BAC-based sequences :, 1A BAC sequencing and assembly :, Pozniak, C. J., Sharpe, A. G., Walkowiak, S., Budak, H., Condie, J., Ens, J., Koh, C., Maclachlan, R., Tan, Y., Wicker, T., 1B BAC sequencing and assembly :, Choulet, F., Paux, E., Alberti, A., Aury, J.-M., Balfourier, F., Barbe, V., Couloux, A., Cruaud, C., Labadie, K., Mangenot, S., Wincker, P., 1D, 4D, and 6D physical mapping :, Gill, B., Kaur, G., Luo, M., Sehgal, S., 2AL physical mapping :, Singh, K., Chhuneja, P., Gupta, O. P., Jindal, S., Kaur, P., Malik, P., Sharma, P., Yadav, B.,

2AS physical mapping :, Singh, N. K., Khurana, J.,. *Shifting the limits in wheat research and breeding using a fully annotated reference genome*. Science (New York, N.Y.), 361(6403) (2018). doi :10.1126/science.aar7191. (Cité page 32.)

Jaffe, D. B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J. P., Zody, M. C., and Lander, E. S. *Whole-genome sequence assembly for mammalian genomes : Arachne 2*. Genome Research, 13(1) :91–96 (2003). doi :10.1101/gr.828403. (Cité page 82.)

Jaillon, O., Bouhouche, K., Gout, J.-F., Aury, J.-M., Noel, B., Saudemont, B., Nowacki, M., Serrano, V., Porcel, B. M., Ségurens, B., Le Mouël, A., Lepère, G., Schächter, V., Bétermier, M., Cohen, J., Wincker, P., Sperling, L., Duret, L., and Meyer, E. *Translational control of intron splicing in eukaryotes*. Nature, 451(7176) :359–362 (2008). doi :10.1038/nature06495. (Cité pages 73, 83 et 84.)

Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B., and Akeson, M. *Improved data analysis for the MinION nanopore sequencer*. Nature Methods, 12(4) :351–356 (2015). doi :10.1038/nmeth.3290. (Cité page 23.)

Jeck, W. R., Reinhardt, J. A., Baltrus, D. A., Hickenbotham, M. T., Magrini, V., Mardis, E. R., Dangl, J. L., and Jones, C. D. *Extending assembly of short DNA sequences to handle error*. Bioinformatics (Oxford, England), 23(21) :2942–2944 (2007). doi : 10.1093/bioinformatics/btm451. (Cité page 26.)

Johri, P., Marinov, G. K., Doak, T. G., and Lynch, M. *Population Genetics of Paramecium Mitochondrial Genomes : Recombination, Mutation Spectrum, and Efficacy of Selection*. Genome Biology and Evolution, 11(5) :1398–1416 (2019). doi :10.1093/gbe/evz081. (Cité page 56.)

Jones, W. *Nuclear differentiation in Paramecium.* Ph.D. thesis, University of Wales, Aberystwyth, UK. (1956). (Cité page 66.)

Jukes, T. H. and Osawa, S. *Evolutionary changes in the genetic code*. Comparative Biochemistry and Physiology. B, Comparative Biochemistry, 106(3) :489–494 (1993). (Cité page 15.)

Jurka, J., Kapitonov, V. V., Kohany, O., and Jurka, M. V. *Repetitive sequences in complex genomes : structure and evolution*. Annual Review of Genomics and Human Genetics, 8 :241–259 (2007). doi :10.1146/annurev.genom.8.080706.092416. (Cité pages 10, 12 et 33.)

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. *Repbase Update, a database of eukaryotic repetitive elements*. Cytogenetic and Genome Research, 110(1-4) :462–467 (2005). doi :10.1159/000084979. (Cité page 48.)

Jurka, J., Klonowski, P., Dagman, V., and Pelton, P. *CENSOR–a program for identification and elimination of repetitive elements from DNA sequences*. Computers & Chemistry, 20(1) :119–121 (1996). (Cité page 48.)

Juven-Gershon, T., Hsu, J.-Y., Theisen, J. W., and Kadonaga, J. T. *The RNA polymerase II core promoter - the gateway to transcription*. Current Opinion in Cell Biology, 20(3) :253–259 (2008). doi :10.1016/j.ceb.2008.03.003. (Cité page 13.)

Kalendar, R., Vicient, C. M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A., and Schulman, A. H. *Large retrotransposon derivatives : abundant, conserved but nonautonomous retroelements of barley and related genomes*. Genetics, 166(3) :1437–1450 (2004). (Cité page 44.)

Kapitonov, V. V. and Jurka, J. *Helitrons on a roll : eukaryotic rolling-circle transposons*. Trends in genetics : TIG, 23(10) :521–529 (2007). doi :10.1016/j.tig.2007.08.004. (Cité page 45.)

Kapusta, A., Matsuda, A., Marmignon, A., Ku, M., Silve, A., Meyer, E., Forney, J. D., Malinsky, S., and Bétermier, M. *Highly Precise and Developmentally Programmed Genome Assembly in Paramecium Requires Ligase IV–Dependent End Joining*. PLoS Genet, 7(4) :e1002049 (2011). doi :10.1371/journal.pgen.1002049. (Cité pages 71 et 79.)

Karamysheva, Z., Wang, L., Shrode, T., Bednenko, J., Hurley, L. A., and Shippen, D. E. *Developmentally programmed gene elimination in Euplotes crassus facilitates a switch in the telomerase catalytic subunit*. Cell, 113(5) :565–576 (2003). (Cité page 179.)

Kellis, M., Birren, B. W., and Lander, E. S. *Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae*. Nature, 428(6983) :617–624 (2004). doi : 10.1038/nature02424. (Cité page 89.)

Kennedy, G. C., German, M. S., and Rutter, W. J. *The minisatellite in the diabetes susceptibility locus IDDM2 regulates insulin transcription*. Nature Genetics, 9(3) :293–298 (1995). doi : 10.1038/ng0395-293. (Cité page 12.)

Kim, D., Langmead, B., and Salzberg, S. L. *HISAT : a fast spliced aligner with low memory requirements*. Nature Methods, 12(4) :357–360 (2015). doi :10.1038/nmeth.3317. (Cité page 38.)

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. *TopHat2 : accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions*. Genome Biology, 14(4) :R36 (2013). doi :10.1186/gb-2013-14-4-r36. (Cité page 38.)

Klobutcher, L. A. and Herrick, G. *Consensus inverted terminal repeat sequence of Paramecium IESs : resemblance to termini of Tc1-related and Euplotes Tec transposons*. Nucleic Acids Research, 23(11) :2006–2013 (1995). (Cité pages 69 et 106.)

Klobutcher, L. A. and Herrick, G. *Developmental genome reorganization in ciliated protozoa : the transposon link*. Progress in Nucleic Acid Research and Molecular Biology, 56 :1–62 (1997). (Cité pages 69 et 70.)

Koizumi, S. and Kobayashi, S. *Microinjection of plasmid DNA encoding the A surface antigen of Paramecium tetraurelia restores the ability to regenerate a wild-type macronucleus*. Molecular and Cellular Biology, 9(10) :4398–4401 (1989). doi :10.1128/mcb.9.10.4398. (Cité page 75.)

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. *Assembly of long, error-prone reads using repeat graphs*. Nature Biotechnology, 37(5) :540–546 (2019). doi :10.1038/s41587-019-0072-8. (Cité page 27.)

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. *Canu : scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation*. Genome Research, 27(5) :722–736 (2017). doi :10.1101/gr.215087.116. (Cité page 27.)

Korf, I. *Gene finding in novel genomes*. BMC bioinformatics, 5 :59 (2004). doi :10.1186/1471-2105-5-59. (Cité page 34.)

Kornblihtt, A. R., Schor, I. E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M. J. *Alternative splicing : a pivotal step between eukaryotic transcription and translation*. Nature Reviews. Molecular Cell Biology, 14(3) :153–165 (2013). doi :10.1038/nrm3525. (Cité page 15.)

Kovalenko, T. F. and Patrushev, L. I. *Pseudogenes as Functionally Significant Elements of the Genome*. Biochemistry. Biokhimiia, 83(11) :1332–1349 (2018). doi :10.1134/S0006297918110044. (Cité page 15.)

Kozomara, A. and Griffiths-Jones, S. *miRBase : annotating high confidence microRNAs using deep sequencing data*. Nucleic Acids Research, 42(Database issue) :D68–73 (2014). doi : 10.1093/nar/gkt1181. (Cité pages 35 et 36.)

Kreplak, J., Madoui, M.-A., Cápal, P., Novák, P., Labadie, K., Aubert, G., Bayer, P. E., Gali, K. K., Syme, R. A., Main, D., Klein, A., Bérard, A., Vrbová, I., Fournier, C., d'Agata, L., Belser, C., Berrabah, W., Toegelová, H., Milec, Z., Vrána, J., Lee, H., Kougbeadjo, A., Térézol, M., Huneau, C., Turo, C. J., Mohellibi, N., Neumann, P., Falque, M., Gallardo, K., McGee, R., Tar'an, B., Bendahmane, A., Aury, J.-M., Batley, J., Le Paslier, M.-C., Ellis, N., Warkentin, T. D., Coyne, C. J., Salse, J., Edwards, D., Lichtenzveig, J., Macas, J., Doležel, J., Wincker, P., and Burstin, J. *A reference genome for pea provides insight into legume genome evolution*. Nature Genetics, 51(9) :1411–1422 (2019). doi :10.1038/s41588-019-0480-1. (Cité pages 5 et 33.)

Kumazaki, T., Hori, H., Osawa, S., Mita, T., and Higashinakagawa, T. *The nucleotide sequences of 5s rRNAs from three ciliated protozoa*. Nucleic Acids Research, 10(14) :4409–4412 (1982). doi :10.1093/nar/10.14.4409. (Cité page 167.)

Kurtz, S., Narechania, A., Stein, J. C., and Ware, D. *A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes*. BMC genomics, 9 :517 (2008). doi :10.1186/1471-2164-9-517. (Cité page 49.)

Kurtz, S. and Schleiermacher, C. *REPuter : fast computation of maximal repeats in complete genomes*. Bioinformatics (Oxford, England), 15(5) :426–427 (1999). doi :10.1093/bioinformatics/15.5.426. (Cité page 49.)

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G.,

Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and International Human Genome Sequencing Consortium. *Initial sequencing and analysis of the human genome*. Nature, 409(6822) :860–921 (2001). doi : 10.1038/35057062. (Cité pages 16, 33 et 42.)

Lanzoni, O., Fokin, S. I., Lebedeva, N., Migunova, A., Petroni, G., and Potekhin, A. *Rare Freshwater Ciliate Paramecium chlorelligerum Kahl, 1935 and Its Macronuclear Symbiotic Bacterium "Candidatus Holospora parva"*. PLoS ONE, 11(12) (2016). doi :10.1371/journal.pone.0167928. (Cité page 166.)

Lara, E. and Acosta-Mercado, D. *A molecular perspective on ciliates as soil bioindicators*. European Journal of Soil Biology, 49 :107 – 111 (2012). doi :https://doi.org/10.1016/j.ejsobi.2011.11.001. (Cité page 53.)

Lasda, E. L. and Blumenthal, T. *Trans-splicing*. Wiley interdisciplinary reviews. RNA, 2(3) :417–434 (2011). doi :10.1002/wrna.71. (Cité page 15.)

Le Mouël, A., Butler, A., Caron, F., and Meyer, E. *Developmentally regulated chromosome fragmentation linked to imprecise elimination of repeated sequences in paramecia*. Eukaryotic Cell, 2(5) :1076–1090 (2003). doi :10.1128/ec.2.5.1076-1090.2003. (Cité pages 67 et 74.)

Lepere, G., Betermier, M., Meyer, E., and Duharcourt, S. *Maternal noncoding transcripts antagonize the targeting of DNA elimination by scanRNAs in Paramecium tetraurelia*. Genes & Development, 22(11) :1501–1512 (2008). doi :10.1101/gad.473008. (Cité page 77.)

Lepère, G., Nowacki, M., Serrano, V., Gout, J.-F., Guglielmi, G., Duharcourt, S., and Meyer, E. *Silencing-associated and meiosis-specific small RNA pathways in Paramecium tetraurelia*. Nucleic Acids Res, 37(3) :903–15 (2009). doi :10.1093/nar/gkn1018. (Cité page 77.)

Lerat, E. *Identifying repeats and transposable elements in sequenced genomes : how to find your way through the dense forest of programs*. Heredity, 104(6) :520–533 (2010). doi :10.1038/hdy.2009.165. (Cité pages 46 et 49.)

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., and Webb, W. W. *Zero-mode waveguides for single-molecule analysis at high concentrations*. Science (New York, N.Y.), 299(5607) :682–686 (2003). doi :10.1126/science.1079700. (Cité page 23.)

Lhuillier-Akakpo, M., Frapporti, A., Denby Wilkes, C., Matelot, M., Vervoort, M., Sperling, L., and Duharcourt, S. *Local effect of enhancer of zeste-like reveals cooperation of epigenetic and*

cis-acting determinants for zygotic genome rearrangements. PLoS genetics, 10(9) :e1004665 (2014). doi :10.1371/journal.pgen.1004665. (Cité pages 77, 79, 80, 129 et 176.)

Lhuillier-Akakpo, M., Guérin, F., Frapporti, A., and Duharcourt, S. *DNA deletion as a mechanism for developmentally programmed centromere loss*. Nucleic Acids Research, 44(4) :1553–1565 (2016). doi :10.1093/nar/gkv1110. (Cité pages 63, 64, 135, 162 et 171.)

Li, H. *Minimap and miniasm : fast mapping and de novo assembly for noisy long sequences*. Bioinformatics (Oxford, England), 32(14) :2103–2110 (2016). doi :10.1093/bioinformatics/btw152. (Cité page 27.)

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. *The Sequence Alignment/Map format and SAMtools*. Bioinformatics (Oxford, England), 25(16) :2078–2079 (2009). doi :10.1093/bioinformatics/btp352. (Cité page 40.)

Li, R., Ye, J., Li, S., Wang, J., Han, Y., Ye, C., Wang, J., Yang, H., Yu, J., Wong, G. K.-S., and Wang, J. *ReAS : Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun*. PLoS computational biology, 1(4) :e43 (2005). doi :10.1371/journal.pcbi.0010043. (Cité page 49.)

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., and Wang, J. *De novo assembly of human genomes with massively parallel short read sequencing*. Genome Research, 20(2) :265–272 (2010). doi : 10.1101/gr.097261.109. (Cité pages 26 et 27.)

Lin, C.-Y. G., Lin, I.-T., and Yao, M.-C. *Programmed Minichromosome Elimination as a Mechanism for Somatic Genome Reduction in Tetrahymena thermophila*. PLoS genetics, 12(11) :e1006403 (2016). doi :10.1371/journal.pgen.1006403. (Cité page 179.)

Lin, M. F., Jungreis, I., and Kellis, M. *PhyloCSF : a comparative genomics method to distinguish protein coding and non-coding regions*. Bioinformatics (Oxford, England), 27(13) :i275–282 (2011). doi :10.1093/bioinformatics/btr209. (Cité page 35.)

Liu, Q., Mackey, A. J., Roos, D. S., and Pereira, F. C. N. *Evigan : a hidden variable model for integrating gene evidence for eukaryotic gene prediction*. Bioinformatics (Oxford, England), 24(5) :597–605 (2008). doi :10.1093/bioinformatics/btn004. (Cité page 39.)

Liu, Y., Taverna, S. D., Muratore, T. L., Shabanowitz, J., Hunt, D. F., and Allis, C. D. *RNAi-dependent H3K27 methylation is required for heterochromatin formation and DNA elimination in Tetrahymena*. Genes Dev, 21(12) :1530–45 (2007). doi :10.1101/gad.1544207. (Cité page 77.)

Love, M. I., Huber, W., and Anders, S. *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2*. Genome Biology, 15(12) :550 (2014). doi : 10.1186/s13059-014-0550-8. (Cité page 38.)

Lowe, T. M. and Eddy, S. R. *tRNAscan-SE : a program for improved detection of transfer RNA genes in genomic sequence*. Nucleic Acids Research, 25(5) :955–964 (1997). doi : 10.1093/nar/25.5.955. (Cité page 36.)

Lucier, J.-F., Perreault, J., Noël, J.-F., Boire, G., and Perreault, J.-P. *RTAnalyzer : a web application for finding new retrotransposons and detecting L1 retrotransposition signatures*. Nucleic Acids Research, 35(Web Server issue) :W269–274 (2007). doi :10.1093/nar/gkm313. (Cité page 50.)

Mahillon, J. and Chandler, M. *Insertion sequences*. Microbiology and molecular biology reviews : MMBR, 62(3) :725–774 (1998). (Cité page 44.)

Majoros, W. H., Pertea, M., and Salzberg, S. L. *TigrScan and GlimmerHMM : two open source ab initio eukaryotic gene-finders*. Bioinformatics (Oxford, England), 20(16) :2878–2879 (2004). doi :10.1093/bioinformatics/bth315. (Cité page 34.)

Maliszewska-Olejniczak, K., Gruchota, J., Gromadka, R., Denby Wilkes, C., Arnaiz, O., Mathy, N., Duharcourt, S., Bétermier, M., and Nowak, J. K. *TFIIS-Dependent Non-coding Transcription Regulates Developmental Genome Rearrangements*. PLoS genetics, 11(7) :e1005383 (2015). doi :10.1371/journal.pgen.1005383. (Cité pages xii, 77, 79 et 80.)

Maquat, L. E. *Nonsense-mediated mRNA decay splicing, translation and mRNP dynamics*. Nature Reviews Molecular Cell Biology, 5(2) :89–99 (2004). doi :10.1038/nrm1310. (Cité page 73.)

Marçais, G. and Kingsford, C. *A fast, lock-free approach for efficient parallel counting of occurrences of k-mers*. Bioinformatics (Oxford, England), 27(6) :764–770 (2011). doi : 10.1093/bioinformatics/btr011. (Cité page 49.)

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. *Genome sequencing in microfabricated high-density picolitre reactors*. Nature, 437(7057) :376–380 (2005). doi :10.1038/nature03959. (Cité page 21.)

Marker, S., Carradec, Q., Tanty, V., Arnaiz, O., and Meyer, E. *A forward genetic screen reveals essential and non-essential RNAi factors in Paramecium tetraurelia*. Nucleic Acids Research, 42(11) :7268–7280 (2014). doi :10.1093/nar/gku223. (Cité pages xi et 63.)

Marmignon, A., Bischerour, J., Silve, A., Fojcik, C., Dubois, E., Arnaiz, O., Kapusta, A., Malinsky, S., and Bétermier, M. *Ku-mediated coupling of DNA cleavage and repair during programmed genome rearrangements in the ciliate Paramecium tetraurelia*. PLoS genetics, 10(8) :e1004552 (2014). doi :10.1371/journal.pgen.1004552. (Cité pages xiii, 71, 79 et 178.)

Matera, A. G. and Wang, Z. *A day in the life of the spliceosome*. Nature Reviews. Molecular Cell Biology, 15(2) :108–121 (2014). doi :10.1038/nrm3742. (Cité pages 13 et 15.)

Maumus, F. and Quesneville, H. *Deep Investigation of Arabidopsis thaliana Junk DNA Reveals a Continuum between Repetitive Elements and Genomic Dark Matter*. PLoS ONE, 9(4) (2014). doi :10.1371/journal.pone.0094101. (Cité page 179.)

Maxam, A. M. and Gilbert, W. *A new method for sequencing DNA*. Proceedings of the National Academy of Sciences of the United States of America, 74(2) :560–564 (1977). doi :10.1073/pnas.74.2.560. (Cité page 20.)

Mayer, K. M. and Forney, J. D. *A Mutation in the Flanking 5-TA-3 Dinucleotide Prevents Excision of an Internal Eliminated Sequence From the Paramecium tetraurelia Genome*. Genetics, 151(2) :597–604 (1999). (Cité page 68.)

McClintock, B. *The origin and behavior of mutable loci in maize*. Proceedings of the National Academy of Sciences of the United States of America, 36(6) :344–355 (1950). doi :10.1073/pnas.36.6.344. (Cité pages 12 et 42.)

McClure, M. A., Richardson, H. S., Clinton, R. A., Hepp, C. M., Crowther, B. A., and Donaldson, E. F. *Automated characterization of potentially active retroid agents in the human genome*. Genomics, 85(4) :512–523 (2005). doi :10.1016/j.ygeno.2004.12.006. (Cité page 33.)

McCormick-Graham, M. and Romero, D. P. *A single telomerase RNA is sufficient for the synthesis of variable telomeric DNA repeats in ciliates of the genus Paramecium*. Molecular and Cellular Biology, 16(4) :1871–1879 (1996). doi :10.1128/mcb.16.4.1871. (Cité pages 74 et 167.)

McGrath, C. L., Gout, J.-F., Doak, T. G., Yanagi, A., and Lynch, M. *Insights into three whole-genome duplications gleaned from the Paramecium caudatum genome sequence*. Genetics, 197(4) :1417–1428 (2014a). doi :10.1534/genetics.114.163287. (Cité pages 56, 85, 89, 160 et 165.)

McGrath, C. L., Gout, J.-F., Johri, P., Doak, T. G., and Lynch, M. *Differential retention and divergent resolution of duplicate genes following whole-genome duplication*. Genome Research, 24(10) :1665–1675 (2014b). doi :10.1101/gr.173740.114. (Cité pages 56, 85, 89, 160 et 165.)

Metzker, M. L. *Sequencing technologies - the next generation*. Nature Reviews. Genetics, 11(1) :31–46 (2010). doi :10.1038/nrg2626. (Cité page 22.)

Meyer, E. *Induction of specific macronuclear developmental mutations by microinjection of a cloned telomeric gene in Paramecium primaurelia.* Genes & Development, 6(2) :211–222 (1992). doi :10.1101/gad.6.2.211. (Cité page 76.)

Miller, W. J. and Capy, P. *Applying mobile genetic elements for genome analysis and evolution*. Molecular Biotechnology, 33(2) :161–174 (2006). doi :10.1385/MB:33:2:161. (Cité page 42.)

Mitchison, H. M. and Valente, E. M. *Motile and non-motile cilia in human pathology : from function to phenotypes*. The Journal of Pathology, 241(2) :294–309 (2017). doi :10.1002/path.4843. (Cité page 53.)

Mochizuki, K., Fine, N. A., Fujisawa, T., and Gorovsky, M. A. *Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena*. Cell, 110(6) :689–99 (2002). (Cité page 77.)

Mudge, J. M. and Harrow, J. *The state of play in higher eukaryote gene annotation*. Nature Reviews. Genetics, 17(12) :758–772 (2016). doi :10.1038/nrg.2016.119. (Cité pages 33 et 35.)

Mungall, C. J., Emmert, D. B., and FlyBase Consortium. *A Chado case study : an ontology-based modular schema for representing genome-associated biological information*. Bioinformatics (Oxford, England), 23(13) :i337–346 (2007). doi :10.1093/bioinformatics/btm189. (Cité page 40.)

Munoz-Torres, M. C., Reese, J. T., Childers, C. P., Bennett, A. K., Sundaram, J. P., Childs, K. L., Anzola, J. M., Milshina, N., and Elsik, C. G. *Hymenoptera Genome Database : integrated community resources for insect species of the order Hymenoptera*. Nucleic Acids Research, 39(Database issue) :D658–662 (2011). doi :10.1093/nar/gkq1145. (Cité page 41.)

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., Anson, E. L., Bolanos, R. A., Chou, H. H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., Dunn, P. J., Lai, Z., Liang, Y., Nusskern, D. R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G. M., Adams, M. D., and Venter, J. C. *A whole-genome assembly of Drosophila*. Science (New York, N.Y.), 287(5461) :2196–2204 (2000). (Cité page 26.)

Nagarajan, N. and Pop, M. *Sequence assembly demystified*. Nature Reviews. Genetics, 14(3) :157–167 (2013). doi :10.1038/nrg3367. (Cité page 26.)

Nanney, D. L. *Mating Type Determination in Paramecium Aurelia, a Model of Nucleo-Cytoplasmic Interaction*. Proceedings of the National Academy of Sciences of the United States of America, 39(2) :113 (1953). (Cité page 61.)

Narzisi, G. and Mishra, B. *Comparing de novo genome assembly : the long and short of it*. PloS One, 6(4) :e19175 (2011). doi :10.1371/journal.pone.0019175. (Cité page 28.)

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., and Finn, R. D. *Rfam 12.0 : updates to the RNA families database*. Nucleic Acids Research, 43(Database issue) :D130–137 (2015). doi :10.1093/nar/gku1063. (Cité pages 35 et 36.)

Nawrocki, E. P. and Eddy, S. R. *Infernal 1.1 : 100-fold faster RNA homology searches*. Bioinformatics (Oxford, England), 29(22) :2933–2935 (2013). doi :10.1093/bioinformatics/btt509. (Cité page 36.)

Nekrasova, I., Nikitashina, V., Bhullar, S., Arnaiz, O., Singh, D. P., Meyer, E., and Potekhin, A. *Loss of a Fragile Chromosome Region leads to the Screwy Phenotype in Paramecium tetraurelia*. Genes, 10(7) (2019). doi :10.3390/genes10070513. (Cité page xi.)

Novák, P., Neumann, P., Pech, J., Steinhaisl, J., and Macas, J. *RepeatExplorer : a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads*. Bioinformatics, 29(6) :792–793 (2013). doi :10.1093/bioinformatics/btt054. (Cité page 49.)

Novák, P., Ávila Robledillo, L., Koblížková, A., Vrbová, I., Neumann, P., and Macas, J. *TAREAN : a computational tool for identification and characterization of satellite DNA from unassembled short reads*. Nucleic Acids Research, 45(12) :e111 (2017). doi :10.1093/nar/gkx257. (Cité page 33.)

Nowacki, M., Haye, J. E., Fang, W., Vijayan, V., and Landweber, L. F. *RNA-mediated epigenetic regulation of DNA copy number*. Proc Natl Acad Sci U S A, 107(51) :22140–4 (2010). doi :10.1073/pnas.1012236107. (Cité page 65.)

Nowacki, M., Higgins, B. P., Maquilan, G. M., Swart, E. C., Doak, T. G., and Landweber, L. F. *A functional role for transposases in a large eukaryotic genome*. Science, 324(5929) :935–8 (2009). doi :10.1126/science.1170023. (Cité page xi.)

Nowacki, M., Vijayan, V., Zhou, Y., Schotanus, K., Doak, T. G., and Landweber, L. F. *RNA-mediated epigenetic programming of a genome-rearrangement pathway*. Nature, 451(7175) :153–8 (2008). doi :10.1038/nature06452. (Cité page 66.)

Nowacki, M., Zagorski-Ostoja, W., and Meyer, E. *Nowa1p and Nowa2p : Novel Putative RNA Binding Proteins Involved in trans-Nuclear Crosstalk in Paramecium tetraurelia*. Current

Biology, 15(18) :1616–1628 (2005). doi :10.1016/j.cub.2005.07.033. (Cité pages 76, 77, 79 et 80.)

Nowak, J. K., Gromadka, R., Juszczuk, M., Jerka-Dziadosz, M., Maliszewska, K., Mucchielli, M.-H., Gout, J.-F., Arnaiz, O., Agier, N., Tang, T., Aggerbeck, L. P., Cohen, J., Delacroix, H., Sperling, L., Herbert, C. J., Zagulski, M., and Bétermier, M. *Functional study of genes essential for autogamy and nuclear reorganization in Paramecium*. Eukaryotic Cell, 10(3) :363–372 (2011). doi :10.1128/EC.00258-10. (Cité page 77.)

O'Connor, B. D., Day, A., Cain, S., Arnaiz, O., Sperling, L., and Stein, L. D. *GMODWeb : a web framework for the Generic Model Organism Database*. Genome Biology, 9(6) :R102 (2008). doi :10.1186/gb-2008-9-6-r102. (Cité page ix.)

Ohno, S. *So much "junk" DNA in our genome*. Brookhaven Symposia in Biology, 23 :366–370 (1972). (Cité page 15.)

Orgeur, M., Martens, M., Börno, S. T., Timmermann, B., Duprez, D., and Stricker, S. *A dual transcript-discovery approach to improve the delimitation of gene features from RNA-seq data in the chicken model*. Biology Open, 7(1) (2018). doi :10.1242/bio.028498. (Cité page 38.)

Osawa, S., Jukes, T. H., Watanabe, K., and Muto, A. *Recent evidence for evolution of the genetic code*. Microbiological Reviews, 56(1) :229–264 (1992). (Cité page 15.)

Parra, G., Blanco, E., and Guigó, R. *GeneID in Drosophila*. Genome Research, 10(4) :511–515 (2000). (Cité page 34.)

Pavlicek, A., Gentles, A. J., Paces, J., Paces, V., and Jurka, J. *Retroposition of processed pseudogenes : the impact of RNA stability and translational control*. Trends in genetics : TIG, 22(2) :69–73 (2006). doi :10.1016/j.tig.2005.11.005. (Cité page 15.)

Pellicer, J., FAY, M. F., and LEITCH, I. J. *The largest eukaryotic genome of them all ?* Botanical Journal of the Linnean Society, 164(1) :10–15 (2010). doi :10.1111/j.1095-8339.2010.01072.x. (Cité page 5.)

Pevsner, J. *Bioinformatics and Functional Genomics*. WILEY Blackwell, 3 edition (2015). (Cité page 44.)

Pevzner, P. A., Tang, H., and Waterman, M. S. *An Eulerian path approach to DNA fragment assembly*. Proceedings of the National Academy of Sciences of the United States of America, 98(17) :9748–9753 (2001). doi :10.1073/pnas.171285098. (Cité page 27.)

Piégu, B., Bire, S., Arensburger, P., and Bigot, Y. *A survey of transposable element classification systems–a call for a fundamental update to meet the challenge of their diversity and complexity*. Molecular Phylogenetics and Evolution, 86 :90–109 (2015). doi :10.1016/j.ympev.2015.03.009. (Cité page 42.)

Pinskaya, M., Saci, Z., Gallopin, M., Gabriel, M., Nguyen, H. T., Firlej, V., Descrimes, M., Rapinat, A., Gentien, D., Taille, A. d. l., Londoño-Vallejo, A., Allory, Y., Gautheret, D., and Morillon, A. *Reference-free transcriptome exploration reveals novel RNAs for prostate cancer diagnosis*. Life Science Alliance, 2(6) (2019). doi :10.26508/lsa.201900449. (Cité page 39.)

Platt, R. N., Vandewege, M. W., and Ray, D. A. *Mammalian transposable elements and their impacts on genome evolution*. Chromosome Research : An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology, 26(1-2) :25–43 (2018). doi :10.1007/s10577-017-9570-z. (Cité page 12.)

Plohl, M., Meštrović, N., and Mravinac, B. *Centromere identity from the DNA point of view*. Chromosoma, 123(4) :313–325 (2014). doi :10.1007/s00412-014-0462-0. (Cité page 7.)

Preer, J. R., Preer, L. B., Rudman, B. M., and Barnett, A. J. *Deviation from the universal code shown by the gene for surface protein 51a in Paramecium*. Nature, 314(6007) :188–190 (1985). doi :10.1038/314188a0. (Cité page 15.)

Preer, L. B., Hamilton, G., and Preer, J. R. *Micronuclear DNA from Paramecium tetraurelia : serotype 51 A gene has internally eliminated sequences*. The Journal of Protozoology, 39(6) :678–682 (1992). (Cité page 68.)

Price, A. L., Jones, N. C., and Pevzner, P. A. *De novo identification of repeat families in large genomes*. Bioinformatics (Oxford, England), 21 Suppl 1 :i351–358 (2005). doi : 10.1093/bioinformatics/bti1018. (Cité page 33.)

Pritchard, A. E., Sable, C. L., Venuti, S. E., and Cummings, D. J. *Analysis of NADH dehydrogenase proteins, ATPase subunit 9, cytochrome b, and ribosomal protein L14 encoded in the mitochondrial DNA of Paramecium*. Nucleic Acids Research, 18(1) :163–171 (1990). doi :10.1093/nar/18.1.163. (Cité page 15.)

Quadrana, L., Bortolini Silveira, A., Mayhew, G. F., LeBlanc, C., Martienssen, R. A., Jeddeloh, J. A., and Colot, V. *The Arabidopsis thaliana mobilome and its impact at the species level*. eLife, 5 (2016). doi :10.7554/eLife.15716. (Cité page 50.)

Quax, T. E. F., Claassens, N. J., Söll, D., and van der Oost, J. *Codon Bias as a Means to Fine-Tune Gene Expression*. Molecular Cell, 59(2) :149–161 (2015). doi :10.1016/j.molcel.2015.05.035. (Cité page 13.)

Quesneville, H., Bergman, C. M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., and Anxolabehere, D. *Combined evidence annotation of transposable elements in genome sequences*. PLoS computational biology, 1(2) :166–175 (2005). doi :10.1371/journal.pcbi.0010022. (Cité pages 46 et 49.)

Quesneville, H., Nouaud, D., and Anxolabéhère, D. *Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes*. Journal of Molecular Evolution, 57 Suppl 1 :S50–59 (2003). doi :10.1007/s00239-003-0007-2. (Cité page 48.)

Richard, G.-F., Kerrest, A., and Dujon, B. *Comparative genomics and molecular dynamics of DNA repeats in eukaryotes*. Microbiology and molecular biology reviews : MMBR, 72(4) :686–727 (2008). doi :10.1128/MMBR.00011-08. (Cité pages 10 et 11.)

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. *Integrative genomics viewer*. Nature Biotechnology, 29(1) :24–26 (2011). doi :10.1038/nbt.1754. (Cité page 40.)

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. *edgeR : a Bioconductor package for differential expression analysis of digital gene expression data*. Bioinformatics (Oxford, England), 26(1) :139–140 (2010). doi :10.1093/bioinformatics/btp616. (Cité page 38.)

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., and Nyrén, P. *Real-time DNA sequencing using detection of pyrophosphate release*. Analytical Biochemistry, 242(1) :84–89 (1996). doi :10.1006/abio.1996.0432. (Cité page 20.)

Ronaghi, M., Uhlén, M., and Nyrén, P. *A sequencing method based on real-time pyrophosphate*. Science (New York, N.Y.), 281(5375) :363, 365 (1998). (Cité page 20.)

Ruan, J. and Li, H. *Fast and accurate long-read assembly with wtdbg2*. Nature Methods, 17(2) :155–158 (2020). doi :10.1038/s41592-019-0669-3. (Cité page 27.)

Ruehle, M. D., Orias, E., and Pearson, C. G. *Tetrahymena as a Unicellular Model Eukaryote : Genetic and Genomic Tools*. Genetics, 203(2) :649 (2016). doi :10.1534/genetics.114.169748. (Cité page 165.)

Ruggiero, M. A., Gordon, D. P., Orrell, T. M., Bailly, N., Bourgoin, T., Brusca, R. C., Cavalier-Smith, T., Guiry, M. D., and Kirk, P. M. *A higher level classification of all living organisms*. PloS One, 10(4) :e0119248 (2015). doi :10.1371/journal.pone.0119248. (Cité page 4.)

Ruiz, F., Krzywicka, A., Klotz, C., Keller, A., Cohen, J., Koll, F., Balavoine, G., and Beisson, J. *The SM19 gene, required for duplication of basal bodies in Paramecium, encodes a novel tubulin, eta-tubulin*. Current biology : CB, 10(22) :1451–1454 (2000). (Cité page 68.)

Ruiz, F., Vayssié, L., Klotz, C., Sperling, L., and Madeddu, L. *Homology-dependent gene silencing in Paramecium*. Molecular biology of the cell, 9(4) :931–943 (1998). (Cité page 63.)

Russell, J. and Zomerdijk, J. C. B. M. *The RNA polymerase I transcription machinery*. Biochemical Society Symposium, (73) :203–216 (2006). (Cité page 36.)

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., and Barrell, B. *Artemis : sequence visualization and annotation*. Bioinformatics (Oxford, England), 16(10) :944–945 (2000). doi :10.1093/bioinformatics/16.10.944. (Cité page 40.)

Salzberg, S. L. *Next-generation genome annotation : we still struggle to get it right*. Genome Biology, 20(1) :92 (2019). doi :10.1186/s13059-019-1715-2. (Cité pages 31 et 36.)

Sandoval, P. Y., Swart, E. C., Arambasic, M., and Nowacki, M. *Functional diversification of Dicer-like proteins and small RNAs required for genome sculpting*. Developmental cell, 28(2) :174–188 (2014). doi :10.1016/j.devcel.2013.12.010. (Cité pages 77, 79, 80 et 176.)

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M., and Smith, M. *Nucleotide sequence of bacteriophage phi X174 DNA*. Nature, 265(5596) :687–695 (1977). (Cité pages 16, 17 et 20.)

Santos-Rosa, H., Schneider, R., Bannister, A. J., Sherriff, J., Bernstein, B. E., Emre, N. C. T., Schreiber, S. L., Mellor, J., and Kouzarides, T. *Active genes are tri-methylated at K4 of histone H3*. Nature, 419(6905) :407–411 (2002). doi :10.1038/nature01080. (Cité page 8.)

Saudemont, B., Popa, A., Parmley, J. L., Rocher, V., Blugeon, C., Necsulea, A., Meyer, E., and Duret, L. *The fitness cost of mis-splicing is the main determinant of alternative splicing patterns*. Genome Biology, 18(1) :208 (2017). doi :10.1186/s13059-017-1344-6. (Cité pages 73 et 83.)

Scarano, E., Iaccarino, M., Grippo, P., and Parisi, E. *The heterogeneity of thymine methyl group origin in DNA pyrimidine isostichs of developing sea urchin embryos*. Proceedings of the National Academy of Sciences of the United States of America, 57(5) :1394–1400 (1967). doi :10.1073/pnas.57.5.1394. (Cité page 82.)

Schadt, E. E., Turner, S., and Kasarskis, A. *A window into third-generation sequencing*. Human Molecular Genetics, 19(R2) :R227–240 (2010). doi :10.1093/hmg/ddq416. (Cité page 17.)

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Van Buren, P., Vaughn, M. W., Ying, K., Yeh, C.-T., Emrich, S. J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W. B., Baucom, R. S., Brutnell, T. P., Carpita, N. C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J. C., Fu, Y., Jeddeloh, J. A., Han, Y., Lee, H., Li, P., Lisch, D. R., Liu, S., Liu, Z., Nagel, D. H., McCann, M. C., SanMiguel, P., Myers, A. M., Nettleton, D., Nguyen, J., Penning, B. W., Ponnala, L., Schneider, K. L., Schwartz, D. C., Sharma, A., Soderlund, C., Springer, N. M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T. K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J. L., Dawe, R. K., Jiang, J., Jiang, N., Presting, G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R., Wing, R. A., and Wilson, R. K. *The B73 maize genome : complexity, diversity, and dynamics*. Science (New York, N.Y.), 326(5956) :1112–1115 (2009). doi :10.1126/science.1178534. (Cité page 42.)

Schones, D. E. and Zhao, K. *Genome-wide approaches to studying chromatin modifications*. Nature Reviews. Genetics, 9(3) :179–191 (2008). doi :10.1038/nrg2270. (Cité pages 8 et 9.)

Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. *Oases : robust de novo RNA-seq assembly across the dynamic range of expression levels*. Bioinformatics (Oxford, England), 28(8) :1086–1092 (2012). doi :10.1093/bioinformatics/bts094. (Cité page 38.)

Shafee, R., T ; Lowe. *Eukaryotic and prokaryotic gene structure*. WikiJournal of Medicine (2017). doi :10.15347/wjm/2017.002. (Cité pages 13 et 14.)

Shi, L., Koll, F., Arnaiz, O., and Cohen, J. *The Ciliary Protein IFT57 in the Macronucleus of Paramecium*. The Journal of Eukaryotic Microbiology (2017). doi :10.1111/jeu.12423. (Cité page ix.)

Sieber, P., Platzer, M., and Schuster, S. *The Definition of Open Reading Frame Revisited*. Trends in genetics : TIG, 34(3) :167–170 (2018). doi :10.1016/j.tig.2017.12.009. (Cité page 13.)

Simpson, J. T. and Durbin, R. *Efficient de novo assembly of large genomes using compressed data structures*. Genome Research, 22(3) :549–556 (2012). doi :10.1101/gr.126953.111. (Cité page 27.)

Singh, A., Vancura, A., Woycicki, R. K., Hogan, D. J., Hendrick, A. G., and Nowacki, M. *Determination of the presence of 5-methylcytosine in Paramecium tetraurelia*. PloS One, 13(10) :e0206667 (2018). doi :10.1371/journal.pone.0206667. (Cité page 82.)

Singh, D. P., Saudemont, B., Guglielmi, G., Arnaiz, O., Goût, J.-F., Prajer, M., Potekhin, A., Przybòs, E., Aubusson-Fleury, A., Bhullar, S., Bouhouche, K., Lhuillier-Akakpo, M.,

Tanty, V., Blugeon, C., Alberti, A., Labadie, K., Aury, J.-M., Sperling, L., Duharcourt, S., and Meyer, E. *Genome-defence small RNAs exapted for epigenetic mating-type inheritance*. Nature, 509(7501) :447–452 (2014). doi :10.1038/nature13318. (Cité pages xii, 61, 62, 77 et 179.)

Skipper, K. A., Andersen, P. R., Sharma, N., and Mikkelsen, J. G. *DNA transposon-based gene vehicles - scenes from an evolutionary drive*. Journal of Biomedical Science, 20 :92 (2013). doi :10.1186/1423-0127-20-92. (Cité page 45.)

Slater, G. S. C. and Birney, E. *Automated generation of heuristics for biological sequence comparison*. BMC bioinformatics, 6 :31 (2005). doi :10.1186/1471-2105-6-31. (Cité page 36.)

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M. H., Baldock, R., Barbiera, G., Bardou, P., Beck, T., Blake, A., Bonierbale, M., Brookes, A. J., Bucci, G., Buetti, I., Burge, S., Cabau, C., Carlson, J. W., Chelala, C., Chrysostomou, C., Cittaro, D., Collin, O., Cordova, R., Cutts, R. J., Dassi, E., Di Genova, A., Djari, A., Esposito, A., Estrella, H., Eyras, E., Fernandez-Banet, J., Forbes, S., Free, R. C., Fujisawa, T., Gadaleta, E., Garcia-Manteiga, J. M., Goodstein, D., Gray, K., Guerra-Assunção, J. A., Haggarty, B., Han, D.-J., Han, B. W., Harris, T., Harshbarger, J., Hastings, R. K., Hayes, R. D., Hoede, C., Hu, S., Hu, Z.-L., Hutchins, L., Kan, Z., Kawaji, H., Keliet, A., Kerhornou, A., Kim, S., Kinsella, R., Klopp, C., Kong, L., Lawson, D., Lazarevic, D., Lee, J.-H., Letellier, T., Li, C.-Y., Lio, P., Liu, C.-J., Luo, J., Maass, A., Mariette, J., Maurel, T., Merella, S., Mohamed, A. M., Moreews, F., Nabihoudine, I., Ndegwa, N., Noirot, C., Perez-Llamas, C., Primig, M., Quattrone, A., Quesneville, H., Rambaldi, D., Reecy, J., Riba, M., Rosanoff, S., Saddiq, A. A., Salas, E., Sallou, O., Shepherd, R., Simon, R., Sperling, L., Spooner, W., Staines, D. M., Steinbach, D., Stone, K., Stupka, E., Teague, J. W., Dayem Ullah, A. Z., Wang, J., Ware, D., Wong-Erasmus, M., Youens-Clark, K., Zadissa, A., Zhang, S.-J., and Kasprzyk, A. *The BioMart community portal : an innovative alternative to large, centralized data repositories*. Nucleic Acids Research, 43(W1) :W589–598 (2015). doi :10.1093/nar/gkv350. (Cité page x.)

Smit, G., Hubley. *RepeatMasker*. repeatmasker org (1996). (Cité pages 34 et 48.)

Sohn, J.-I. and Nam, J.-W. *The present and future of de novo whole-genome assembly*. Briefings in Bioinformatics, 19(1) :23–40 (2018). doi :10.1093/bib/bbw096. (Cité page 29.)

Solovyev, V. V., Salamov, A. A., and Lawrence, C. B. *Identification of human gene structure using linear discriminant functions and dynamic programming*. Proceedings. International Conference on Intelligent Systems for Molecular Biology, 3 :367–375 (1995). (Cité page 34.)

Sonneborn. *Recent advances in the genetics of Paramecium and Euplotes*. Advances in genetics, 1 :263–358 (1947). (Cité page 61.)

Sonneborn, T. M. *The Paramecium aurelia Complex of Fourteen Sibling Species*. Transactions of the American Microscopical Society, 94(2) :155–178 (1975). (Cité page 56.)

Sperling, L. *Remembrance of things past retrieved from the Paramecium genome*. Research in Microbiology, 162(6) :587–597 (2011). doi :10.1016/j.resmic.2011.02.012. (Cité page 52.)

Staden, R. *A strategy of DNA sequencing employing computer programs*. Nucleic Acids Research, 6(7) :2601–2610 (1979). (Cité page 16.)

Stanke, M., Schöffmann, O., Morgenstern, B., and Waack, S. *Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources*. BMC bioinformatics, 7 :62 (2006). doi :10.1186/1471-2105-7-62. (Cité pages 34 et 39.)

Stanke, M. and Waack, S. *Gene prediction with a hidden Markov model and a new intron submodel*. Bioinformatics (Oxford, England), 19 Suppl 2 :ii215–225 (2003). (Cité page 34.)

Steele, C. J., Barkocy-Gallagher, G. A., Preer, L. B., and Preer, J. R. *Developmentally excised sequences in micronuclear DNA of Paramecium*. Proceedings of the National Academy of Sciences, 91(6) :2255–2259 (1994). (Cité page 68.)

Stein, L. D. *Using GBrowse 2.0 to visualize and share next-generation sequence data*. Briefings in Bioinformatics, 14(2) :162–171 (2013). doi :10.1093/bib/bbt001. (Cité page 40.)

Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A., and Lewis, S. *The generic genome browser : a building block for a model organism system database*. Genome Research, 12(10) :1599–1610 (2002). doi : 10.1101/gr.403602. (Cité page 40.)

Strahl, B. D. and Allis, C. D. *The language of covalent histone modifications*. Nature, 403(6765) :41–45 (2000). doi :10.1038/47412. (Cité page 8.)

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M. J. P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H. Y. K., Mu, X. J., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J. M., Kong, Y., Lameijer, E.-W., McCarthy, S., Flicek, P., Gibbs, R. A., Marth, G., Mason, C. E., Menelaou, A., Muzny, D. M., Nelson, B. J., Noor, A., Parrish, N. F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E. E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A. A., Untergasser, A., Walker, J. A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M. A., McCarroll, S. A., 1000 Genomes Project Consortium, Mills, R. E., Gerstein, M. B., Bashir, A., Stegle, O., Devine, S. E., Lee, C., Eichler, E. E., and Korbel, J. O. *An integrated map of structural variation in 2,504 human genomes*. Nature, 526(7571) :75–81 (2015). doi :10.1038/nature15394. (Cité page 18.)

Sutton, G. G., WHITE, O., ADAMS, M. D., and KERLAVAGE, A. R. *TIGR Assembler : A New Tool for Assembling Large Shotgun Sequencing Projects*. Genome Science and Technology, 1 :9–19 (1995). doi :10.1089/gst.1995.1.9. (Cité page 26.)

Swart, E., Denby Wilkes, C., Sandoval, P., Hoehener, C., Singh, A., Furrer, D., Arambasic, M., Ignarski, M., and Nowacki, M. *Identification and analysis of functional associations among natural eukaryotic genome editing components*. F1000Research, 6(1374) (2017). doi : 10.12688/f1000research.12121.1. (Cité pages 79 et 80.)

Swart, E. C., Bracht, J. R., Magrini, V., Minx, P., Chen, X., Zhou, Y., Khurana, J. S., Goldman, A. D., Nowacki, M., Schotanus, K., Jung, S., Fulton, R. S., Ly, A., McGrath, S., Haub, K., Wiggins, J. L., Storton, D., Matese, J. C., Parsons, L., Chang, W.-J., Bowen, M. S., Stover, N. A., Jones, T. A., Eddy, S. R., Herrick, G. A., Doak, T. G., Wilson, R. K., Mardis, E. R., and Landweber, L. F. *The Oxytricha trifallax macronuclear genome : a complex eukaryotic genome with 16,000 tiny chromosomes*. PLoS Biol, 11(1) :e1001473 (2013). doi :10.1371/journal.pbio.1001473. (Cité pages 65 et 66.)

Swart, E. C., Wilkes, C. D., Sandoval, P. Y., Arambasic, M., Sperling, L., and Nowacki, M. *Genome-wide analysis of genetic and epigenetic control of programmed DNA deletion*. Nucleic Acids Research, 42(14) :8970–8983 (2014). doi :10.1093/nar/gku619. (Cité pages 71, 79 et 80.)

Takahashi, H., Kato, S., Murata, M., and Carninci, P. *CAGE (cap analysis of gene expression) : a protocol for the detection of promoter and transcriptional networks*. Methods in Molecular Biology (Clifton, N.J.), 786 :181–200 (2012). doi :10.1007/978-1-61779-292-2_11. (Cité page 38.)

Tanabe, H. *Paramecium tetraurelia MS2 which expresses non-coding RNAs related to aging, life span and autogamy (sexual reproduction)* (2006). { :itemType : dataset}. (Cité page 167.)

Taverna, S. D., Coyne, R. S., and Allis, C. D. *Methylation of histone h3 at lysine 9 targets programmed DNA elimination in tetrahymena*. Cell, 110(6) :701–11 (2002). (Cité page 77.)

Thibaud-Nissen, Souvorov, Murphy, DiCuccio, and Kitts. *Eukaryotic Genome Annotation Pipeline*. The NCBI Handbook. 2nd edition. Bethesda (2013). (Cité page 40.)

Thomas, J. and Pritham, E. J. *Helitrons, the Eukaryotic Rolling-circle Transposable Elements*. Microbiology Spectrum, 3(4) (2015). doi :10.1128/microbiolspec.MDNA3-0049-2014. (Cité page 45.)

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation*. Nature Biotechnology, 28(5) :511–515 (2010). doi :10.1038/nbt.1621. (Cité page 38.)

Treangen, T. J. and Salzberg, S. L. *Repetitive DNA and next-generation sequencing : computational challenges and solutions*. Nature Reviews. Genetics, 13(1) :36–46 (2011). doi : 10.1038/nrg3117. (Cité pages 33 et 44.)

Tu, Z. *Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, Anopheles gambiae*. Proceedings of the National Academy of Sciences of the United States of America, 98(4) :1699–1704 (2001). doi :10.1073/pnas.041593198. (Cité page 50.)

Tu, Z., Li, S., and Mao, C. *The changing tails of a novel short interspersed element in Aedes aegypti : genomic evidence for slippage retrotransposition and the relationship between 3' tandem repeats and the poly(dA) tail*. Genetics, 168(4) :2037–2047 (2004). doi :10.1534/genetics.104.032045. (Cité page 50.)

Tucker, J. B., Beisson, J., Roche, D. L., and Cohen, J. *Microtubules and control of macronuclear 'amitosis' in Paramecium*. Journal of Cell Science, 44 :135–151 (1980). (Cité page 58.)

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. *The Third Revolution in Sequencing Technology*. Trends in genetics : TIG, 34(9) :666–681 (2018). doi :10.1016/j.tig.2018.05.008. (Cité page 29.)

Vanin, E. F. *Processed pseudogenes : characteristics and evolution*. Annual Review of Genetics, 19 :253–272 (1985). doi :10.1146/annurev.ge.19.120185.001345. (Cité page 15.)

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L.,

Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. *The sequence of the human genome*. Science (New York, N.Y.), 291(5507) :1304–1351 (2001). doi :10.1126/science.1058040. (Cité pages 16 et 31.)

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., and Earl, A. M. *Pilon : an integrated tool for comprehensive microbial variant detection and genome assembly improvement*. PloS One, 9(11) :e112963 (2014). doi :10.1371/journal.pone.0112963. (Cité pages 29 et 107.)

Walker, P. M. *Origin of satellite DNA*. Nature, 229(5283) :306–308 (1971). (Cité page 10.)

Wang, A. and Au, K. F. *Performance difference of graph-based and alignment-based hybrid error correction methods for error-prone long reads*. Genome Biology, 21(1) :14 (2020). doi : 10.1186/s13059-019-1885-y. (Cité pages 27 et 29.)

Wang, J. and Davis, R. E. *Programmed DNA elimination in multicellular organisms*. Current Opinion in Genetics & Development, 27 :26–34 (2014). doi :10.1016/j.gde.2014.03.012. (Cité page 53.)

Wang, Q., Arighi, C. N., King, B. L., Polson, S. W., Vincent, J., Chen, C., Huang, H., Kingham, B. F., Page, S. T., Rendino, M. F., Thomas, W. K., Udwary, D. W., Wu, C. H., and North East Bioinformatics Collaborative Curation Team. *Community annotation and bioinformatics workforce development in concert–Little Skate Genome Annotation Workshops and Jamborees*. Database : The Journal of Biological Databases and Curation, 2012 :bar064 (2012). doi :10.1093/database/bar064. (Cité page 41.)

Wang, Y., Wang, Y., Sheng, Y., Huang, J., Chen, X., Al-Rasheid, K. A. S., and Gao, S. *A comparative study of genome organization and epigenetic mechanisms in model ciliates, with an emphasis on Tetrahymena, Paramecium and Oxytricha*. European Journal of Protistology, 61(Pt B) :376–387 (2017). doi :10.1016/j.ejop.2017.06.006. (Cité pages 55, 64 et 82.)

Wang, Z., Gerstein, M., and Snyder, M. *RNA-Seq : a revolutionary tool for transcriptomics*. Nature Reviews. Genetics, 10(1) :57–63 (2009). doi :10.1038/nrg2484. (Cité page 19.)

Warren, A., Patterson, D. J., Dunthorn, M., Clamp, J. C., Achilles-Day, U. E. M., Aescht, E., Al-Farraj, S. A., Al-Quraishy, S., Al-Rasheid, K., Carr, M., Day, J. G., Dellinger, M., El-Serehy, H. A., Fan, Y., Gao, F., Gao, S., Gong, J., Gupta, R., Hu, X., Kamra, K., Langlois, G., Lin, X., Lipscomb, D., Lobban, C. S., Luporini, P., Lynn, D. H., Ma, H., Macek, M., Mackenzie-Dodds, J., Makhija, S., Mansergh, R. I., Martín-Cereceda, M., McMiller, N., Montagnes, D. J. S., Nikolaeva, S., Ong'ondo, G. O., Pérez-Uz, B., Purushothaman, J., Quintela-Alonso, P., Rotterová, J., Santoferrara, L., Shao, C., Shen, Z., Shi, X., Song, W., Stoeck, T., La Terza, A., Vallesi, A., Wang, M., Weisse, T., Wiackowski, K., Wu, L., Xu, K., Yi, Z., Zufall, R., and Agatha, S. *Beyond the "Code" : A Guide to the Description and Documentation of Biodiversity in Ciliated Protists (Alveolata, Ciliophora)*. The Journal of Eukaryotic Microbiology, 64(4) :539–554 (2017). doi :10.1111/jeu.12391. (Cité page 4.)

Westesson, O., Skinner, M., and Holmes, I. *Visualizing next-generation sequencing data with JBrowse*. Briefings in Bioinformatics, 14(2) :172–177 (2013). doi :10.1093/bib/bbr078. (Cité page 40.)

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., and Schulman, A. H. *A unified classification system for eukaryotic transposable elements*. Nature Reviews. Genetics, 8(12) :973–982 (2007). doi :10.1038/nrg2165. (Cité pages 12, 42, 43, 45 et 173.)

Witte, C. P., Le, Q. H., Bureau, T., and Kumar, A. *Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes*. Proceedings of the National Academy of Sciences of the United States of America, 98(24) :13778–13783 (2001). doi :10.1073/pnas.241341898. (Cité page 44.)

Woese, C. R., Kandler, O., and Wheelis, M. L. *Towards a natural system of organisms : proposal for the domains Archaea, Bacteria, and Eucarya*. Proceedings of the National Academy of Sciences of the United States of America, 87(12) :4576–4579 (1990). (Cité page 3.)

Wu, X. and Brewer, G. *The regulation of mRNA stability in mammalian cells : 2.0*. Gene, 500(1) :10–21 (2012). doi :10.1016/j.gene.2012.03.021. (Cité pages 13 et 15.)

Xiong, J., Yang, W., Chen, K., Jiang, C., Ma, Y., Chai, X., Yan, G., Wang, G., Yuan, D., Liu, Y., Bidwell, S. L., Zafar, N., Hadjithomas, M., Krishnakumar, V., Coyne, R. S., Orias, E., and Miao, W. *Hidden genomic evolution in a morphospecies-The landscape of rapidly evolving genes in Tetrahymena*. PLoS biology, 17(6) :e3000294 (2019). doi :10.1371/journal.pbio.3000294. (Cité page 51.)

Yandell, M. and Ence, D. *A beginner's guide to eukaryotic genome annotation*. Nature Reviews. Genetics, 13(5) :329–342 (2012). doi :10.1038/nrg3174. (Cité page 40.)

Yao, M. C., Yao, C. H., and Monks, B. *The controlling sequence for site-specific chromosome breakage in Tetrahymena*. Cell, 63(4) :763–772 (1990). (Cité page 65.)

Yi, Z., Strüder-Kypke, M., Hu, X., Lin, X., and Song, W. *Sampling strategies for improving tree accuracy and phylogenetic analyses : a case study in ciliate protists, with notes on the genus Paramecium*. Molecular Phylogenetics and Evolution, 71 :142–148 (2014). doi : 10.1016/j.ympev.2013.11.013. (Cité page 56.)

Yu, J., Hu, S., Wang, J., Wong, G. K.-S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Huang, X., Li, W., Li, J., Liu, Z., Li, L., Liu, J., Qi, Q., Liu, J., Li, L., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Zhang, J., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Ren, X., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W., Li, P., Chen, W., Wang, X., Zhang, Y., Hu, J., Wang, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Li, G., Liu, S., Tao, M., Wang, J., Zhu, L., Yuan, L., and Yang, H. *A draft sequence of the rice genome (Oryza sativa L. ssp. indica)*. Science (New York, N.Y.), 296(5565) :79–92 (2002). doi :10.1126/science.1068037. (Cité page 42.)

Zagulski, M., Nowak, J. K., Le Mouël, A., Nowacki, M., Migdalski, A., Gromadka, R., Noël, B., Blanc, I., Dessen, P., Wincker, P., Keller, A.-M., Cohen, J., Meyer, E., and Sperling, L. *High coding density on the largest Paramecium tetraurelia somatic chromosome*. Current biology : CB, 14(15) :1397–1404 (2004). doi :10.1016/j.cub.2004.07.029. (Cité page 82.)

Zahler, A. M., Neeb, Z. T., Lin, A., and Katzman, S. *Mating of the stichotrichous ciliate Oxytricha trifallax induces production of a class of 27 nt small RNAs derived from the parental macronucleus*. PLoS One, 7(8) :e42371 (2012). doi :10.1371/journal.pone.0042371. (Cité page 66.)

Zerbino, D. R. and Birney, E. *Velvet : algorithms for de novo short read assembly using de Bruijn graphs*. Genome Research, 18(5) :821–829 (2008). doi :10.1101/gr.074492.107. (Cité pages 27 et 127.)

Zytnicki, M., Akhunov, E., and Quesneville, H. *Tedna : a transposable element de novo assembler*. Bioinformatics (Oxford, England), 30(18) :2656–2658 (2014). doi :10.1093/bioinformatics/btu365. (Cité page 49.)

Ce document a été préparé à l'aide de l'éditeur de texte TexMaker et du logiciel de composition typographique LaTeX 2ε.

Titre Annotation de génomes de paramécies

Résumé Les nouvelles technologies de séquençage (NGS) ont démocratisé le séquençage conduisant à une augmentation du nombre de génomes disponibles. Le décryptage de l'information qu'ils contiennent représente une étape cruciale. Dans ce manuscrit, je me focalise sur l'impact des NGS sur l'annotation des génomes de paramécies. Ces eucaryotes unicellulaires possèdent deux types de noyau, un noyau germinal (MIC) qui transmet l'information génétique à la génération suivante et un noyau somatique (MAC) qui assure l'expression des gènes. A chaque génération sexuelle, un nouveau MAC est produit à partir d'une copie du MIC suite à des réarrangements programmés de génome, assurant l'élimination d'éléments transposables (ET) et de petites séquences à copie unique appelées IES (*Internal Eliminated Sequence*).
Mon travail a impliqué le développement de nouvelles procédures, utilisant notamment des données ARN-seq, pour l'annotation de gènes de paramécies, prenant en compte leurs caractéristiques particulières, comme la taille minuscule des introns (20 - 30 nt) et la forte densité en gènes des chromosomes.
J'ai développé le logiciel ParTIES pour annoter les 45,000 IES du génome de *Paramecium tetraurelia* et montré que les IES sont des reliques d'ET. Nous avons pu décrire la dynamique évolutive des IES en exploitant une série de trois duplications globales de génome dans l'histoire de cette espèce.

Mots-clés Génome Annotation Paramecium Séquençage IES

Title Annotation of paramecium genomes

Abstract Next generation sequencing technologies (NGS) have democratized sequencing making more and more genomes available. Deciphering the information they contain remains a critical step. In this manuscript, I focus on the impact of NGS on the annotation of *Paramecium* genomes. These unicellular eukaryotes have two kinds of nuclei, a germline nucleus (MIC) transmits the genetic information to the next generation and a somatic nucleus (MAC) assures gene expression. At each sexual generation, a new MAC develops from a copy of the MIC after programmed genome rearrangements that eliminate transposable elements (TE) and short unique copy sequences called IESs (Internal Eliminated Sequences).
My work involved development of new procedures to annotate Paramecium genes based on the use of RNA-Seq data. The pipeline takes into account specific characteristics of *Paramecium* genomes such as tiny introns (20-30 nt) and high gene density along chromosomes.
I developed the ParTIES software suite to annotate IESs. We found 45,000 IESs in the genome of *Paramecium tetraurelia* and showed that they are relics of TE. A series of three ancient whole genome duplications in the lineage made it possible to use a comparative genomics approach to describe the evolutionary dynamics of IESs.

Keywords Genome Annotation Paramecium Sequencing IES

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Annotation de génomes de paramécies

Mots clés : Génome Annotation *Paramecium* Séquençage IES

Résumé : Les nouvelles technologies de séquençage (NGS) ont démocratisé le séquençage conduisant à une augmentation du nombre de génomes disponibles. Le décryptage de l'information qu'ils contiennent représente une étape cruciale. Dans ce manuscrit, je me focalise sur l'impact des NGS sur l'annotation des génomes de paramécies. Ces eucaryotes unicellulaires possèdent deux types de noyau, un noyau germinal (MIC) qui transmet l'information génétique à la génération suivante et un noyau somatique (MAC) qui assure l'expression des gènes. A chaque génération sexuelle, un nouveau MAC est produit à partir d'une copie du MIC suite à des réarrangements programmés de génome, assurant l'élimination d'éléments transposables (ET) et de petites séquences à copie unique appelées IES (*Internal Eliminated Sequence*).

Mon travail a impliqué le développement de nouvelles procédures, utilisant notamment des données ARN-seq, pour l'annotation de gènes de paramécies, prenant en compte leurs caractéristiques particulières, comme la taille minuscule des introns (20 – 30 nt) et la forte densité en gènes des chromosomes.

J'ai développé le logiciel ParTIES pour annoter les 45,000 IES du génome de *Paramecium tetraurelia* et montré que les IES sont des reliques d'ET. Nous avons pu décrire la dynamique évolutive des IES en exploitant une série de trois duplications globales de génome dans l'histoire de cette espèce.

Title : Annotation of paramecium genomes

Keywords : Genome Annotation *Paramecium* Sequencing IES

Abstract : Next generation sequencing technologies (NGS) have democratized sequencing making more and more genomes available. Deciphering the information they contain remains a critical step. In this manuscript, I focus on the impact of NGS on the annotation of *Paramecium* genomes. These unicellular eukaryotes have two kinds of nuclei, a germline nucleus (MIC) transmits the genetic information to the next generation and a somatic nucleus (MAC) assures gene expression. At each sexual generation, a new MAC develops from a copy of the MIC after programmed genome rearrangements that eliminate transposable elements (TE) and short unique copy sequences called IESs (Internal Eliminated Sequences).

My work involved development of new procedures to annotate *Paramecium* genes based on the use of RNA-Seq data. The pipeline takes into account specific characteristics of *Paramecium* genomes such as tiny introns (20-30 nt) and high gene density along chromosomes.

I developed the ParTIES software suite to annotate IESs. We found 45,000 IESs in the genome of *Paramecium tetraurelia* and showed that they are relics of TE. A series of three ancient whole genome duplications in the lineage made it possible to use a comparative genomics approach to describe the evolutionary dynamics of IESs.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France