Université de Montréal

Séquençage des génomes nucléaires d'eucaryotes unicellulaires 'primitifs' : les jakobides

Par

Samuel Prince

Département de biochimie et médecine moléculaire, Faculté de médecine

Mémoire présenté en vue de l'obtention du grade de maîtrise

en biochimie, option générale

Novembre 2022

Université de Montréal

Département de biochimie et médecine moléculaire

Ce mémoire intitulé

Séquençage des génomes nucléaires d'eucaryotes unicellulaires 'primitifs' : les jakobides

Présenté par

Samuel Prince

A été évalué par un jury composé des personnes suivantes

Adrian Serohijos
Président-rapporteur

B. Franz Lang
Directeur de recherche

Martin Smith
Membre du jury

Résumé

Les eucaryotes sont des organismes chimériques issus de l'endosymbiose entre une archéobactérie et une α-protéobactérie. Au cours de ce processus, ces organismes ont évolué de sorte à obtenir un grand nombre de caractéristiques observées chez les eucaryotes modernes, notamment une mitochondrie, un noyau, un système endomembranaire, un système d'épissage ou encore des chromosomes linéaires terminés par un télomère. Bien que les caractéristiques du dernier ancêtre commun des eucaryotes aient majoritairement été identifié, la suite des évènements évolutifs ayant mené à l'apparition de cet organisme demeure peu compris. Afin de mieux reconstruire cette suite d'évènements, l'analyse des génomes d'organismes basals aux eucaryotes sera nécessaire pour identifier des traces de cette évolution. Ainsi, nous proposons que l'analyse d'une collection de génomes d'eucaryotes « primitifs », les jakobides et malawimonades, des eucaryotes unicellulaires flagellés se nourrissant de bactéries, pourrait permettre une meilleure compréhension de ce processus. De plus, il a été supposé que le génome d'un de ces organismes, *Andalucia godoyi*, pourrait posséder des chromosomes circulaires, une caractéristique atypique chez les eucaryotes, une caractéristique qui pourra être confirmée par la production d'assemblage génomique de haute contigüité.

Afin d'obtenir des assemblages génomiques de haute qualité, les jakobides *A. godoyi*, *Jakoba bahamiensis*, *Seculamonas ecuadoriensis, Stygiella incarcerata* et le malawimonades *Malawimonas californiana* ont été séquencés par nanopore. Le séquençage nanopore a présenté des résultats mitigés et les organismes *J. bahamiensis* et *M. californiana* ont présentés un faible rendement de séquençage, possiblement dû à la contamination par des polysaccharides. Pour les autres organismes, nous avons développé un pipeline d'assemblage utilisant les assembleurs Flye et Shasta qui nous a permis de produire des assemblages génomiques. L'analyse du génome de *A. godoyi* a permis d'identifier la présence de quatre chromosomes circulaires, possiblement localisés dans le noyau, contenant plusieurs gènes liés au métabolisme, au transport et à la signalisation et qui constituent possiblement un type de chromosome circulaire différent de ceux observés précédemment chez les eucaryotes. Dans l'ensemble, ces travaux ont permis la mise en

place d’une collection de génome d’eucaryotes « primitifs » qui pourront être utilisés pour des analyse de génomique comparative afin de mieux comprendre l’évolution des eucaryotes.

Mots-clés : protistes, jakobides, malawimonades, évolution, origine des eucaryotes, chromosome circulaire, nanopore

Abstract

Eucaryotes are chimeric organisms that are the product of an endosymbiotic event between an archaebacteria and an α-proteobacteria. During the eukaryogenesis, these organisms have gained many characteristics that defines modern eucaryotes such as a mitochondrion, a nucleus, an endomembrane system, the splicing machinery, and linear chromosome with telomeres. While most characteristics of the last common eukaryote ancestor have mostly been identified, most of the evolutionary process that led to this organism is still unknown. To reconstruct this string of event, we must analyse the genome of "primitive" basal eukaryotes with a slow evolutionary rate and a lifestyle like that of the last common eukaryotes ancestor, and thus are most likely to contain remains of ancestral mechanisms that have been lost in most known eukaryotes. We propose that this analysis of the genome of the jakobids and malawimonads, two groups are free-living flagellate that feeds on bacteria, could provide such clues on the evolution of eukaryotes. Using nanopore sequencing, a collection of high-quality genomes has been built to help in this analysis. Furthermore, it has been supposed that the genome of the jakobid *Andalucia godoyi* could be composed to both linear and circular chromosomes, a genomic structure that have not been identified in other eukaryotes, which was investigated using the high quality nanopore assembly.

To generate a collection of high-quality genome assemblies, we have sequenced the genomes of the jakobids *A. godoyi*, *Jakoba bahamiensis*, *Seculamonas ecuadoriensis* and *Stygiella incarcerata* as well as the malawimonad *Malawimonas californiana* by nanopore. While the yields were too low for *J. bahamiensis* and *M. californiana*, probably due to a contamination by polysaccharides, we were able to assemble chromosome level genome for *A. godoyi* and *S. incarcerata* and high-quality draft genome for *S. ecuadoriensis* et *R. americana*. Using this assembly, we were able to identify four circular chromosomes in the genome of *A. godoyi*. The circular chromosomes are likely to be located in the nucleus and encodes genes with functions related to the metabolism, ions and macromolecules transport as well as signaling. Furthermore, these molecules differ from known circular chromosome in eukaryotes as they are unlikely to be

selfish DNA elements, such as known eucaryotes plasmids, or circular by-product of replication identified in other eukaryotes. Overall, this work sets the bases for larger scale comparative genomics of the jakobids and malawimonads, by generating a small collection of genomes that will be used in future studies to better understand the origin of the eukaryotes.

Keywords: protists, jakobids, malawimonads, evolution, eukaryote origin, circular chromosome, nanopore

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

ADN : Acide désoxyribonucléique

ADNecc : ADN circulaire extrachromosomal

ARN : Acide ribonucléique

ARNlnc : ARN long non codant

ARNnc : ARN non codant

CTAB : Bromure de cétyltriméthylammonium

CM : Modèle de covariance

EDTA : Éthylènediaminetétraacétique

HCl : Acide chloridrique

HMM : Modèle caché de Markov

Indel : Insertion et délétion

LECA : Dernier ancêtres commun des eucaryotes

MWCO : Limite de poids moléculaire

ONT : Oxford Nanopore Technologies

SDS : Dodécylsulfate de sodium

SNP : *Single nucleotide variant* (variant mononucléotidique)

TAT : *Twin-arginine translocation*

Tris : 2-amino-2-hydroxyméthylpropane-1,3-diol

Remerciements

Tout d'abord, j'aimerais remercier Dr. B. Franz Lang. Merci de m'avoir ouvert les portes de votre laboratoire pour un stage en 2019 puis en 2020 pour ma maitrise. Merci d'avoir pris le temps d'échanger avec moi autour de plusieurs idées. Ces échanges ont toujours été très passionnants et formateurs et m'ont permis d'arriver là où j'en suis.

J'aimerais aussi remercier les membres du laboratoire, entre autres Matt Sarasin qui m'a aidé sur plusieurs aspects de mon projet et qui a aussi pris le temps de s'occuper de mes problèmes informatiques, mais aussi Lise Forget qui m'a accompagné dans mon travail au laboratoire tout en me faisant la genèse des 30 dernières années du département.

Merci aux membres de mon jury, Dr. Adrian Serohijos et Dr. Martin Smith, d'avoir accepté d'évaluer mon mémoire.

Merci à mes parents pour le soutient qu'ils ont pu m'apporter.

Merci à Kamélia Maguemoun pour tout le soutiens qu'elle m'a apporté à travers ce parcours.

Chapitre 1 – Introduction

L'évolution de la complexité de la vie observée sur terre est un processus complexe marqué par plusieurs transitions évolutives majeures, et notamment par l'apparition d'un groupe de cellules à l'organisation complexe : les eucaryotes (1). Bien que l'identification des caractéristiques et de la position phylogénique de ces organismes datent de la moitié du 20ème siècle (2,3), plusieurs aspects de cette transition demeurent incompris. Les avancées technologiques des dernières décennies en séquençage, assemblage de génomes et identification de gènes ont permis de mieux définir la position évolutive des différents organismes, mais aussi de commencer à comprendre l'histoire évolutive ayant mené à l'apparition des eucaryotes (4). Pourtant, plusieurs questions restent toujours sans réponse, les principales étants : quel est l'ordre des évènements évolutifs qui ont permis l'apparition des eucaryotes, et quel est l'ensemble de gènes caractéristiques dans les ancêtres des eucaryotes?

L'origine des eucaryotes

Apparition des eucaryotes

Les eucaryotes sont souvent décrits comme des organismes chimériques issus de l'endosymbiose entre une archéobactérie et une α-protéobactérie (5,6). Cette dernière est devenue la mitochondrie, et occupe ainsi à la fois le rôle de central énergétique, dans la synthèse de métabolites et des acides aminés en plus d'avoir agi comme moteur évolutif chez les premiers eucaryotes (5,6). Ce faisant, l'endosymbionte a contribué significativement au bagage génétique de la cellule eucaryote par le transfert des gènes mitochondriaux vers le noyau (7–9). De plus, l'eucaryogénèse, processus symbiotique ayant permis l'apparition des premiers eucaryotes, est marquée par un nombre important d'échanges horizontaux de gènes bactériens vers l'ancêtre des eucaryotes (6,10,11). La combinaison de transfert de gènes horizontaux ou avec le génome mitochondrial, ont permis au proto-eucaryote de combiner des gènes bactériens, principalement impliqués dans le métabolisme, avec des gènes d'archéobactéries, majoritairement impliqués dans les voies de signalisations et les mécanismes de réplication et de transcription (9,12). En même temps, l'apparition de plusieurs gènes paralogues aurait permis à l'ancêtre commun des eucaryotes d'évoluer rapidement en modifiant des gènes procaryotes pour former de nouveaux gènes spécifiques aux eucaryotes (13,14). Ultimement, la combinaison de tous ces gènes aura

permis la formation de structures spécifiques aux eucaryotes tel que le noyau, le cytosquelette d'actine tubuline et les mitochondries. Il est à noter que les détails de l'eucaryogénèse et l'identité des partenaires symbiotiques restent toujours des hypothèses controversées, malgré l'identification et la caractérisation récente de génomes d'archéobactéries qui sont phylogénétiquement rapprochées aux eucaryotes (15–17).

Caractéristiques des premiers eucaryotes

Au niveau de l'ultrastructure (*i.e.*, la structure de leur composants cellulaires), il a été postulé que les premiers eucaryotes devaient être des organismes unicellulaires, bactérivore et flagellés pouvant ressembler à plusieurs eucaryotes unicellulaires modernes (18–21). De récentes études de génomique comparative ont permis à plusieurs groupes de recherche d'inférer les caractéristiques génétiques de l'ancêtre commun des eucaryotes, puisque les protéines et ARN structurés communes à la majorité des taxons d'eucaryotes connus sont susceptibles de provenir de l'ancêtre commun (4,13). Ces études ont postulé que l'ancêtre commun des eucaryotes était un organisme complexe, potentiellement même plus complexe que la plupart des eucaryotes modernes. Cette cellules ancestrale possédait notamment un noyau (22), un système endomembranaire (réticulum endoplasmique, appareil de golgi et lysosomes) (23,24), un cytosquelette d'actine/tubuline qui entre autre permet la formation de structures spécialisées comme des flagelles, et qui permet la ségrégation des chromosomes (13), un système d'épissage spliceosomal (25) et un cycle cellulaire sexué (méiose) (26). Pourtant, ces études comparatives viennent avec un degré d'incertitude élevée, comme ils ne comprenait pas de génomes de jakobides et malawimonades, lignées de protistes centrales dans l'hypothèse ultrastructurelle de l'ancêtre d'eucaryotes (18,21).

Les jakobides et les malawimonades

Les jakobides et malawimonades sont des groupes d'eucaryotes unicellulaires, non-parasitaires, bactérivores (prédateurs) et flagellés (21). Cinq genres de *Jakobidae* ont été identifiés : *Histiona* (21,27), *Jakoba* (28), *Reclinomonas* (29), *Andalucia* (30) et *Seculamonas* (31), et seulement trois isolats de malawimonades (Malawimonadea) sont présentement disponibles, et décrit en détail : *Malawimonas californiana* (20,31–33), *Malawimonas jakobiformis* (34) et *Gefionella okellyi* (32). Les premières analyse de ces organismes, principalement basées sur la comparaison de l'ultrastructure ont mis en lumière leur caractère ancestral. De plus, en raison de leur similarité avec celle attendue pour les eucaryotes primitifs,

ces études les ont décrits comme des organismes importants pour la compréhension de l'origine des eucaryotes (18,21). Le placement phylogénétique de ces organismes va dans le même sens, plaçant les jakobides dans les Discoba, un sous-règne d'organismes situé à la base des plantes et des algues rouges, et les Malawimonadea à la base d'un groupement d'eucaryotes qui réunit entre autres les animaux, les mycètes et les amœbozoaires (Figure 1) (20). Ainsi, à la fois les Jakobida et les Malawimonadea pourrait retenir des caractéristiques génétiques ancestrales (4,27,33,35).

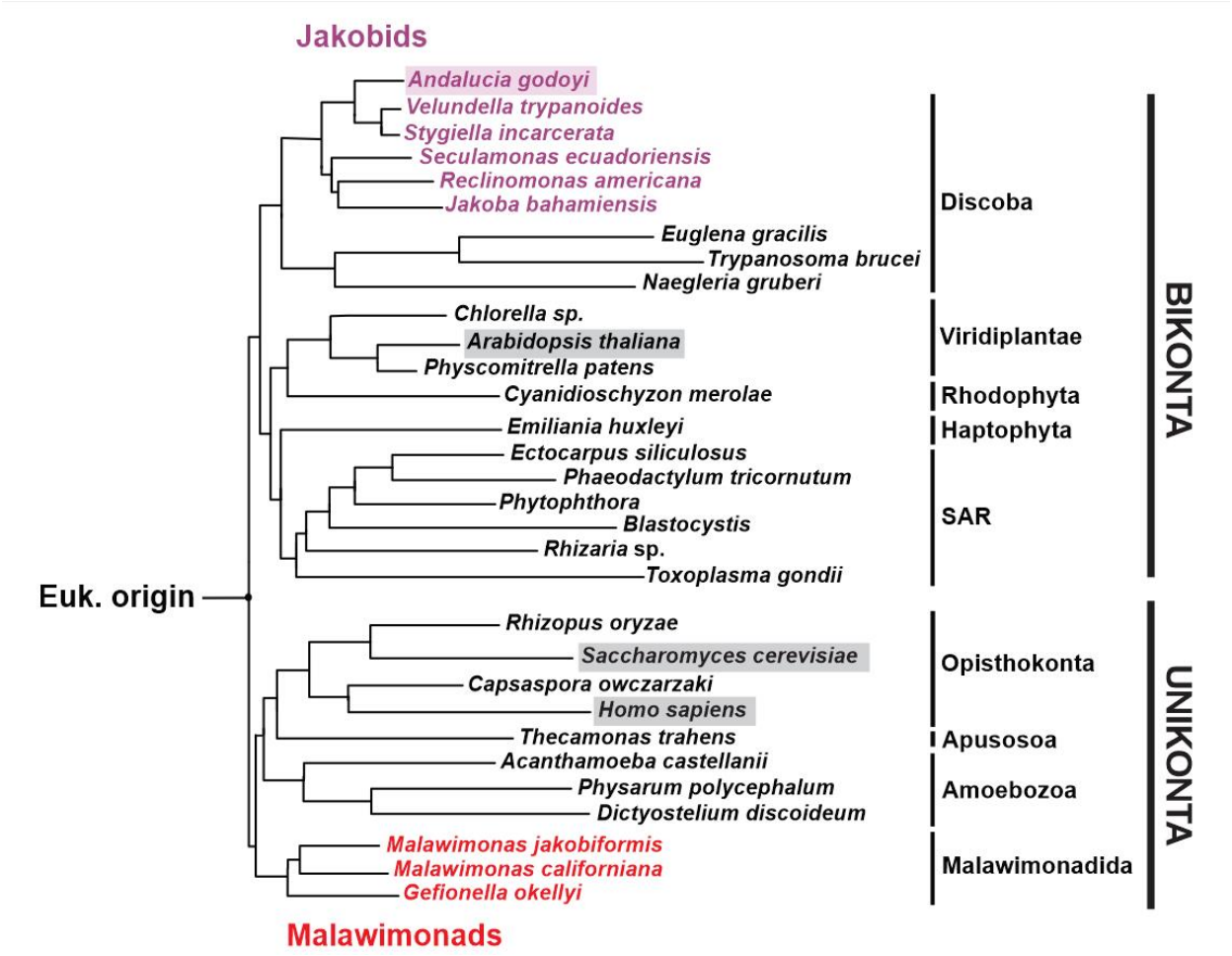

Figure 1. – Placement phylogénétique des jakobides et malawimonades dans les eucaryotes. La topologie de la phylogénie et la longueur des branches sont été obtenues de (31,36,37).

En fait, les génomes mitochondriaux (ADNmt) des jakobides sont particulièrement intéressant d'un point de vue évolutif, puisqu'ils possèdent le plus grand nombre de caractéristiques similaires aux bactéries, parmi tous les eucaryotes connus (31,38). Ces ADNmt possèdent les gènes *rpoA-D* qui encodent une ARN polymérase de type bactérienne et le facteur régulateur de transcription sigma, alors que ce système a été remplacé par une ARN polymérase de type phagique (T3/T7) chez tous les autres eucaryotes

(39,40). De plus, la plupart de gènes qui encodent des protéines dans les ADNmt des jakobides contiennent des motifs Shine-Dalgarno, participant au recrutement de l'ARN polymérase bactérienne. Seule une autre espèce d'eucaryote connue, *Palpitomonas bilix*, un organisme du supergroupe SAR (incluant notamment les algues brunes et les diatomées), partage la présence de motifs Shine-Dalgarno, ainsi qu'une organisation de gènes en opérons bactériens qui a été initialement retrouvée dans les jakobides (31,38,41). Finalement, le contenu génique de ces génomes mitochondriaux est plus important que celui de tous les organismes connus, avec jusqu'à 100 gènes codant chez *Andalucia* (31). Ces caractéristiques ont permis à certains auteurs de poser l'hypothèse que les génomes nucléaires de ces organismes pourraient aussi contenir des caractéristiques ancestrales (42,43).

Les malawimonades n'ont pas d'ADNmt si riches en gènes que les jakobides, mais se caractérisent par un autre trait génétique fort intéressant pour explorer l'origine des eucaryotes : un taux d'évolution de séquences géniques exceptionnellement réduit, par rapport à tous les autres groupes d'eucaryotes (résultats non-publiés de notre laboratoire). Malgré cette propriété fort intéressante, nous avons décidé de faire d'abord focus sur la génomique des jakobides, et de seulement explorer la faisabilité d'un projet génomique sur les malawimonades.

Andalucia godoyi

Parmi les jakobides, une espèce est particulièrement intéressante pour débuter l'analyse des génomes nucléaires: *Andalucia godoyi*. En effet, il possède le plus grand nombre de protéines encodées par le génome mitochondrial de tous les eucaryotes connus (31), son génome nucléaire est de petite taille (~20 Mb) et les études phylogénétiques le classent sans ambiguïté à la base des jakobides (31,42). Cette espèce a d'ailleurs fait l'objet de la première analyse à grande échelle des génomes nucléaires des jakobides qui était orientée sur le mitoprotéome par Gray et *al.* (42). Le contenu génique du mitoprotéome nucléaire (894 protéines) de *A. godoyi* est similaire à celui d'autres eucaryotes soit la levure (901 protéines) et l'humain (1158 protéines) et contient la majorité des protéines impliquées dans les voies métaboliques conservées chez les eucaryotes mais encode certains gènes d'origine α-proteobacterienne (42,44,45). En ce sens, ce mitoprotéome est représentatif de ceux des eucaryote et ne représente pas un stade minimaliste transitionnel entre les eucaryotes primitifs et les eucaryotes plus dérivés (42). Cette caractéristique n'en est pas moins intéressante puisqu'elle supporte fortement l'hypothèse que le dernier ancêtre commun des eucaryotes (LECA) possédait déjà un mitoprotéome complexe (42). Il contient

cependant plusieurs caractéristiques ancestrales, tel que la présence des 20 protéines de la petite sous unité ribosomique suggérée dans LECA par Desmond et al. (42,46). Le mitoprotéome de *A. godoyi* contient aussi quelques particularités qui le distinguent des autres eucaryotes, notamment tous les composants d'un système TAT minimal de translocation des protéines dans la mitochondrie (42,47). Cette première étude s'imbrique dans une étude plus générale des génomes des jakobides (42), cependant elle ne couvre qu'une fraction des gènes du protéome entier de *A. godoyi* et l'origine évolutive de la majorité de ces protéines n'a encore fait l'objet d'aucune étude. De plus, l'assemblage du génome d'*A. godoyi* qui a été utilisé pour les analyses mentionnées ci-haut n'est qu'une version préliminaire, et donc est potentiellement incomplet.

Architecture du génome de A. godoyi

En plus d'améliorer la connaissance du contenu génique de ces organismes, l'assemblage de génome de haute qualité des jakobides et malawimonades, et plus spécifiquement d'*A. godoyi*, pourra permettre de définir la structure de leurs chromosomes et ainsi de déterminer si leur structure est similaire à celle des autres eucaryotes.

La seule étude visant à identifier la structure des chromosomes de ces organismes a identifié la présence de télomères aux extrémités des chromosomes des malawimonades et du jakobide *R. americana* (48). Ceci suggère que la structure des chromosomes nucléaires devrait, a priori, être similaire à celle des autres eucaryotes connus (*i.e.,* des chromosomes linéaires terminés par des télomères). Cependant, en raison de leur structure dans l'assemblage public de *A. godoyi*, il a été suggéré que certains chromosomes pourraient être des molécules circulaires (Marek Eliáš, communication personnelle). Les données de séquençages ne permettaient cependant pas d'établir avec certitude la structure des chromosomes. En effet, il pourrait s'agir d'un artéfact d'assemblage puisque la présence de répétitions en tandem peut produire une séquence qui semble circulaire. Ainsi, la production d'un assemblage de haute qualité, avec de longues lectures nanopore qui permettent à mieux résoudre l'assemblage des chromosomes, est nécessaire pour déterminer avec certitude la structure du génome.

Séquençage et assemblage des génomes

Depuis la parution du premier génome bactérien complet (49), les avancées dans le domaine du séquençage ont permis d'accumuler une importante quantité de données multi-omiques sur les organismes vivants transférant ainsi la limite analytique du laboratoire à la bio-informatique (50). Les

organismes moins étudiés sont particulièrement affectés par ce phénomène puisque plusieurs outils sont développés et optimisé pour les organismes modèles ce qui limite particulièrement la capacité à analyser ces organismes (51,52). En effet, plusieurs analyses de génomique comparative nécessitent un assemblage génomique et une annotation de haute qualité qui ne sont pas disponibles chez ces organismes (53). L'analyse génomique de jakobides et malawimonades pose encore une autre difficulté spécifique à ces organismes : leur nourriture exclusive consiste en bactéries vivantes qui co-existent en grand nombre dans les isolats provenant de laboratoires de recherche ou les collections d'espèces comme le ATCC (54). Toutes tentatives des chercheurs d'établir des lignées de jakobides ou malawimonades axéniques ont jusqu'à date échouées. Même la réduction de la diversité de cette « contamination » bactérienne s'est avérée difficile, à cause de la dépendance de survie de certains protistes sur des bactéries spécifiques (résultats non-publiés du laboratoire). Par conséquent, les cultures de jakobides et malawimonades contiennent une panoplie de bactéries qui rendent l'assemblage et l'identification du génome eucaryote dans le mélange très difficile.

Séquençage

Première étape dans la production d'un assemblage génomique, le séquençage permet de déterminer la séquence de fragments d'ADN, de longueur variable selon la technologie de séquençage, à partir desquels il est possible de reconstruire le génome en entier (55). La technologie de séquençage la plus couramment utilisée est le séquençage Illumina (56). Il s'agit d'une technologie de séquençage de deuxième génération basée sur le séquençage par synthèse (55). Elle consiste en plusieurs cycles d'incorporation de nucléotides fluorescents qui permettent d'identifier précisément la séquence. Cette technologie permet ainsi de générer de courts fragments (100-300 bases) d'une très haute précision (55). Plus récemment, des méthodes de séquençage de troisième génération ont été mises en marché (56,57). Il s'agit de technologies de séquençage en temps réel qui permettent de séquencer des molécules plus longues (jusqu'à quelques mégabases) et dont les principales sont PacBio (Pacific Bioscience) et nanopore (Oxford Nanopore Technologies) (58–60). Ces technologie permettent de séquencer des molécules d'ADN native et ainsi d'identifier non seulement la séquence, mais aussi d'identifier directement les bases modifiées dans l'ADN (ex. méthylation) (61). Cependant, le séquençage nanopore est la seule technologie qui permet de séquencer des molécules d'ADN sans limite de longueur.

Le concept de séquençage nanopore est basé sur l'utilisation d'une protéine pore ancrée dans une membrane de phospholipide pour identifier la séquence nucléotidique (62). Lors du séquençage, l'ADN est transloqué dans le pore par électrophorèse, ce qui cause une fluctuation dans le courant spécifique aux nucléotides présents dans le canal de la protéine et qui est mesuré par le séquenceur (62,63). La transformation du signal électrique en séquence nucléotidique est effectuée lors du *basecalling* (64,65). À cette fin, plusieurs algorithmes ont été développés (66,67), dont les plus efficaces sont basés sur des méthodes d'apprentissage machine comme les réseaux de neurones (57,58). Ces algorithmes permettent à la fois la détection des nucléotides canoniques (ACGT) et de certaines bases modifiées (*e.g.* m5C aux îlots CpG) (69).

Assemblage de génomes eucaryotes

Malgré les progrès dans la longueur des molécules d'ADN séquencées, aucune méthode ne parviennent à séquencer de chromosomes entiers (à l'exception des chromosomes courts tel que les chromosomes mitochondriaux) (70). Pour obtenir une séquence génomique de haute qualité, il est donc nécessaire d'assembler le génome, soit identifier la succession de fragments séquencés qui donne la séquence d'un chromosome (70). Il s'agit d'une problématique complexe et différents algorithmes ont été produits pour tenter de la résoudre, en tirant profit des avantages de chaque technologie de séquençage (70). Pour les données Illumina, les assembleurs utilisent généralement des graphes de Bruijn construits à partir d'une suite de courts fragments exacts (*k-mers*) identifiés à partir des séquences Illumina. Ce graphe est ensuite utilisé pour déterminer l'ordre optimal des fragments et reconstruire la séquence génomique (Figure 2C) (71). Bien que la courte taille de ces séquences limite la capacité des assembleurs à résoudre des régions répétées, tels que les transposons, l'utilisation de *k-mers* exacts permet de reconstruire une séquence qui ne contient presque aucune erreur (70). Les séquences PacBio ou nanopore étant beaucoup plus longues, une approche différente est utilisée pour procéder à l'assemblage (70). Plusieurs assembleurs utilisent un algorithme *Overlap-Layout-Consensus* qui se base sur l'identification de régions se chevauchant entre les séquences pour identifier l'ordre des séquences et reconstruire la séquence génomique entière (Figure 2B) (72,73). Cette méthode permet ainsi de résoudre plusieurs régions répétées, cependant, le taux d'erreur élevée des séquences initiales empêche la production d'un assemblage initial sans erreurs (74,75).

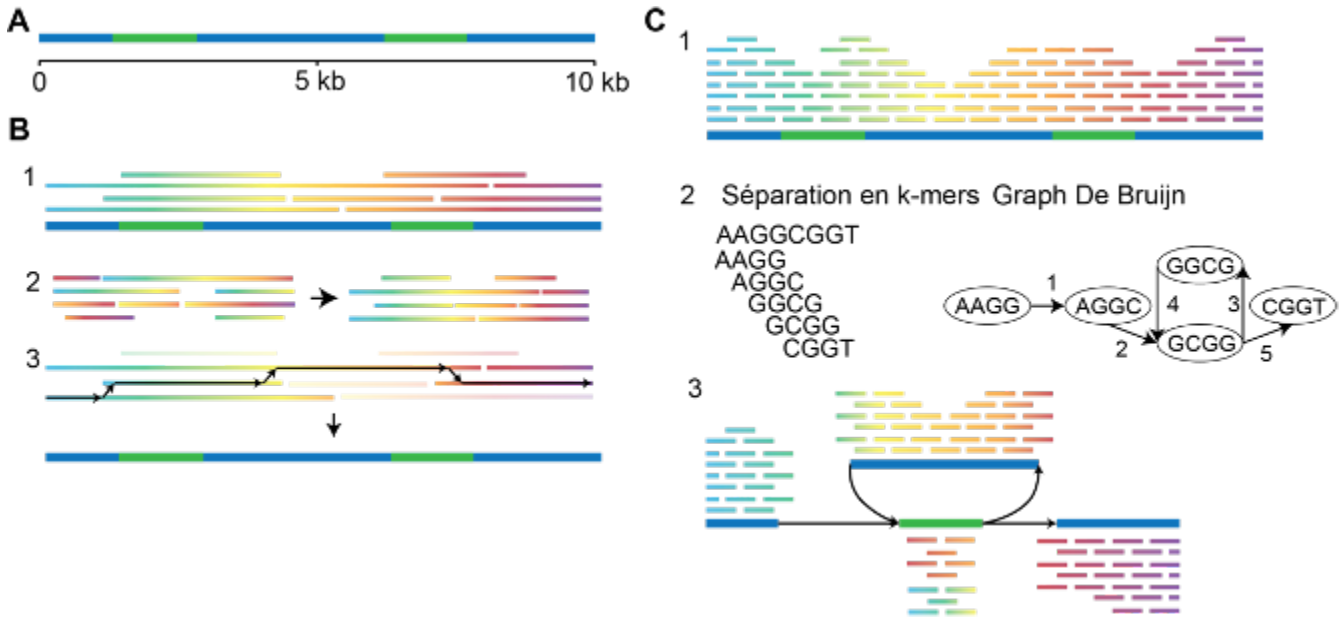

Figure 2. – Stratégie d'assemblage de génomes. Séquence hypothétique d'ADN (bleu : région unique, vert : région répétée dispersée). B. Assemblage du génome avec des données nanopore par une stratégie *Overlap-Layout-Consensus*. Les séquences produites par le séquençage nanopore (1) sont alignées les unes contre les autres pour identifier les chevauchements entre celles-ci (2). Un graphe est construit à partir de ces chevauchements qui permet d'identifier le meilleur enchainement de séquences pour reconstruire la séquence génomique (3). Puisque les séquences nanopore traversent les régions répétées, ces régions sont correctement assemblées et l'assemblage final est une seule région contiguë. C. Assemblage du génome avec des données Illumina à l'aide d'un graphe De Bruijn. Les séquences Illumina (1) sont séparées en k-mers qui sont utilisés pour construire le graphe De Bruijn (2). Dans ce graphe, le meilleur enchaînement de k-mers est identifié pour reconstruire la séquence initiale. À l'échelle génomique, la résolution du graph permet de reconstruire les séquences uniques et les séquences répétées entières, cependant, l'ordre de ces séquences ne peut être résolu, ce qui cause une fragmentation de l'assemblage (les fragments sont séparés par les flèches noires).

Les assembleurs nanopore emploient diverses stratégies différentes pour minimiser ces erreurs. Une première stratégie, utilisée par les assembleurs « hybrides » (tel que SPAdes-Hybrid, MaSuRCA ou Unicycler) consiste à utiliser à la fois des données nanopore et Illumina pour produire l'assemblage (76–78). Les données Illumina sont d'abord utilisées pour obtenir un assemblage sans erreurs mais hautement fragmenté avec un algorithme De Bruijn puis les données nanopore sont utilisées pour joindre les

différents fragments en un assemblage plus contigüe (76–78). D'autres assembleurs, cependant, utilisent exclusivement les données nanopore pour produire les assemblages (*e.g.,* Canu, Flye ou wtdbg2/Redbean) (52,79,80). Afin de limiter le nombre d'erreurs dans l'assemblage final, ils utilisent deux stratégies. D'abord, certains effectuent une correction et un prétraitement des lectures nanopore (*e.g.,* Canu et NECAT) avant de procéder à l'assemblage (80). Ensuite, les programmes d'assemblage effectuent généralement un polissage sur l'assemblage final, une étape qui sert à déterminer la séquence consensus à partir des données nanopore (52,79,80). Cette étape ne permet cependant pas de retirer l'ensemble des erreurs dans l'assemblage. Notamment, les erreurs systématiques du séquençage nanopore (principalement dans les homopolymères) ne peuvent être corrigées (81–83). Une étape supplémentaire de polissage utilisant de données Illumina doit donc être effectuée avec l'assemblage afin de réduire encore plus le nombre d'erreurs dans la séquence (84,85). Cependant, aucune stratégie ne permet, à ce jour, de produire un assemblage sans erreur et un certain nombre d'erreurs persistent (principalement dans les régions répétées) (74).

Alors que le nombre d'assembleurs et de stratégie d'assemblage publié augmente rapidement avec la publication de nouveau assembleurs tel que HASLR (86), Raven (87) ou Shasta (88) en plus des assembleurs établis comme Canu (80) ou Flye (79), peu d'articles offrent une comparaison de ces différents logiciels (la plus complète étant limité aux procaryotes dont le génome moins répétitif est plus simple à assembler (89)) ce qui complexifie l'identification de la suite d'outils produisant les meilleurs assemblages. De plus, la qualité de l'assemblage peut aussi être influencée par les étapes de prétraitement des données (i.e., *basecalling* et correction) ainsi que par le polissage. Or, l'impact de ces étapes sur la présence d'erreurs dans les assemblages n'est généralement pas analysé dans les articles de comparaison de ces outils (90,91). De plus, malgré ces avancées technologiques, l'assemblage de génome *de novo* reste complexe et plusieurs régions répétées d'une certaine longueur sont toujours difficile, voire impossible à assembler avec les technologies actuelles (74).

Estimation de la qualité des assemblages

En absence de génome de haute qualité (*gold standard*) il est difficile d'évaluer la qualité des assemblages comme c'est nécessaire, par exemple, pour comparer différents programmes d'assemblages (92). Une difficulté supplémentaire est présente chez les protistes dont les assemblages sont contaminés par les séquences de bactéries provenant des cultures (54). Des outils tel que QUAST ou GAGE produisent des métriques qui permettent d'évaluer la contiguïté des assemblages comme le nombre de contig ou le

N50 (plus petit contig tel que 50% du génome est inclus dans des contigs plus grands) (92,93). Ces métriques sont cependant peu informatives de l'exactitude et de la qualité du génome, deux facteurs importants pour l'identification de contenu génique du génome lors de l'annotation (50,94). De plus, pour obtenir plus de métriques tel que les régions mal assemblées, manquantes ou encore le taux d'erreur, ces programmes ont besoins d'un génome de référence (92,93). Il a été proposé que l'identification de gènes orthologues conservés pourrait permettre de pallier partiellement cette problématique (94). Cette idée a été développé par les auteurs du programme BUSCO, cependant ce programme est principalement développé pour les Opisthokonta (Metazoa et les eucaryotes unicellulaires apparentés, ainsi que Fungi) (95).

Problématique et hypothèse

Les récents développements technologiques ont permis le séquençage d'un nombre important d'organismes (principalement des métazoaires, champignons et bactéries d'intérêt commercial). Cependant, le manque de données génomiques, transcriptomiques et protéomiques de qualité chez les eucaryotes basals, depuis longtemps identifié comme un facteur limitant dans l'analyse de l'origine des eucaryotes et des caractéristiques communes à l'ensemble des eucaryotes (4), n'a toujours pas été résolu.

Ainsi, le séquençage et l'annotation d'une collection de génomes des eucaryotes « primitifs », les jakobides et malawimonades, pourrait fournir des données essentielles à une meilleure compréhension de l'évolution des eucaryotes. En effet, cette collection rendra possible des études de génomiques comparatives du contenu protéique, des ARN non codants ou de toutes autres caractéristiques génomiques intéressantes. De plus, la présence de plusieurs caractéristiques « ancestrales » chez certains de ces organismes nous permet de croire que ces assemblages permettront d'identifier des éléments ancestraux n'ayant pas été identifiés auparavant.

De plus, la production d'assemblages de haute qualité permettra de définir la structure des chromosomes. En effet, bien qu'il ait été suggéré que le génome de *A. godoyi* pourrait contenir un mélange de chromosomes circulaires et linéaires, les données disponibles précédemment ne permettent pas de déterminer la structure des chromosomes avec confiance. Il est donc possible que la circularité observée soit plutôt un artéfact. Ainsi, cette recherche visera aussi à déterminer la structure du génome nucléaire de *A. godoyi*. Notamment, elle permettra d'identifier si le génome nucléaire de *A. godoyi* et d'autres jakobides contient des chromosomes circulaires. Pour ce faire, il sera nécessaire de confirmer la

structure des chromosomes de *A. godoyi*, mais aussi de confirmer qu'il s'agisse bien de chromosomes eucaryotes et non de contaminants bactériens. Cette validation pourra être effectuée en déterminant l'origine des gènes présents sur ces molécules.

Chapitre 2 – Matériel et méthodes

Culture des protistes

Les souches *A. godoyi* PRA-185, *J. bahamiensis* ATCC 50695, *M. californiana* ATCC 50740, *R. americana* ATCC 50394, *S. ecuadoriensis* ATCC 50688 réduites en bactéries (tel que décrit dans (31)) ont été utilisées. *S. incarcerata* a été fournis par Andrew J. Roger (96). Les cultures de *A. godoyi*, *R. americana* et *S. ecuadoriensis* ont été effectuées dans du milieu de culture WCL et les cultures de *J. bahamiensis* et *M. californiana* dans du milieu de culture F/2. Les informations sur ces milieux de culture sont disponibles à http://megasun.bch.umontreal.ca/People/lang/FMGP/methods.html. La culture de *S. incarcerata* été effectuée dans du milieu marin artificiel à 40 g/L (Sigma, S9883) supplémenté avec 3% de milieu LB (10g/L peptone, 5g/L extrait de levure, 5g/L chlorure de sodium). Toutes les cultures, à l'exception de *S. incarcerata* étaient nourries avec *Enterobacter aerogenes* ATCC 13048 tous les deux à trois jours.

Extraction de l'ADN génomique

Les cellules des cultures de protistes en phase exponentielle ont été prélevées par centrifugation (12,000 x g; 20 min; 4°C) lorsque la majorité des bactéries avaient été consommées. Les culots étaient resuspendus dans un faible volume de milieu de culture avant de procéder à l'extraction.

SDS Protéinase K

Afin de lyser les cellules, 100 µg/ml de protéinase K et 0,2% de SDS ont étaient ajoutés au culot resuspendu. Le lysat a été incubé pendant 2h à 65°C pour permettre la digestion des protéines puis dialysé contre 1mM EDTA dans une membrane Spectra/Por membrane (MWCO 12-14,000; Fisher 08-667B) pendant 2, 4 et 8 heures. Après la dialyse, l'ADN a été purifié sur une colonne Qiagen G20 selon la méthode de fabriquant. Brièvement, 50 µg/ml de RNase A a été ajouté à la solution puis incubé à 37°C pendant 15 minutes. 100 µg/ml de protéinase K et 2 volumes de G2 3X (2 400 mM chlorure de guanidium, 90 mM Tris-HCl (pH 8,0), 90 mM EDTA (pH 8,0), 15% Tween,

1,5% Triton X-100) ont été ajoutés à la solution. La solution a été incubée pendant 1h à 65°C puis chargée sur la colonne. L'ADN a été lavé avec du tampon QC (Qiagen) puis élué avec du tampon QF (Qiagen). L'ADN a été précipité avec de l'isopropanol puis resuspendu dans du TE (10 mM Tris-HCl pH8, 1mM EDTA pH8).

CTAB

Les cellules ont été nettoyées par resuspension dans 0,6M sorbitol, TE (pH 8.0), puis centrifugées (1 000 x g; 20 mins). Le culot a été resuspendu dans du tampon CTAB (100 mM Tris-HCl (pH 7,5), 25 mM EDTA, 1,5 M NaCl, 2% (w/v) CTAB, 0,3% (v/v) β-mercaptoéthanol) pré-chauffé (10 ml / 1g de cellules) puis incubé à 65°C pendant 1 heure. L'ADN a ensuite été séparé des peptides par deux extractions avec 1 volume de chloroforme suivit d'une centrifugation (8 500 x g; 10 mins). Suite à la deuxième extraction, la phase aqueuse était dialysée contre 1 mM EDTA, 100 mM NaCl pendant 2h puis 1 mM EDTA pendant 2, 4 et 8 heures. Les débris cellulaires ont été retirés par centrifugation (17 000 x g; 10 mins) puis le surnageant a été traité avec 50 µg/ml de RNase A pendant 15 minutes à 37°C. L'ADN a ensuite été purifié avec une colonne Qiagen G20 tel que décrit dans la section précédente.

Évaluation de la qualité de l'ADN extrait

La qualité de l'ADN extrait a été évaluée en utilisant un spectrophotomètre NanoDrop 2000 (Thermo Scientific, USA). La fragmentation de l'ADN a été évaluée par séparation sur un gel d'agarose 1%. La concentration de l'ADN a été mesurée en utilisant l'essai dsDNA BR sur un fluoromètre Qubit (Invitrogen, USA).

Préparation des librairies et séquençage par nanopore

Les librairies de séquençage nanopore basée sur la ligation (ONT, cat # SQK-LSK110) ont été préparées selon une procédure similaire à celle du fabriquant. Environ 2µg d'ADN génomique ont été réparé et préparé pour la ligation (NEBNext FFPE (NEB, cat #M6630) et Ultra II End-prep (NEB, act #E7546)). L'ADN a ensuite été nettoyé en avec 0,5X de billes AMPure XP (Beckman Coulter, cat #A63880), puis les adaptateurs ont été ligués aux séquences. L'ADN a ensuite été nettoyé avec 0,4X de billes AMPure XP dans le tampon LFB. La librairie a ensuite été éluée puis 12 µL de librairie

ont été séquencés sur une *flow cell* MinIONR9.4.1. Le *basecalling* a ensuite été effectué avec Guppy (ONT; v6.1.2) utilisant le modèle *super high accuracy*.

Analyse de l'état des pores

L'état des pores a été obtenu à partir du rapport de *mux scan* généré par MinION et visualisé dans R (v4.2.0) en utilisant les librairies ggplot2 (v3.3.6), ggpubr (v0.4.0) et dplyr (v1.0.9). Le temps de survie des pores a été obtenu à partir des données de *mux scan* et analysé dans R (v4.2.0) en utilisant la librairie survival (v3.4). Le temps de survie des pores a été comparé au moyen d'un test Kaplan-Meier.

Séquençage Illumina

L'ADN extrait de *A. godoyi* et *S. incarcerata* a été envoyé à la plateforme de génomique de Génome Québec pour séquençage Illumina. Pour *A. godoyi*, des lectures à embouts pairés de 300 nucléotides ont été séquencées avec un séquenceur MiSeq en utilisant une librairie Illumina TruSeq DNA et une fragmentation à 600 nucléotides. Pour *S. incarcerata*, des lectures à embouts pairés de 150 nucléotides ont été séquencées avec un séquenceur Illumina NovaSeq en utilisant une librairie Illumina TruSeq DNA et une fragmentation à 500 nucléotides. Pour les autres organismes, les données lllumina obtenues préalablement au laboratoire ont été utilisées.

Contrôle de la qualité des lectures Illumina

Les adapteurs et les lectures de faible qualité ont été filtrées avec fastp (v0.23.2) (97). La qualité des jeux de données a été validée à partir du rapport produit par ce même programme. Les erreurs de séquençage présentes dans les séquences ont ensuite été corrigées avec Rcorrector (v1.0.4) (98). Après cette correction, les séquences n'ayant pas pu être corrigées ont été retirées.

Séquençage RNA-Seq

L'ARN total de *A. godoyi* a été extrait au moyen du kit RNeasy (Qiagen, USA). La qualité de l'ARN extrait a été évaluée avec un spectrophotomètre NanoDrop 2000 (Thermo Scientific, USA) et par séparation par électrophorèse sur gel d'agarose. L'ARN total a été envoyé à Génome Québec pour

le séquençage des ARNlnc. Pour les autres organismes, les données RNA-Seq disponibles au laboratoire ont été utilisés.

Assemblage des génomes

La méthode utilisée pour l'assemblage des génomes varie selon la ploïdie de l'organisme. Afin de déterminer la ploïdie, un premier assemblage préliminaire a été généré avec Shasta (v0.9.0) (88). Les lectures nanopore ont ensuite été alignées contre cet assemblage préliminaire avec minimap2 (v2.24) (99). Cet alignement a été visualisé dans IGV (v2.13.1) afin de déterminer la ploïdie. La présence d'indels et de SNPs hétérozygotes a été utilisée pour déterminer la ploïdie. Cette observation a ensuite été confirmée par l'analyse du profil de k-mers avec KAT (v2.4.1) (100) (Figure 3).

Pour les organismes haploïdes, un assemblage a d'abord été généré avec Shasta (v0.9.0) (88). Cet assemblage a été corrigés avec Medaka (ONT; v1.6.1), suivi par trois itérations de Pilon (v1.24) (84) (Figure 3). Les erreurs structurelles visibles dans l'assemblage ont ensuite été corrigé manuellement (selon la méthode décrite ci-dessous).

Pour les organismes diploïdes, deux stratégies ont été utilisées. Dans la première stratégies (utilisée pour assembler le génome de *A. godoyi*), l'assemblage a été généré avec Shasta en utilisant le mode « *phased diploid* ». Cet assemblage a ensuite été corrigé manuellement (selon la méthode décrite ci-dessous). Dans la deuxième stratégie (utilisée pour assembler le génome de *R. americana*), les lectures nanopore sont d'abord séparées en deux selon leur haplotype en utilisant Whatshap (101) (Figure 3). Chaque haplotypes sont ensuite assemblés individuellement tel que décrit pour les organismes haploïdes.

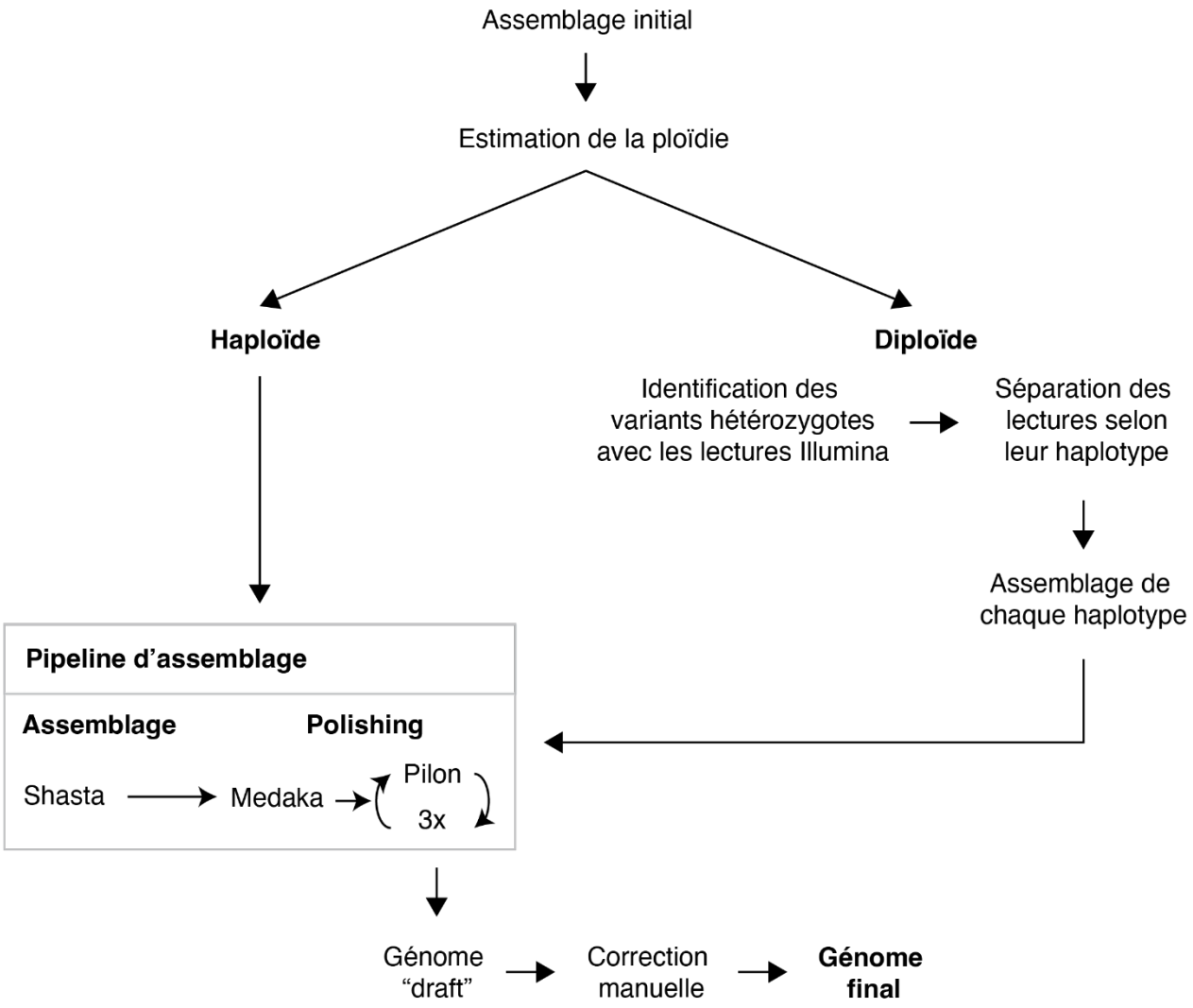

Figure 3. – Sommaire de la méthodologie d'assemblage des génomes des jakobides et malawimonades.

La version préliminaire des assemblages a ensuite été corrigée manuellement pour retirer les erreurs structurelles visibles. Pour ce faire, les lectures nanopore ont été alignées contre l'assemblage nanopore avec minimap2 en utilisant le préréglage map-ont (v2.24) (102) et cet alignement a été visualisé dans IGV. Les erreurs structurelles potentielles ont été identifiées dans IGV par la présence de *clipping* de plus de 100 nucléotides (assemblage incorrect), d'insertion (région présente dans l'assemblage mais pas dans les lectures) ou de délétion (région présente dans les lectures mais pas dans l'assemblage) dans l'alignement. Lorsque ces variations étaient

présentes dans une majorité des lectures (80-90%), la région génomique a été analysée pour tenter de corriger la structure de l'assemblage :

- Les **délétions** ont été corrigées en ajoutant la séquence d'une lecture nanopore.
- Les **insertions** ont été corrigées en retirant la section qui n'est pas présente dans les lectures nanopore de l'assemblage.
- Les **séquences chimériques** (fusion de deux chromosomes) ont été corrigées en séparant les deux séquences.
- Les **chromosomes fragmentés** en plusieurs contigs sont reconstruits manuellement en plaçant les contigs dans le bon ordre. Cette correction a été effectuée uniquement dans les cas où les lectures nanopore alignées aux extrémités d'un contig présente un *clipping* et que la région supplémentaire s'aligne sur un seul autre contig. L'alignement des lectures nanopore a été utilisé pour déterminer l'orientation des deux contigs puis les contigs ont été joints pour former un seul contig.

Pour toutes les corrections décrites ci-dessus, suite à la correction, les lectures nanopore étaient alignées contre le contig corrigé avec minimap2 en utilisant le préréglage map-ont (v2.24) (102) puis visualisé dans IGV pour confirmer que la structure corrigée est supportée par les lectures nanopore.

Comparaison des assemblages nanopore et Illumina

Les assemblages Illumina (et Illumina/454 pour *A. godoyi*) des jakobides et malawimonades ont été obtenus de https://megasun.bch.umontreal.ca/jakmals-downloads/. La contiguïté de ces assemblage ainsi que ces assemblages produits avec les données nanopore a été déterminée en utilisant QUAST (v5.2.0) (93). Afin de comparer le contenu génique de ces assemblages, ils ont d'abord tous été annotés avec BRAKER2 (103) (v2.1.6) en utilisant les données RNA-Seq disponibles au laboratoire. La complétude du contenu génique des assemblages a ensuite été déterminée avec BUSCO à partir de ces annotations en utilisant les données représentatives de l'ensemble des eucaryotes (95) (v5.2.1).

Afin d'identifier de potentielles améliorations dans la structure de l'assemblage nanopore de *A. godoyi* par rapport à l'assemblage Illumina/454, ces deux assemblages ont été alignés avec NUCmer (v4.0.0) (104). Cet alignement a ensuite été visualisé dans IGV (v2.13.1).

Annotation du génome de *A. godoyi*

Annotation structurelle

Le pipeline d'annotation structurel utilisé a été mis en place de manière similaire à ce qui est décrit dans (42) (Figure 4A). Brièvement, les séquences RNA-Seq ont été alignés contre le génome en utilisant STAR (v2.7.3) (105). Le transcriptome a ensuite été assemblé avec Trinity (v2.1.0) (106) en mode *de novo* et *genome-guided*. Dans les deux cas, l'assemblage par Trinity incluait une étape de *jaccard clipping* afin de diminuer le risque de fusion artificielle de transcrits. Les assemblages de transcriptome ont ensuite été combinés avec PASA (v2.4.1) (107). Des modèles de gènes ont ensuite été produits avec les outils de prédiction *ab initio* Augustus (v3.3.3) (108), SNAP (109), GeneMark (v4.33) (110) et CodingQuarry (v2.0) (111). Finalement, l'assemblage de PASA, les alignements de Spaln et les modèles de gènes de Augustus, SNAP, Genemark et Codingquarry ont été combinés en une annotation consensus avec EvidenceModeler (v1.1.1) (112).

Les répétitions ont été annotées en utilisant la suite EDTA (v2.0.1) (113) pour effectuer l'identification *de novo* des séquences répétées. La librairie produite par EDTA a ensuite été utilisée dans RepeatMasker (114) (v4.1.2) pour effectuer l'annotation des séquences répétées. Les séquences d'ARN de transfert ont été identifiées avec tRNAscan-SE (v2.0.9) (115) et les séquences d'ARN ribosomique avec barrnap (v0.9) (https://github.com/tseemann/barrnap).

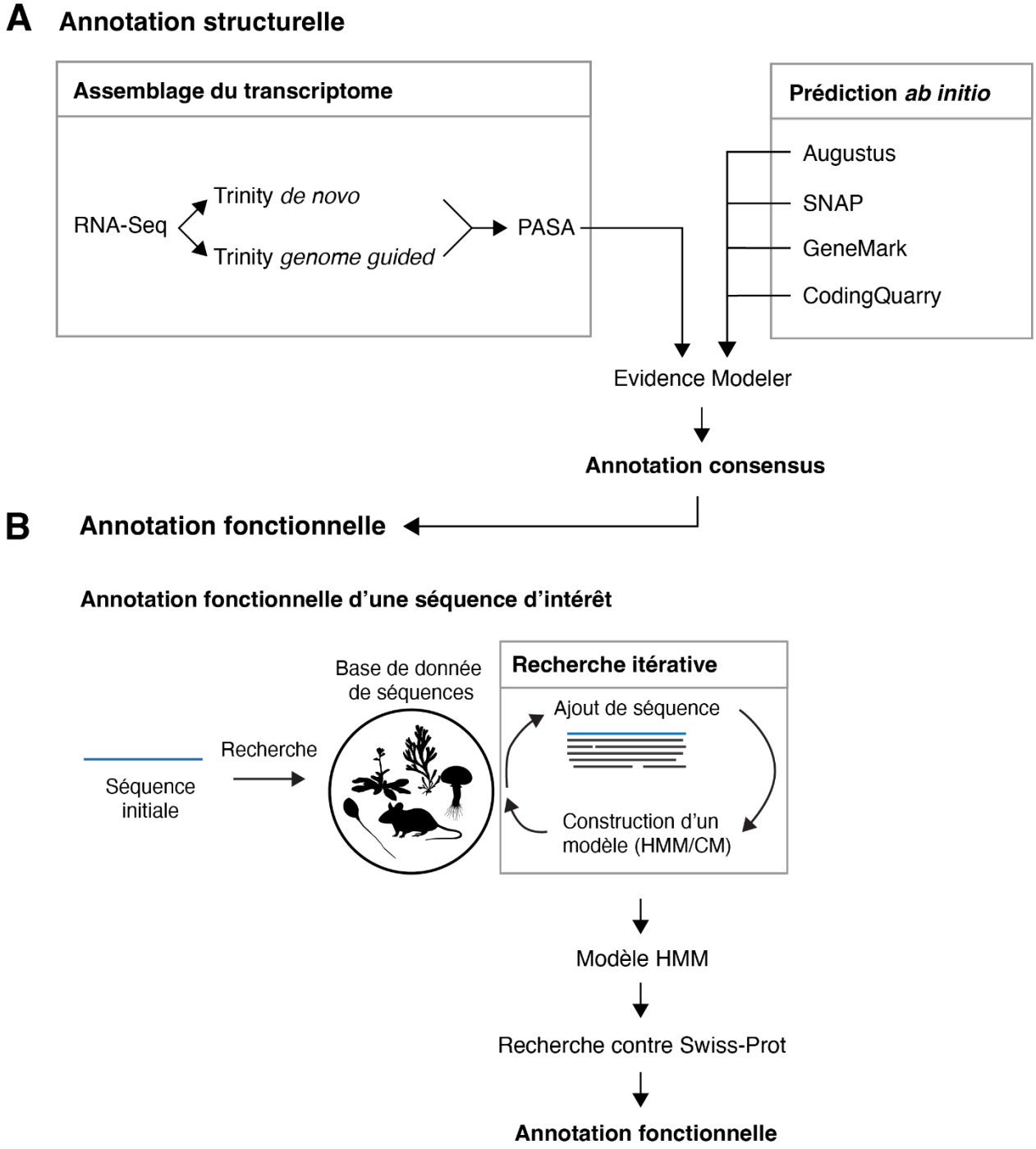

Figure 4. – Sommaire de la méthodologie pour l'annotation du protéome des jakobides et malawimonades. (A) L'annotation fonctionnelle a été effectuée en combinant l'assemblage du transcriptome avec les prédictions *ab initio* avec EvidenceModeler (112). (B) L'annotation fonctionnelle a été effectuée pour chaque protéine d'intérêt au moyen d'un pipeline basé sur

la construction automatique de HMM. L'annotation fonctionnelle était ensuite dérivée de Swiss-Prot.

Annotation fonctionnelle

En raison de l'importante distance évolutive entre les jakobides et les malawimonades et la majorité des eucaryotes séquencés et annotés, un modèle HMM a d'abord été construit pour chaque protéine puis utilisé pour effectuer l'annotation (Figure 4B). Ce modèle HMM a été construit par une recherche itérative contre une collection de protéomes couvrant l'ensemble des eucaryotes (191 organismes) et procaryotes (134 organismes) au moyen de l'outil HMMER (v3.3.2) (116). Une fois le modèle construit, celui-ci a été recherché contre la base de données Swiss-Prot qui contient des séquences protéiques dont l'annotation est validée par des experts (Figure 4B). Afin de standardiser la nomenclature utilisée pour l'annotation fonctionnelle de *A. godoyi* avec celle des principaux organismes modèles (notamment *A. thaliana*, *Homo sapiens* et *S. cerevisiae*), la nomenclature du gène chez ces organismes a ensuite été utilisée pour dériver l'annotation fonctionnelle du gène de *A. godoyi* (Figure 4B).

Analyse des chromosomes circulaires

Ploïdie des chromosomes circulaires

Puisque l'alignement des lectures nanopore contre l'assemblage des chromosomes circulaires semblait suggérer qu'ils ne présentent aucune variation hétérozygote (contrairement aux chromosomes linéaires) la couverture de ces molécules a été comparée à celle des molécules linéaires afin d'évaluer s'ils sont tout de même présents en deux copies par cellules. Pour ce faire, les lectures Illumina ont été alignées contre l'assemblage avec Bowtie2 (v2.4.5). La couverture a ensuite été calculée avec l'outil genomecov de la suite bedtools (v2.30.0). La couverture a ensuite été visualisée dans R (v4.2.0) en utilisant la librairie ggplot2 (v3.3.6).

Origine évolutive des protéines

L'origine évolutive des protéines des chromosomes circulaires a été évaluée à partir des résultats de la recherche avec phmmer contre la collection de protéomes eucaryotes et procaryotes

utilisée pour effectuer l'annotation fonctionnelle. Le meilleur résultat pour chaque organisme (meilleure e-value) a été sélectionné et les séquences protéiques ont été alignées avec MAFFT (v7.490) (117). L'alignement a été validé dans AliView (v2021). Une phylogénie a été construite par *maximum likelihood* avec IQ-TREE (v2.2.0) (118).

Visualisation des annotations

Les annotations structurelles et fonctionnelles pour les chromosomes circulaires ont été visualisée en utilisant Circos (v0.69.9) (119). La position des répétitions inversées a été identifiée à partir de l'alignement NUCmer de chaque chromosome circulaire contre lui-même grâce à l'outil show-coords (v4.0.0) (104).

Comparaison des outils d'assemblage de génomes avec les données nanopore

Origine des données

Les jeux de données utilisés pour la comparaison des outils ont été identifiés dans la base de données NCBI SRA. Pour toutes ces données, des librairies de séquençage par nanopore SQK-LSK108 ou SQK-LSK109 ont été séquencées sur des *flow cell* R9.4.1. Pour *S. cerevisiae*, les données nanopore brutes (format FAST5) de la souche CVy61 (120), une souche dérivée de *S. cerevisiae* W303 (PRJNA730563) ainsi que les données Illumina HiSeq 2000 de *S. cerevisiae* W303 (102 nucléotides, pairées; PRJNA260311) (121) ont été utilisées. Pour *C. elegans*, les données nanopore et Illumina de la souche N2 ont été utilisées. Les données nanopore traitées (*basecalling* par Guppy v3.1.5; modèle *high accuracy*) (122) ont été obtenues de la base de données NCBI SRA (PRJNA562392) puisque les données brutes n'étaient pas disponibles. Les données Illumina pour *C. elegans* N2 ont été obtenues de la base de données CeNDR (123). Les données Illumina pour *A. thaliana* écotype Col-0 ont été obtenues de la base de données NCBI SRA (PRJNA643548). Les données nanopore brutes pour *A. thaliana* proviennent du groupe qui a produit les données Illumina et ont été fournies par Brieanne Vaillancourt du laboratoire de Robin Buell à *Michigan State University*. Le *basecalling* de ces données a été effectué avec Guppy (ONT; v6.1.2) en utilisant le modèle *super high accuracy*. Les génomes de référence ont été obtenus à

partir de la base de données NCBI Assembly pour *S. cerevisiae* W303 (GCA_002163515.1), *C. elegans* N2 (GCF_000002985.6) et *A. thaliana* Col-0 (GCF_000001735.4).

Pré-traitement des lectures nanopore

Basecalling

Les données nanopore de *S. cerevisiae* W303 ont été traitées avec Guppy avec le modèle *super high accuracy* (ONT; v6.1.2), Chiron (v0.6.1) (66), Halcyon (v0.0.1) (124) et DeepNano-blitz (v0.1) en utilisant le mode --network-type 256 pour augmenter la précision (67) (Tableau 1). Les paramètres par défaut ont été utilisés pour l'ensemble des analyses.

Tableau 1. – Sommaire des outils de basecalling

Outils	Description	Version	Référence
Chiron	Cinq couches de réseau neuronal convolutif (CNN) et trois couches de réseau neuronal récurrent (RNN) suivies d'une couche entièrement connectée	0.6.1	(66)
DeepNano-blitz	Réseau neuronal récurrent bidirectionnel (BRNN) optimisé pour la performance	0.1	(67)
Guppy	Réseau de neurones profond (DNN) propriétaire	6.1.2	N/A
Halcyon	Réseau neuronal convolutif (CNN) suivi d'un encodeur et décodeur basé sur un réseau neuronal récurrent (RNN)	0.0.1	(124)

Correction des lectures

Les lectures en format FASTQ produites par Guppy avec le modèle *super high accuracy* (v6.1.2) ont été utilisées pour la comparaison des outils de correction. Ces données ont été corrigées avec deux outils d'auto-correction (i.e., qui effectuent la correction en utilisant exclusivement des données nanopore) (Tableau 2) ainsi que six outils qui effectue une correction hybride et utilisent des données Illumina pour corriger les lectures nanopore (Tableau 3). La correction a été effectuée selon les paramètres et les pipelines par défaut des différents outils.

Tableau 2. – Sommaire des outils d'auto-correction

Outils	Description	Version	Référence
Canu	Alignement tous contre tous des lectures nanopore. Construction d'un graphe acyclique pour chaque lecture à partir de ces chevauchements. Le chemin ayant la pondération la plus élevée est utilisé pour générer le consensus.	2.2	(80)
NECAT	Alignement tous contre tous des lectures nanopore. Les lectures sont corrigées en deux passes. La première corrige les sous-séquences avec une faible fréquence d'erreur et la deuxième les sous-séquences avec une fréquence d'erreur élevée. La correction est effectuée en utilisant le consensus de l'alignement.	0.0.1	(125)

Tableau 3. – Sommaire des outils de correction hybride

Outils	Description	Version	Référence
ECTools	Alignement des lectures nanopore contre les unitigs Illumina avec NUCmer (104). Les alignements sont filtrés pour identifier l'alignement optimal et les différences entre l'unitig et la lecture sont ensuite utilisées pour déterminer la séquence consensus.	0.1	(126)
LoRDEC	Construction d'un graph de Bruijn avec les lectures Illumina. Le graph est ensuite aligné contre les lectures nanopore pour corriger les régions erronées en les remplaçant par la séquence du graph.	0.9	(127)

Jabba	Similaire à LoRDEC, mais basé sur un algorithme *seed-and-extend* utilisant l'homologie exacte maximale pour accélérer la correction.	Commit 02ed64e	(128)
FMLRC2	Similaire à LoRDEC, mais utilise un FM-index pour représenter un graph de Bruijn avec des k -mers de longueur arbitraire. Les lectures nanopore sont corrigées en deux passes en utilisant différentes longueurs de k-mers pour corriger plus efficacement les répétitions de faible complexité.	0.1.7	(129)
HG-CoLoR	Similaire à FMLRC2, mais utilise un graph de Bruijn d'ordre variable pour représenter la séquence Illumina. De plus, les 'seeds' utilisées pour l'alignement du graph par l'algorithme *seed-and-extend* proviennent de l'alignement des lectures Illumina contre les lectures nanopore.	1.1.1	(130)
PBrC	Les lectures Illumina sont alignées contre les lectures nanopore. La séquence consensus est calculée à partir de cet alignement.	8.3	(131)

Comparaison des méthodes de pré-traitement

Le taux d'erreur des lectures a été calculé avec LightQC (v0.1.1) (https://github.com/drs/lightqc) en comparant les lectures au génome de référence de *S. cerevisiae* W303. Les données ont été visualisées dans R (v4.2.0) en utilisant les librairies ggplot2 (v3.3.6), lvplot (v0.2.0) et reshape2 (v1.4.4). Le taux d'erreur de la séquence consensus a été déterminé avec l'outil dnadiff de la suite MUMmer (v4.0.0) (104) en comparant un assemblage Flye (v2.9) (79) construit avec les lectures générées par le *basecalling* ou la correction contre le génome de référence de *S. cerevisiae* W303. La contiguïté de ces assemblages a quant à elle été déterminée avec QUAST (v5.2.0) (93). Pour comparer la vitesse de *basecalling*, les outils ont été utilisés sur le GPU (sauf DeepNano qui ne supporte que le CPU) et la vitesse des outils a été mesurée en cinq réplicas sur un poste Ubuntu

22.04 avec 192Gb de RAM, deux processeurs Xeon E5-2680V4, une carte graphique NVIDIA GTX 1080 et un disque NVME de 512 GB.

Assemblage des génomes

Les génomes de *Arabidopsis thaliana*, *Caenorhabditis elegans* et *Saccharomyces cerevisiae* ont été assemblés avec neuf assembleurs (Tableau 4). Pour *A. thaliana* et *S. cerevisiae*, les données nanopore traitées avec Guppy (v6.1.2) en utilisant le modèle *super high accuracy* ont été utilisées alors que celles disponibles sur NCBI ont été utilisées pour *C. elegans*. Les données nanopore de *A. thaliana* et *S. cerevisiae* ont été sous-échantillonnées aléatoirement avec l'outil fastq-sample de la collection fastq-tools (https://github.com/dcjones/fastq-tools) pour obtenir une couverture d'environ 100X. Pour *C. elegans* l'ensemble des données a été utilisé (couverture d'environ 30X). L'assemblage a été effectué en utilisant les options par défaut des assembleurs, à l'exception des options spécifiques aux jeux de données (tel que la taille du génome ou les paramètres spécifiques spécifiant le *basecalling* des données nanopore) qui ont été spécifiés selon l'organisme et le jeu de données.

Tableau 4. – Sommaire des outils d'assemblage

Outils	Description	Version	Référence
Canu	Alignement tous contre tous des lectures nanopore pour effectuer la correction et le clivage des lectures (retire les segments erronés) avant de calculer le chevauchement des lectures. Construction d'un « *best overlap graph* » pour assembler le génome.	2.2	(80)
Flye	Construction d'un *repeat graph* avec les lectures nanopore et résolution de ce graph pour définir la structure du génome. La séquence consensus est ensuite calculée pour l'assemblage final.	2.9	(79)
HASLR	Construction de contigs avec des lectures Illumina et nettoyage des contigs provenant de répétitions.	0.8	(86)

	Alignement des lectures nanopore sur les contigs et résolution de la structure génomique. La séquence consensus est obtenue par *partial order alignment*.		
MaSuRCA	Assemblage des lectures Illumina pour obtenir des *super-reads*. Assemblage du génome en combinant les *super-reads* avec les lectures nanopore.	4.0.9	(78)
miniasm+Racon	Alignement tous contre tous avec minimap2 (102) suivit de la résolution de la structure génomique par un algorithme « *Overlap-Layout-Consensus* ». Racon est ensuite utilisé pour produire la séquence consensus.	0.3	(99,132)
NECAT	Similaire à Canu, effectue la correction et le nettoyage des lectures nanopore avant l'assemblage. L'assemblage est effectué par la résolution d'un « *directed string graph* ».	0.0.1	(125)
Raven	Similaire à miniasm, mais commence par l'identification de plus longues lectures non redondantes pour accélérer la construction du graph d'assemblage. La librairie Racon (132) est utilisée pour produire l'assemblage consensus.	1.8.1	(87)
Shasta	Afin d'augmenter la vitesse de calcul, les lectures nanopore sont représentées par codage de plage en utilisant des marqueurs (ensemble de k-mers prédéterminé). L'assemblage est obtenu par la résolution d'un « *repeat graph* ». La séquence consensus est obtenue avec MarginPolish et un modèle Bayesian pour déterminer la longueur des homopolymères.	0.10.0	(88)

wtdbg2	Utilise une méthode « Overlap-Layout-Consensus » avec un algorithme de disposition basé sur les « fuzzy-Bruijn graph ».	2.5	(52)

La contiguïté des assemblages a été validée par QUAST (v5.2.0) (93). La présence des protéines représentatives de ces organismes a été évaluée avec BUSCO (v5.2.1) (95) en utilisant l'ensemble de protéine le plus spécifique possible, soit Saccharomycetes pour *S. cerevisiae*, Nematoda pour *C. elegans* et Brassicales pour *A. thaliana*. Afin de comparer la qualité des assemblages produits par les différents assembleurs, la fraction du génome de référence couverte par l'assemblage nanopore, le N50 de l'assemblage nanopore et le nombre d'erreur d'assemblage ont été extraits du rapport de QUAST. Le nombre de gènes BUSCO complets, fragmentés et manquants ont quant à eux été extraits du rapport BUSCO. Afin de permettre la comparaison de la performance des assembleurs entre les différentes espèces et d'identifier des tendances dans la performance, le Z-score de la métrique a été calculé indépendamment pour chaque métrique et chaque organisme. Ces résultats ont été visualisés dans R (v4.2.0) en utilisant la librairie ggplot2 (v3.3.6).

Afin de visualiser les différences structurelles majeures entre les organismes, les génomes des organismes ont été comparés en utilisant l'outil Circos-WGC (v0.1.0) (https://github.com/drs/Circos-WGC). Brièvement, cet outil utilise l'alignement des génomes produit par minimap2 en utilisant le préréglage asm5 (v2.24) (99) pour produire une visualisation des alignements entre les génomes avec Circos (v0.69.9) (119).

Polissage avec les données Illumina

L'assemblage du génome de *S. cerevisiae* produit avec Flye a été polis avec les données Illumina en utilisant six outils de polissage (Tableau 5). Pour tous les outils, à l'exception de ntEdit et POLCA qui utilisent les fichiers FASTQ directement, les lectures Illumina ont été alignées contre l'assemblage nanopore avec Bowtie2 (v2.4.5) (133). Les paramètres par défaut ont été utilisés pour l'ensemble des outils. Les assemblages polis ont ensuite été alignés contre l'assemblage de référence avec l'outil NUCmer (v4.0.0) puis la position des différences a été déterminée avec le

script show-snps de la suite MUMmer (v4.0.0) (104). Les données ont été visualisées dans R (v4.2.0) en utilisant la librairie ggplot2 (v3.3.6).

Tableau 5. – Sommaire des outils de polissage

Outils	Description	Version	Référence
Apollo	Construction et entraînement d'un graph de profile de modèle caché de Markov à partir de l'alignement des lectures contre le contig qui est ensuite utilisé pour corriger la séquence.	Commit 85a4ca3	(134)
Racon	Consensus dérivé à partir d'un graph de « *Partial Order Alignment* » (POA) construit à avec l'alignement des lectures sur l'assemblage.	1.4.3	(132)
HyPo	Similaire à Racon, mais sépare l'assemblage en régions robuste (sans erreurs) et régions faibles. Seules les régions faibles sont polies par POA pour accélérer la correction.	1.0.3	(85)
Pilon	Consensus construit en identifiant la base la plus probable à chaque position en utilisant l'information contenue dans les colonnes de l'alignement des lectures Illumina contre l'assemblage.	1.24	(84)
POLCA	Les variants identifiés avec FreeBayes à partir de l'alignement des lectures Illumina contre l'assemblage sont utilisés pour générer la séquence consensus.	4.0.9	(82)
ntEdit	Balayage des k-mers dans l'assemblage draft et valide la présence dans un Bloom Filter contenant les k-mers des lectures Illumina. Si un k-mer est absent, il est probablement erroné et la séquence	1.3.5	(135)

est corrigée avec le k-mer le plus probable selon les données Illumina.

Chapitre 3 – Résultats et discussion

Séquençage génomique des jakobides et malawimonades

Séquençage nanopore des jakobides et malawimonades

La diminution des coûts de séquençage grâce au développement des méthodes de séquençage de seconde génération (*e.g.* Illumina) a rendu plus accessible l'assemblage *de novo* de génomes pour une large variété de projets (136), incluant un premier séquençage des jakobides et malawimonades. En effet, ces technologies permettent de produire un nombre important de lectures courtes ce qui permet de produire rapidement des génomes « *draft* ». Cependant, ces séquences courtes ne permettent pas d'assembler de longues régions répétés (*e.g.* satellites ou transposons) ce qui cause une importante fragmentation des génomes de eucaryotes dont les génomes sont riches en régions répétées en plus d'induire des erreurs difficiles à corriger dans les assemblages tel que des compressions des répétitions ou des *mis-assembly* (137,138). Les limitations de ces technologies augmentent le nombre d'erreurs dans les modèles de gènes (gènes fragmentés, incorrects ou incomplets) qui sont obtenus à partir des assemblages.

Les premiers assemblages *de novo* des génomes de jakobides et malawimonades obtenus en utilisant des données Illumina sont particulièrement fragmentés (entre 2911 contigs pour *M. jakobiformis* et 37152 contigs pour *J. libera*) ce qui affecte les analyses pouvant être effectuées sur ces génomes. L'amélioration de la contigüité de l'assemblage permettrait d'améliorer la qualité de modèles de gènes ainsi que l'ensemble des analyses subséquentes effectuées sur les génomes. De plus, les cultures de ces protistes contiennent un nombre important d'espèces de bactéries qui ne sont pas toutes identifiées et dont les séquences doivent être retirées de l'assemblage final. Cependant, certaines courtes séquences ne contenant pas de marqueurs bactériens peuvent être difficile à retirer en raison l'importante fragmentation du génome et certains fragments de génome bactérien risquent d'être inclus dans l'assemblage final et considérés comme des séquences de protistes.

Les développements récents des techniques de séquençage de 3e génération (*i.e.* nanopore et PacBio) ont rendu possible le séquençage à haut débit de long fragment d'ADN. Ces longs fragments permettent l'assemblage de région répétées de plusieurs kilobases augmentant ainsi largement la contiguïté des assemblages *de novo* générés avec ces données en plus de faciliter l'identification et le retrait de séquences contaminantes bactériennes. En combinant ces données avec des données de séquençage Illumina, il est donc possible d'obtenir des assemblages génomiques de très haute qualité.

Afin de complémenter les données Illumina déjà produites, l'ADN génomique des jakobides *A. godoyi*, *Stygiella incarcerata, J. bahamiensis, R. americana* et *S. ecuadoriensis* ainsi que du malawimonade *M. californiana* a été séquencé par nanopore. Pour la majorité des organismes (à l'exception de *J. bahamiensis* et *M. californiana*) une quantité suffisante de données a été produite pour permettre l'assemblage du génome (> 40X de couverture) (Tableau 6). De plus, le taux d'erreur pour *A. godoyi*, *J. bahamiensis*, *M. californiana* et *S. incarcerata* est similaire aux estimations de Oxford nanopore Technology (3-5%) suggérant l'absence de biais majeur chez ces groupes de protistes qui pourrait affecter la qualité des séquences (notamment GC élevé ou modification de bases (68,139)). Le taux d'erreur est cependant plus élevé pour *R. americana* et *S. incarcerata.* Puisque ces organismes ont un contenu en GC neutre (43% pour *R. americana* et 42% pour *S. incarcerata*), il est possible que ce taux d'erreur plus élevé soit causé par la présence de modifications de bases. Dans ce cas, il serait possible de réduire le taux d'erreur en entrainant des modèles de *basecalling* spécifiques pour ces organismes.

Tableau 6. – Métriques des séquençages nanopore de jakobides et malawimonades.

Organisme	Rendement (Gb)	NR50	Taux d'erreur (%)
A. godoyi	5.9[†]	30941	4.0
J. bahamiensis	0.9	8400	4.8
M. californiana	1.1	33596	4.5
R. americana	10.7	28291	8.5
S. ecuadoriensis	9.2	26885	3.9
S. incarcerata	15.8	34123	6.8

[†] Le séquençage de *A. godoyi* a été arrêté lorsqu'une quantité suffisante de données a été obtenu.

La quantité de données obtenues pour *J. bahamiensis* et *M. californiana* est largement inférieure à celle obtenue pour les autres protistes (Tableau 6). Cette différence de rendement s'explique principalement par une diminution plus rapide du nombre de pores disponibles lors du séquençage de *J. bahamiensis* et *M. californiana* que lors de celui des autres protistes (Figure 5). Cette différence ne semble pas être associée à une différence majeure dans la qualité des ADN purifiés, puisqu'aucune différence n'a été notée lors du contrôle qualité des échantillons par NanoDrop. De plus, cette diminution est liée à une augmentation du nombre de pores bloqués en état « *unavailable* » (Figure 5), un état dans lequel les pores sont obstrués et non disponibles pour le séquençage. Cet état diffères des autres états de pores non fonctionnels (ex. *multiple, saturated* ou *zero*), puisque les pores *unavailable* être restaurés à la suite d'un nettoyage de la *flow cell* (Figure 6) (140). Cependant, Oxford Nanopore Technology (ONT) ne fournit que peu d'information à ce sujet, il est donc difficile d'identifier avec certitude la cause du blocage des pores. L'information structurelle sur les pores ainsi que les résultats obtenus nous permettent cependant d'identifier certaines causes potentielles qui devront être investiguées.

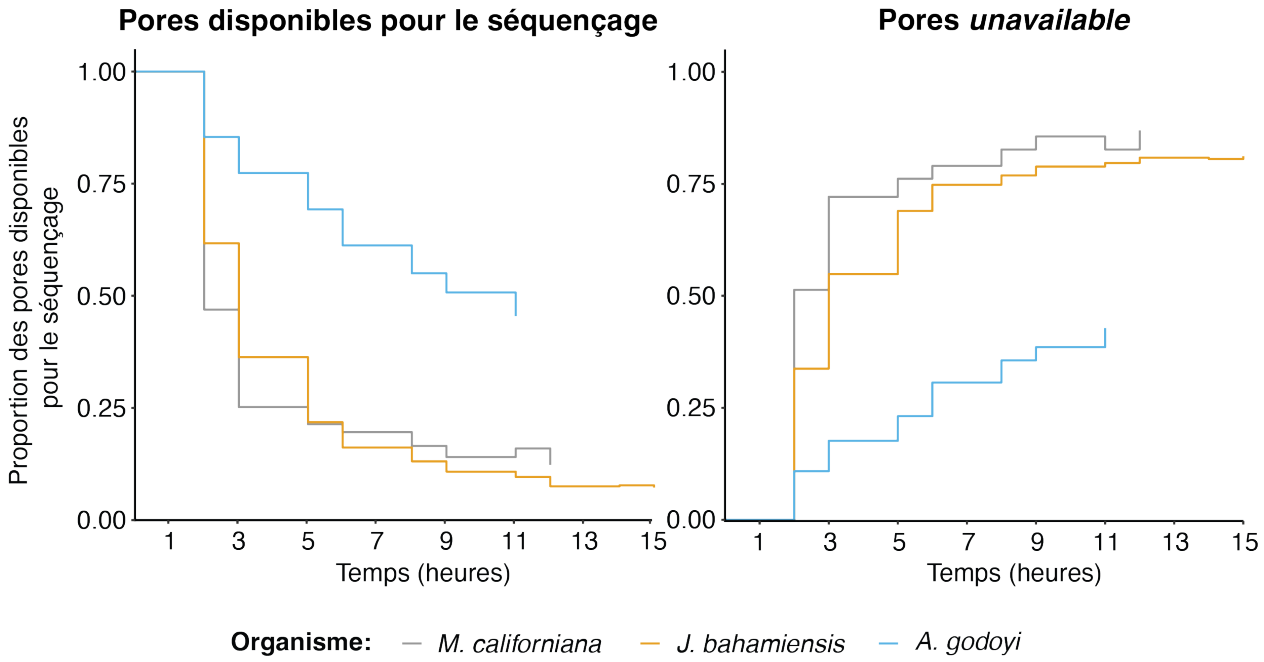

Figure 5. – Diminution rapide du nombre de pores disponibles pour le séquençage nanopore de *J. bahamiensis* et *M. californiana*. Les ADN de *A. godoyi*, *J. bahamiensis* et *M. californiana* ont été extraits par lyse au SDS, dialyse, et suivie d'une colonne par gravité. Le nombre de pores disponible au début du séquençage était de 1220 pour *A. godoyi*, 1008 pour *J. bahamiensis* et 1140 pour *M. californiana*. L'état des pores au cours du séquençage a été obtenu à partir des données de *mux scan* (une validation de l'état des pores effectuée à chaque 1,5h lors du séquençage) produite par MinKNOW. La proportion de pores disponibles a été déterminée à partir du nombre de pores en état *unavailable* lors du premier *mux scan*.

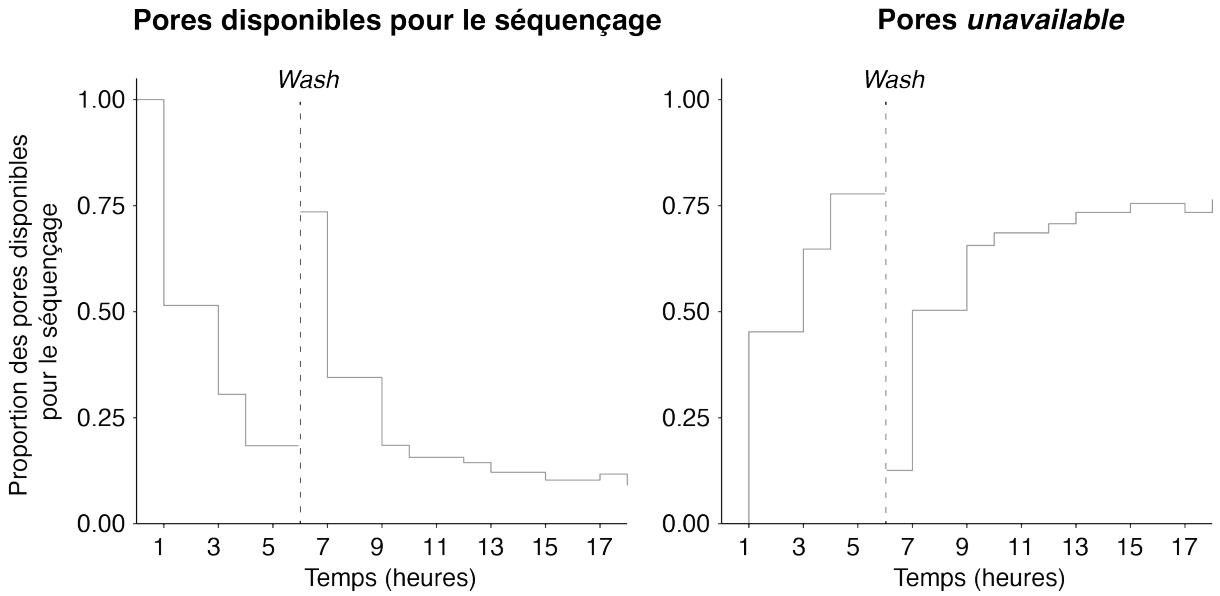

Figure 6. – Le nettoyage de la *flow cell* permet de restaurer les pores en état unavailable. L'état des pores au cours du séquençage de *M. californiana* a été obtenu à partir des données de *mux scan* produites par MinKNOW. Une première librairie a été séquencée pendant 6 heures avant de nettoyer la *flow cell* et de procéder au séquençage d'une deuxième librairie.

Le complexe de séquençage est composé d'un enzyme moteur (hélicase) qui permet la translocation de l'ADN simple brin et d'une protéine pore (dérivée de la protéine CsgG de *Escherichia coli*) qui permet la lecture du signal (Figure 7A) (141,142). Puisque l'état *unavailable* est causé par une obstruction d'un pore, nous avons posé l'hypothèse que le blocage rapide des pores pourrait être causé par la présence de régions génomiques chez ces espèces qui forment des structures secondaires dans le tunnel du pore et bloque le passage de l'ADN (Figure 7B) ou par la présence d'un contaminant co-purifié avec l'ADN (Figure 7C). En effet, en raison de la purification exhaustive effectuée sur l'ADN, il est peu probable qu'un contaminant provenant des cultures et qui affecte le séquençage soit libre dans le milieu (Figure 7C). Il est donc nécessaire que le contaminant forme une interaction suffisamment forte avec l'ADN lors de la purification pour être co-purifié. Nous croyons donc que ce contaminant pourrait être un polysaccharide cationique capable d'interagir avec l'ADN lors de l'extraction. La présence de tels contaminants a

déjà été suggérée chez certaines espèces de plantes (*e.g. Chenopodium quinoa*), et associée à un blocage prématuré des pores lors du séquençage nanopore (143).

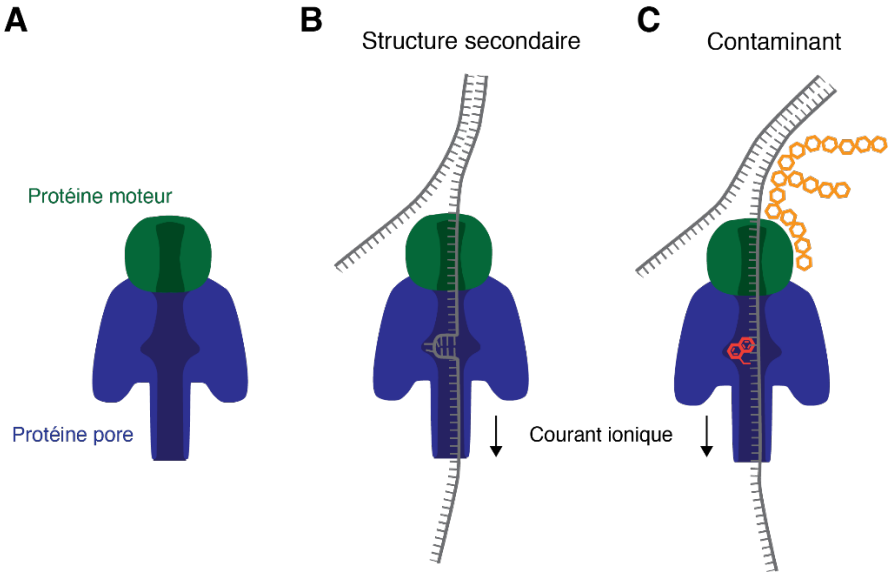

Figure 7. – Structure du complexe de séquençage et des éléments pouvant bloquer le canal. Le complexe de séquençage est composé d'une protéine moteur (en vert) et d'une protéine pore (en bleu). B) La formation de structure secondaire dans le canal peut bloquer le nanopore et empêcher la translocation de l'ADN (en gris). C) La présence de contaminant sur l'ADN (en orange) ou dans le canal (en rouge) pourrait aussi bloquer le séquençage nanopore.

Optimisation du protocole d'extraction avec *Malawimonas californiana*

Le rendement de séquençage nanopore des protistes *M. californiana* et *J. bahamiensis* étant trop faible pour pouvoir assembler le génome à partir de ces données, il est nécessaire d'identifier et d'éliminer la cause du blocage des pores. Pour ce faire, l'extraction de l'ADN de *M. californiana* a été optimisée afin de réduire les contaminants liés à l'ADN qui pourrait affecter le séquençage. Puisque nous croyons que le contaminant qui affecte le séquençage est un polysaccharide, une étape d'extraction au CTAB en présence d'une haute concentration de sel, qui permet de séparer l'ADN des polysaccharides en solution (144,145), a été ajoutée à la purification. L'extraction au CTAB a permis d'augmenter de 2 fois la durée de vie médiane des pores (5.2h) en comparaison à l'ADN obtenu sans extraction au CTAB (2.6h) (Figure 8). Cette différence permet de suggérer que

le contaminant qui affecte le séquençage est en fait un polysaccharide puisque l'extraction au CTAB permet de diminuer la vitesse de blocage des pores. De plus, il est peu probable que la cause du blocage soit la formation de structure secondaire dans le pore étant donné que la méthode d'extraction ne devrait avoir aucun impact sur la formation de ces structures.

Cependant, le rendement du séquençage est similaire entre les deux méthodes d'extraction (Tableau 7). Cette différence pourrait être attribuable au fait que les lectures des extractions au CTAB sont plus courtes et donc que le nombre de molécules séquencées est environ 2 fois plus important pour les ADN extraits au CTAB, ce qui augmente le risque qu'une molécule d'ADN lié à un polysaccharide traverse le pore (Tableau 7). Quoi qu'il en soit, le rendement du séquençage reste trop faible pour pouvoir être utilisé pour le séquençage de *J. bahamiensis* et *M. californiana*.

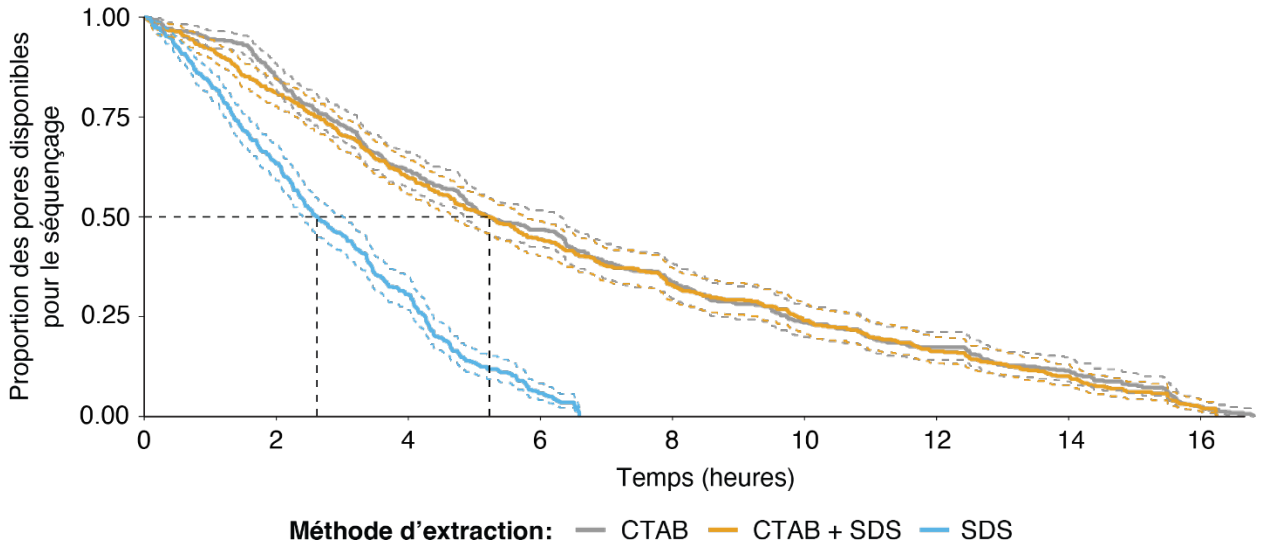

Figure 8. – L'extraction de l'ADN de *M. californiana* au CTAB permet d'augmenter la durée de vie des pores de la *flow cell.* L'ADN de *M. californiana* a été extrait selon le protocole de lyse au SDS, au CTAB ou au CTAB suivit d'une deuxième digestion à la protéinase K dans du SDS. Les ADN ont été séquencés sur un *flow cell* MinION nettoyée entre chaque séquençage. La durée de vie moyenne des pores (représentée par la ligne pointillée noire) est similaire pour les deux extractions par lyse au CTAB et CTAB puis SDS (5.2h; $p = 0.36$; test Kaplan-Meier) alors qu'elle est significativement plus faible pour l'ADN extrait par lyse au SDS (2.6h; $p > 0.0001$; test Kaplan-Meier).

Tableau 7. – Rendement du séquençage de l'ADN de *M. californiana* extrait par lyse au CTAB et au SDS.

Méthode d'extraction de l'ADN	**Rendement du séquençage (Gb)**	**Nombre de lectures**	**Longueur médiane des lectures**
SDS	0,85	65 312	13 039
CTAB	0,80	125 760	6 393
CTAB + SDS	1,49	150 075	9 953

En effet, malgré l'amélioration de la durée de vie des pores observée après l'extraction au CTAB, le rendement du séquençage reste largement inférieur à celui des autres protistes. Plus de développement sera donc nécessaire afin d'identifier une méthode d'extraction de l'ADN qui permet de purifier l'ADN de *M. californiana* suffisamment pour le séquençage nanopore. Puisque les résultats obtenus suggèrent que le contaminant est un polysaccharide, les prochaines optimisations viseront à mieux séparer d'ADN des polysaccharides. Pour ce faire, il serait possible de mieux séparer les cellules de protistes du milieu (contenant des bactéries et des exo-polysaccharides) ou d'évaluer d'autres méthodes permettant de séparer les polysaccharides de l'ADN (*e.g.* séparation de l'ADN par centrifugation différentielle). De plus, il serait possible d'évaluer l'impact de méthodes d'extraction de l'ADN sans précipitation (e.g. *NanoBind HMW DNA Extraction Kit* ou *Monarch® HMW DNA Extraction Kit*). En effet, les polysaccharides qui n'auraient pas été éliminés précipitent avec l'ADN en présence d'éthanol ou d'isopropanol avec lequel ils risquent de former des interactions lors de la précipitation. L'absence de précipitation pourrait donc diminuer le risque que des interactions se forment entre les polysaccharides et l'ADN.

Construction d'un pipeline pour l'assemblage *de novo* de génomes avec les données nanopore

La technologie de séquençage nanopore ainsi que les outils d'analyse en aval évoluent rapidement. En raison de cette évolution rapide, l'identification des outils les plus performants

pour effectuer les différentes étapes nécessaires pour produire un assemblage avec des données nanopore (i.e. *basecalling*, assemblage et polissage) est difficile. En effet, de nouveaux outils sont développés rapidement et les outils existants sont fréquemment mis à jour, il existe donc peu de comparaisons d'outils publiées pour aider la construction d'un pipeline d'assemblage. De plus, la majorité des comparaisons existantes sont spécifiques aux procaryotes (89), n'incluent pas les outils les plus récents (68) ou encore sont publiées avec de nouveaux outils et n'offrent pas nécessairement de revue systématique et indépendantes de l'ensemble des outils existants au moment de leur publication (86,124). L'identification d'outils performants est donc une étape essentielle au développement d'un pipeline performant d'assemblage des génomes eucaryotes.

Comparaison des outils de pré-traitement et d'assemblage avec les données nanopore

L'assemblage des génomes au moyen des données nanopore requiert trois étapes principales, soit le *basecalling* des lectures, l'assemblage du génome et le polissage avec des données de séquençages Illumina. De plus, dans certain cas, il est possible d'ajouter une étape de correction des lectures nanopore avant l'assemblage afin de diminuer leur taux d'erreur. Afin de déterminer les outils les plus performants, différents outils pour chacun de ces étapes ont été comparés en utilisant des données de *Arabidopsis thaliana* écotype Columbia, *Caenorhabditis elegans* N2 et *Saccharomyces cerevisiae* W303. Ces organismes ont été sélectionnés puisqu'ils possèdent un génome de référence de haute qualité contre lesquels peuvent être comparés les assemblages produits par les outils comparés.

Comparaison des outils de *basecalling*

Le *basecalling* est la première étape de traitement des données nanopore et permet de convertir le signal en séquence d'acides nucléiques, qui peut être utilisée pour les étapes d'analyse en aval. Oxford Nanopore Technologies fournit l'outil Guppy pour effectuer cette étape, mais plusieurs autres outils ont été développés par des tierces personnes. Ces outils utilisent des modèles de réseau de neurones différents de celui intégré dans Guppy (un réseau de neurone récurrents) et visent à pallier les limitations de ce modèle afin d'améliorer la qualité des séquences produites.

Cependant, la dernière comparaison d'outils de *basecalling* ayant été publiée en 2019 (68), la performance de nouveaux outils n'a pas été évaluée de manière indépendante dans la littérature.

La performance des outils de *basecalling* Chiron, DeepNano-blitz, Guppy et Halcyon a donc été évaluée ici par rapport à l'identité des lectures et à la vitesse de traitement des données en utilisant les données de *S. cerevisiae*. L'identité des lectures a été calculée à partir de leur alignement contre le génome de référence. Le modèle *super high accuracy* de Guppy produit les lectures avec le taux d'erreur le plus faible suivit de DeepNano-blitz (Figure 9A). De plus, Guppy est 2 fois plus rapide que DeepNano-blitz et Chiron et environ 30 fois plus rapide qu'Halcyon (Figure 9B). Dans l'ensemble, ces résultats indiquent que Guppy est à la fois plus rapide et produit des séquences avec moins d'erreurs que les autres outils comparés.

La différence de performance entre les outils est similaire à ce qui a été rapporté précédemment (68) et s'explique partiellement par le fait que Guppy est fréquemment mis à jour par Oxford Nanopore Technologies, alors que Chiron et Halcyon ont été développés comme preuve de concept pour des algorithmes de *basecalling* et ils sont plus rarement mis à jour (66,124). De plus, le modèle DeepNano-blitz est quant à lui développé pour être performant sur un CPU au détriment de l'identité des séquences, et présente des performances similaire à ce qui a été rapporté précédemment (67).

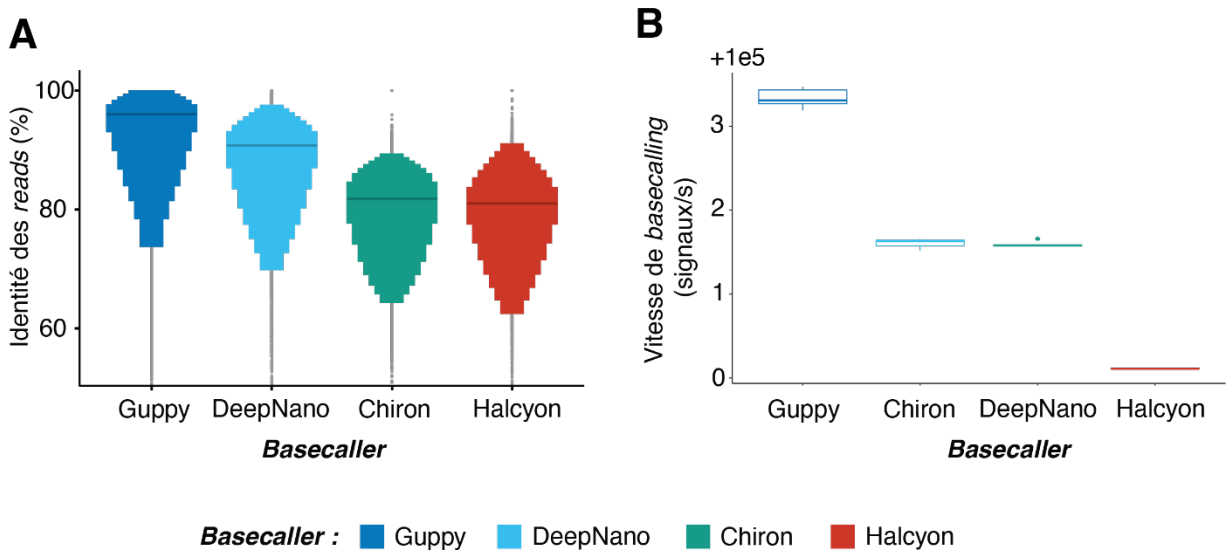

Figure 9. – Comparaison de la performance des outils de *basecalling*. Les données nanopore de *S. cerevisiae* ont été traitées avec Guppy (modèle super high accuracy), DeepNano (--network-type 256), Chiron et Halcyon. La distribution de l'identité des lectures a été déterminée par l'alignement des séquences nanopore contre le génome de référence et est représentée par un graphique *letter-value*. La barre foncée représente la médiane (Guppy : 96%, DeepNano : 91%, Chiron : 82% et Halcyon : 81%). B) La vitesse de *basecalling* a été mesurée en utilisant le GPU (sauf pour DeepNano) sur un sous-ensemble de données nanopore.

Comparaison des outils de correction des lectures nanopore

La correction des lectures est un processus qui permet de diminuer le nombre d'erreurs dans les séquences brutes nanopore. Cette étape permet de simplifier la structure du génome pour les assembleurs, en réduisant la variance dans les lectures, ce qui améliore la qualité de l'alignement des lectures entre eux et simplifie la disposition des lectures dans les graphs d'assemblage (127,129,146,147). Certains assembleurs bénéficient de cette simplification et ont intégré leur propre algorithme de correction dans le pipeline d'assemblage (*e.g.,* Canu et NECAT) (80,125), mais la majorité des assembleurs ne l'effectuent pas et peuvent donc potentiellement bénéficier d'une pré-correction des lectures (90).

Afin d'évaluer la performance des outils de correction des lectures nanopore, la capacité des outils disponibles a d'abord été évaluée. Pour ce faire, les lectures nanopore de *S. cerevisiae* W303 ont été corrigés avec les outils Canu (80), ECTools (126), FMLRC2 (129), HG-CoLoR , Jabba (128), LoRDEC (127), NECAT (125) et PBrC. Le taux d'erreurs dans les lectures a ensuite été déterminé par LightQC à partir de l'alignement des lectures contre le génome de référence. Pour l'ensemble des outils, la correction diminue de manière importante le taux d'erreur dans les lectures nanopore (Figure 10, Tableau 8). De plus, le NR50 (et par conséquent la longueur des lectures) est généralement similaire à celui des données brutes, à l'exception des outils Jabba et PBrC, pour lesquels le N50 diminue de manière importante (Tableau 8). De plus, la majorité des outils (à l'exception de Jabba et FMLRC2) augmente la fraction des lectures alignées sur le génome (Tableau 8). Cependant, l'augmentation du taux d'alignement reste marginale puisque 99% des lectures sont déjà alignées contre le génome avant la correction. Bien que la correction améliore l'alignement des lectures contre le génome, elle a donc peu d'impact sur le taux d'alignement de l'ensemble du jeu de données. Ces résultats contrastent avec ceux de comparaison d'outils de correction effectuée sur d'anciennes technologies de séquençage nanopore (R7 ou R9) ou PacBio qui présente une plus importante augmentation du taux d'alignement (d'environ 75% avant la correction à 99% après correction) (90,127). Cette différence est possiblement attribuable à la diminution de taux d'erreurs dans les lectures nanopore produites par les nouvelles générations de *flow cells* ainsi qu'à l'amélioration de l'outil de *basecalling* Guppy. Puisque l'intérêt principal de la correction des lectures est d'améliorer l'alignement (127,129,146,147), cette étape semble donc moins importante pour des données produites avec les technologies de séquençages nanopore plus récentes.

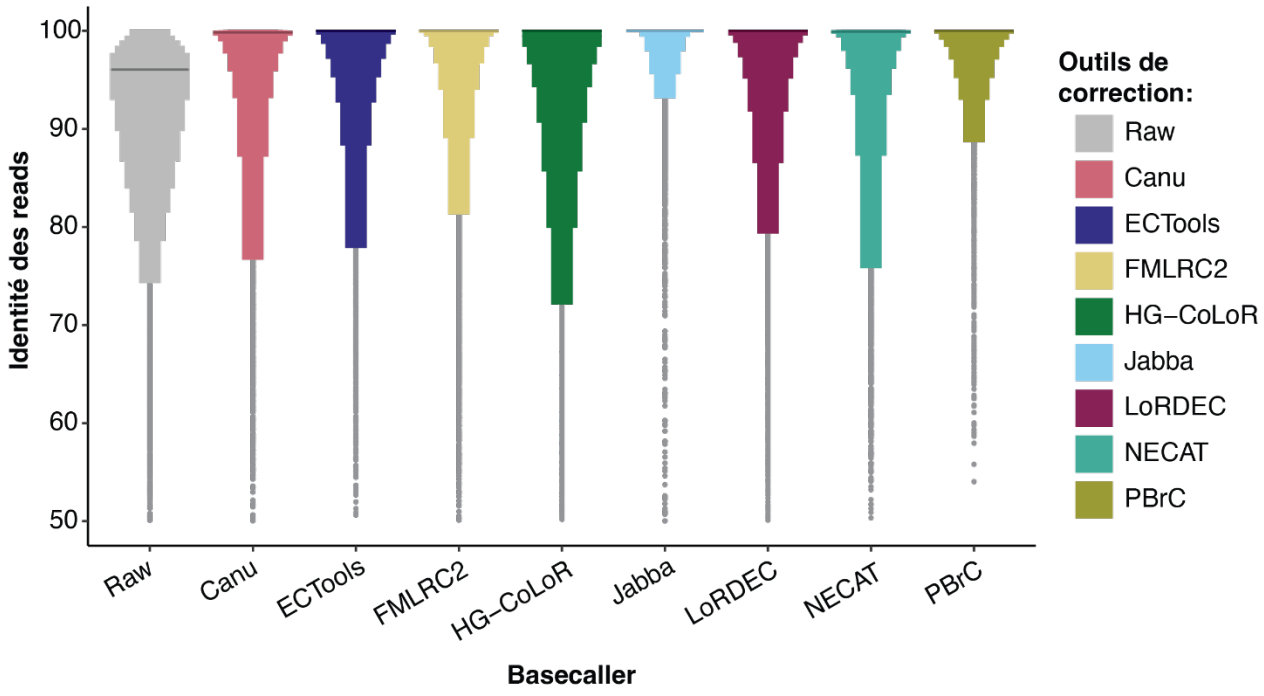

Figure 10. – Taux d'erreurs des lectures nanopore après correction. La distribution de l'identité des lectures avant et après correction a été déterminée par l'alignement des séquences nanopore contre le génome de référence et est représentée par un graphique *letter-value*. La barre foncée représente la médiane (valeurs dans le Tableau 8).

Tableau 8. – Impact de la correction sur la qualité des *reads* nanopore

Outils de correction	NR50 (Kb)	Fraction des lectures alignés	Identité médiane des lectures	Nombre de lectures
Données brutes	18,1	99,05	96,05	252 310
Canu	18,5	99,98	99,83	137 969
ECTools	18,3	100,0	99,99	109 344
FMLRC2	18,3	98,94	100,0	252 310
HG-CoLoR	20,0	99,93	99,45	225 411
Jabba	10,4	99,00	100,0	340 623
LoRDEC	18,3	99,35	100,0	252 310
NECAT	18,7	99,99	99,94	108 583
PBrC	16,3	99,19	100,0	166 506

Puisque l'intérêt principal de la correction des lectures est d'utiliser les données corrigées pour l'assemblage de génome, l'impact de cette étape sur la contiguïté et le taux d'erreur de l'assemblage *de novo* du génome de *S. cerevisiae* a ensuite été évalué. Pour ce faire, trois outils d'assemblage utilisant un algorithme différent (Flye, qui construit et résout un *repeat graph* construit à partir des lectures nanopore; Raven, utilisant une approche *Overlap-Layout-Consensus* et wtdbg2 basé sur un *fuzzy de Bruijn graph*) ont été utilisés pour assembler le génome avec les données corrigées.

La contiguïté des assemblages a été évaluée avec l'outils QUAST (93). À l'exception de ECTools, Jabba et PBrC pour lesquels les assemblages sont plus fragmentés, l'ensemble des outils de correction permettent de produire un assemblage avec une contiguïté similaire à celui produit avec les données nanopore non corrigées (Tableau 9). Cependant, cette comparaison ne permet pas de déterminer si la correction améliore la contiguïté par rapport au données brutes puisque l'ensemble des chromosomes de *S. cerevisiae* peuvent être assemblés en un seul contig à partir des données brutes. Afin de déterminer si la correction peut améliorer la contiguïté, les données nanopore de *C. elegans* ont été corrigées avec les outils Canu, NECAT, FMLRC2 et LoRDEC

(sélectionnés pour la qualité de la correction et leur rapidité d'exécution). Les résultats obtenus suggèrent que, dans certains cas, la correction d'erreur améliore la contiguïté des assemblages (Tableau 10), ce qui est similaire aux résultats obtenus par d'autres groupes (90).

Tableau 9. – Contiguïté de l'assemblage du génome de *S. cerevisiae* générés avec les données nanopore corrigées et non corrigées

Outils	**N50 (Kb)†**		
	Flye	**Raven**	**wtdbg2**
Référence	929	929	929
Données brutes	823	824	801
Canu	830	823	801
ECTools	751	777	719
FMLRC2	813	824	925
HG-CoLoR	816	618	799
Jabba	22	38	25
LoRDEC	811	947	802
NECAT	947	947	935
PBrC	317	230	231

† Le N50 du génome de référence est de 929 Kb.

Tableau 10. – Contiguïté de l'assemblage du génome de *C. elegans* générés avec les données nanopore corrigées et non corrigées

Outils	**N50 (Mb)†**		
	Flye	**Raven**	**wtdbg2**
Données brutes	4,3	5,0	1,5
Canu	5,1	4,4	3,5
FMLRC2	2,7	3,5	2,3
LoRDEC	3,5	4,2	2,4
NECAT	3,6	3,0	2,8

† Le N50 du génome de référence est de 17,5 Mb.

Afin de déterminer si la correction d'erreurs a un impact sur la fréquence d'erreurs dans l'assemblage, le taux de petites erreurs (variation nucléotidique et indels) des assemblages produits avec les données corrigées a été déterminé par l'alignement des contigs contre le génome de référence de *S. cerevisiae* W303 avec NUCmer (104). De plus, les assemblages ont été corrigés par polissage afin de comparer l'impact de la correction des lectures avec celui du polissage. Dans l'ensemble, les méthodes de correction hybride permettent de produire des génomes avec un nombre d'erreur similaire avant et après polissage suggérant que la correction hybride pourrait remplacer le polissage et est suffisante pour produire des assemblages avec un faible nombre d'erreur (Tableau 11). Cependant, puisque la correction d'erreur demande considérablement plus de temps de calcul que le polissage il n'est pas justifiée de l'utiliser uniquement pour réduire le taux d'erreur de l'assemblage final. Les méthodes de *self-correction* quant à elles ne permettent pas de produire des assemblages sans erreurs et nécessite une étape de polissage (Tableau 11). De plus, aucune méthode de semble avoir d'impact négatif sur l'assemblage après polissage (Tableau 11), indiquant que la correction d'erreur n'a pas (ou peu) d'impact négatif sur la qualité de la séquence de l'assemblage.

Il est important de noter qu'il est difficile d'analyser avec précision les petites erreurs puisque les différences observées peuvent provenir d'erreurs dans l'assemblage initial, de variation biologique entre les clones séquencés ou encore d'erreur dans l'alignement des lectures. Il est donc attendu d'observer un certain nombre de différences par rapport au génome de référence lors de cette comparaison. De plus, des différences entre les séquences qui sont assemblées peuvent induire des variations dans le nombre de différences biologiques. Par exemple, un assemblage hautement fragmenté présentera généralement une contraction des répétitions dispersées (notamment les transposons) très similaires alors qu'un assemblage plus complet inclura plus de transposons dans leur contexte génomique (148,149). Or, les transposons étant présents souvent dans des régions du génome évoluant rapidement (150,151), un assemblage plus complet risque de mieux représenter les variations biologiques entre les clones d'une même souche et donc de présenter plus de différences avec l'assemblage de référence. Cette comparaison permet cependant d'identifier clairement les méthodes moins performantes

puisque les génomes produits avec ces méthodes ont un taux d'erreur inférieur à ceux produit avec les autres méthodes.

Tableau 11. – Indels et variant nucléotidique dans les assembles de *S. cerevisiae* générés avec les données nanopore corrigées et non corrigées

Outils de correction	**Indel**		**SNP**	
	Consensus	**Après polissage**	**Consensus**	**Après polissage**
Données brutes	5,631	776	192	69
Canu	5,760	989	209	53
ECTools	786	794	210	87
FMLRC2	928	904	123	115
HG-CoLoR	543	435	239	152
Jabba	181	169	79	83
LoRDEC	1674	982	67	67
NECAT	3,415	891	183	53
PBrC	889	874	91	71

Comparaison des outils d'assemblage

L'identification des outils d'assemblage les plus performants est une étape essentielle pour la construction d'un pipeline d'assemblage *de novo* capable d'assembler des génomes de haute qualité. En effet, la fragmentation du génome, l'absence de segments dans l'assemblage ou encore le nombre d'erreurs dans les assemblages varie d'un assembleur à l'autre (89). La présence d'un nombre plus important d'erreurs dans l'assemblage requiert un temps de correction manuelle plus important du génome final, ce qui risque d'introduire des erreurs par la manipulation manuelle. Encore pire, si elles ne sont pas détectées, ces erreurs peuvent affecter la prédiction des modèles de gènes.

Cependant, l'identification des outils les plus performants est difficile en raison 1) du faible nombre de comparaison exhaustives de ces outils dans la littérature et 2) du développement rapide d'outils d'assemblage. En effet, en raison de l'intérêt croissant pour les technologies de

séquençage de troisième génération pour l'assemblage génomique, un nombre important d'outils d'assemblage capable d'utiliser les données produites par ces technologies ont été développés récemment. De plus, plusieurs outils existants sont fréquemment mis à jour pour utiliser plus efficacement les nouvelles technologies de séquençage ou de *basecalling*.

Afin d'identifier les outils d'assemblage capables de produire les assemblages les plus complets et contenant le moins d'erreur structurelles, les génomes de *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Arabidopsis thaliana* ont été assemblés avec les outils Canu (80), Flye (79), HASLR (86), MaSuRCA (78), miniasm+Racon (99,132), NECAT (125), Raven (87), Shasta (88) et wtdbg2 (52). Ces organismes ont été sélectionnés en raison de la qualité des génomes de références disponibles, ainsi que de la taille de leur génome (12 à 120 Mb) qui est similaire à celle des protistes (20 à 60Mb). La contiguïté des assemblages a été évaluée avec QUAST (93) et les résultats obtenus ont été normalisés pour chaque espèce afin de retirer la variation dans les métriques du au génome et pour permettre une comparaison directe entre les outils (Figure 11). Les assembleurs Canu, Flye, NECAT et Raven produisent les assemblages les plus complets et avec la meilleure contiguïté alors que HASLR et wtdbg2 produisent les assemblages les moins complets et les plus fragmentés (Figure 11). Ces résultats sont similaires à ceux d'autres comparaisons d'outils qui ont déterminées que les outils Flye et NECAT produisent les génomes les complets (78).

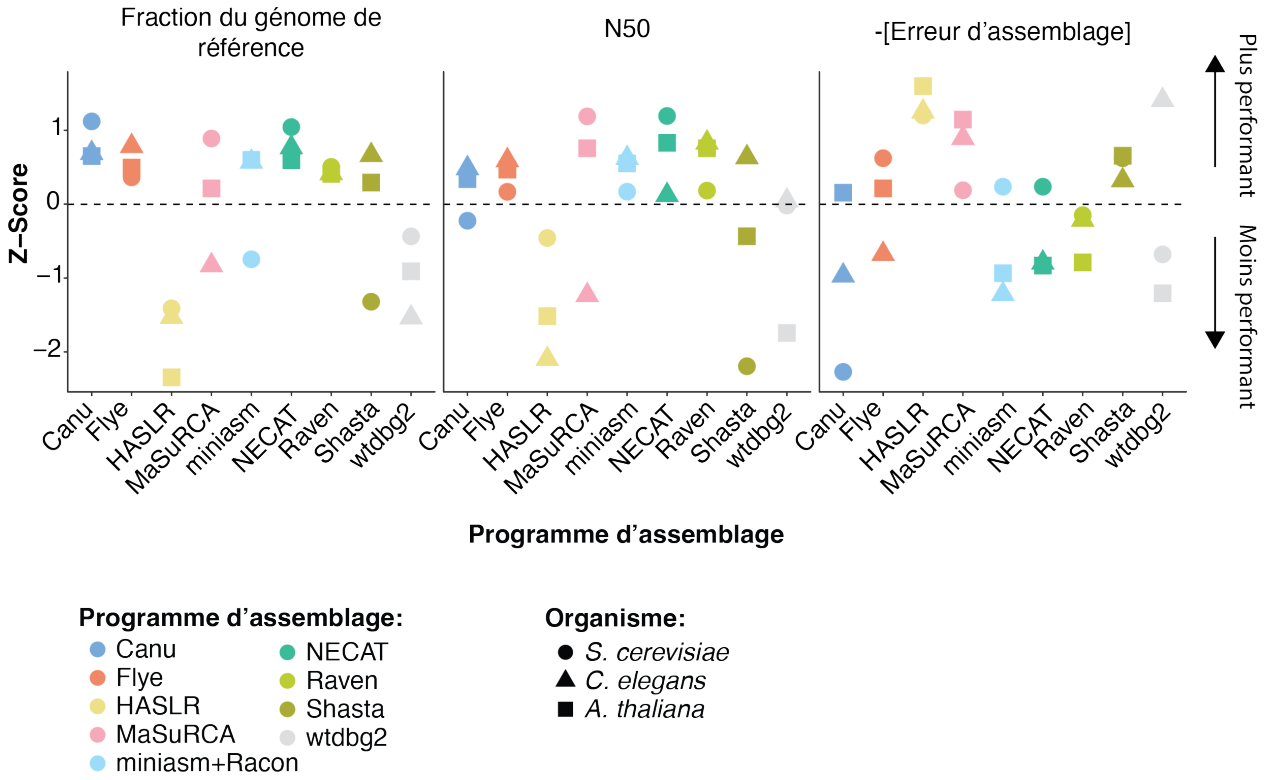

Figure 11. – Comparaison de la contiguïté et de l'exactitude des assemblages produits avec les données nanopore. La fraction du génome de référence des organismes (*S. cerevisiae* W303, *C. elegans* N2 et *A. thaliana* Col-0), le N50 des assemblages nanopore ainsi que le nombre d'erreurs d'assemblages ont été identifiées avec QUAST (93). Le Z-score a été calculé indépendamment pour chaque métrique et pour chaque organisme.

La présence d'erreurs structurelles dans les assemblages a ensuite été évaluée avec QUAST (93). Ces erreurs incluent les régions manquantes dans le génome de référence, les régions pour lesquelles plus d'un contig sont alignés contre le génome de référence, les translocations et les inversions. Les assembleurs HASLR, MaSuRCA et Shasta produisent les assemblages avec le moins d'erreurs tandis que Canu, Raven et wtdbg2 produisent les assemblages avec un nombre d'erreurs élevé (Figure 11). Dans tous les cas, la majorité des erreurs d'assemblages sont courtes (quelques kilobases) et sont situées dans des régions répétées du génome (souvent aux extrémités des contigs des assemblages nanopore) et ont peu d'impact sur la qualité du génome pour les étapes d'analyse en aval (*e.g.* assemblage de transcriptome ou

annotation). Il peut s'agir de répétition assemblée dans deux contigs subséquents, de répétitions absentes dans l'assemblage nanopore ou encore de différences dans le nombre de copies d'une répétition qui cause l'alignement d'une copie supplémentaire dans l'assemblage nanopore sur une autre copie du génome (translocation). Cependant, dans certains cas, des différences structurelles importantes ont été identifiées dans les assemblages nanopore. En effet, certains assemblages produit par Raven présente plusieurs fusions entre des chromosomes (Figure 13D). De plus, certains assemblages produits par Canu, Flye, MaSuRCA, miniasm, NECAT et wtdbg2 présentent des fusions entre deux contigs dans des régions répétés (Figure 13B). Seuls les assembleurs Shasta et HASLR n'ont produit aucune erreur structurelle majeur dans cette comparaison (Figure 13A, C).

Afin d'évaluer si les différences observées au niveau des métriques structuraux ont un impact sur l'analyse en aval effectuée sur les génomes, la qualité des annotations a été comparée. Pour ce faire, le nombre de gènes conservés dans les groupes taxonomique respectifs de *S. cerevisiae*, *C. elegans* et *A. thaliana* et présents dans les assemblages nanopore a été déterminé avec BUSCO (95). Les assembleurs Canu, Flye, NECAT et Raven présentent tous une bonne performance pour l'ensemble des organismes (Figure 12). Ces résultats sont similaires à ceux obtenus par un autre groupe qui a aussi identifié Flye et NECAT comme produisant les génomes avec les meilleurs scores BUSCO (152). La performance des autres assembleurs varie quant à elle d'un organisme à l'autre sans qu'une tendance générale puisse être identifiée (Figure 12).

En somme, les outils d'assemblage Flye et NECAT génèrent des assemblages peu fragmentés et avec peu d'erreurs structurelles. Cependant, puisque Flye est activement développé, cet outil est plus susceptible de tirer profit des améliorations du séquençage nanopore que NECAT qui n'est plus mis à jour depuis 2020. L'assembleur Shasta est aussi intéressant puisqu'il produit les assemblages avec le moins d'erreurs structurelles majeures, bien qu'ils soient plus fragmentés que ceux de Flye. Les outils Flye et Shasta sont donc les meilleurs candidats pour être utilisés dans un *pipeline* d'assemblage. Ainsi, les génomes seront assemblés avec les deux outils et le plus performant (déterminé manuellement en observant notamment la fragmentation et la présence d'erreurs structurelles) sera utilisé.

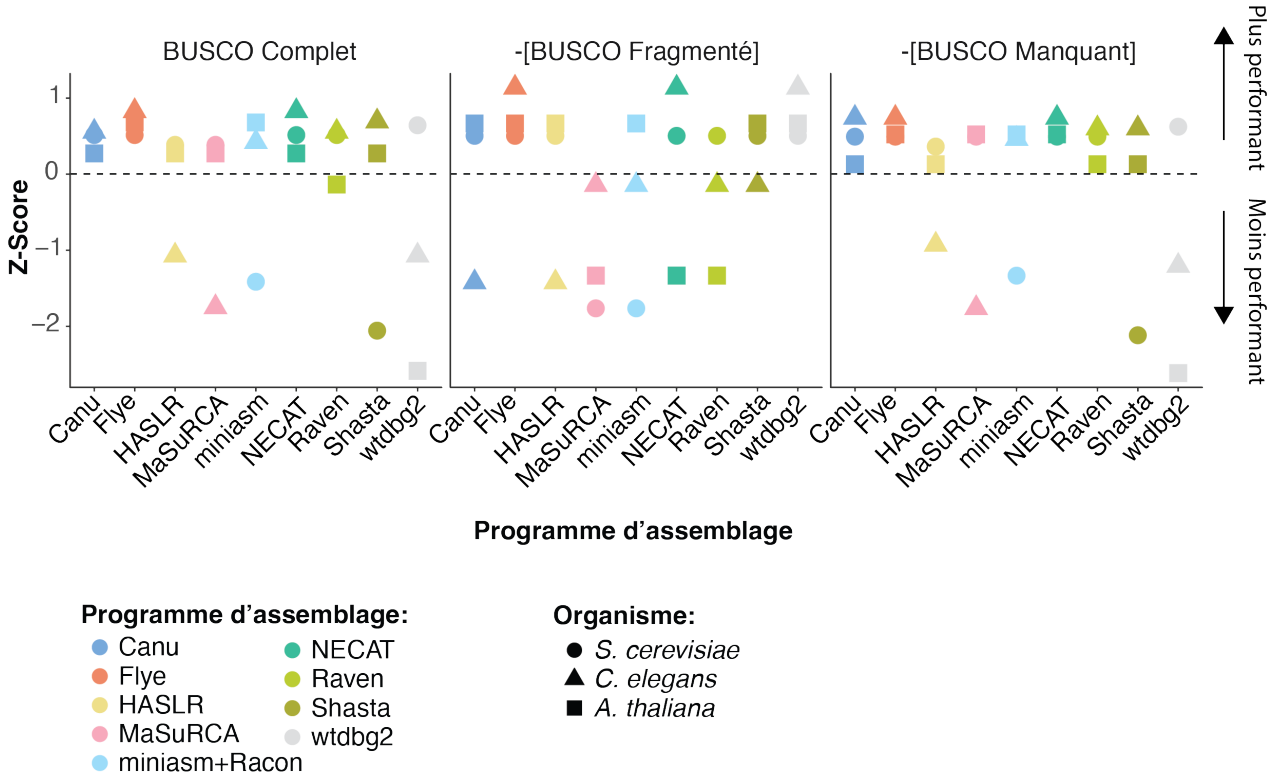

Figure 12. – Comparaison de modèles de gènes BUSCO entre les assemblages produits avec les données nanopore. Le nombre de modèle de gènes complets, fragmentés et manquants a été calculé avec BUSCO en utilisant les données pour les lignées *Saccharomycetes* (*S. cerevisiae*), *Nematoda* (*C. elegans*) et *Brassicales* pour *A. thaliana* (95). Le Z-score a été calculé indépendamment pour chaque métrique et pour chaque organisme.

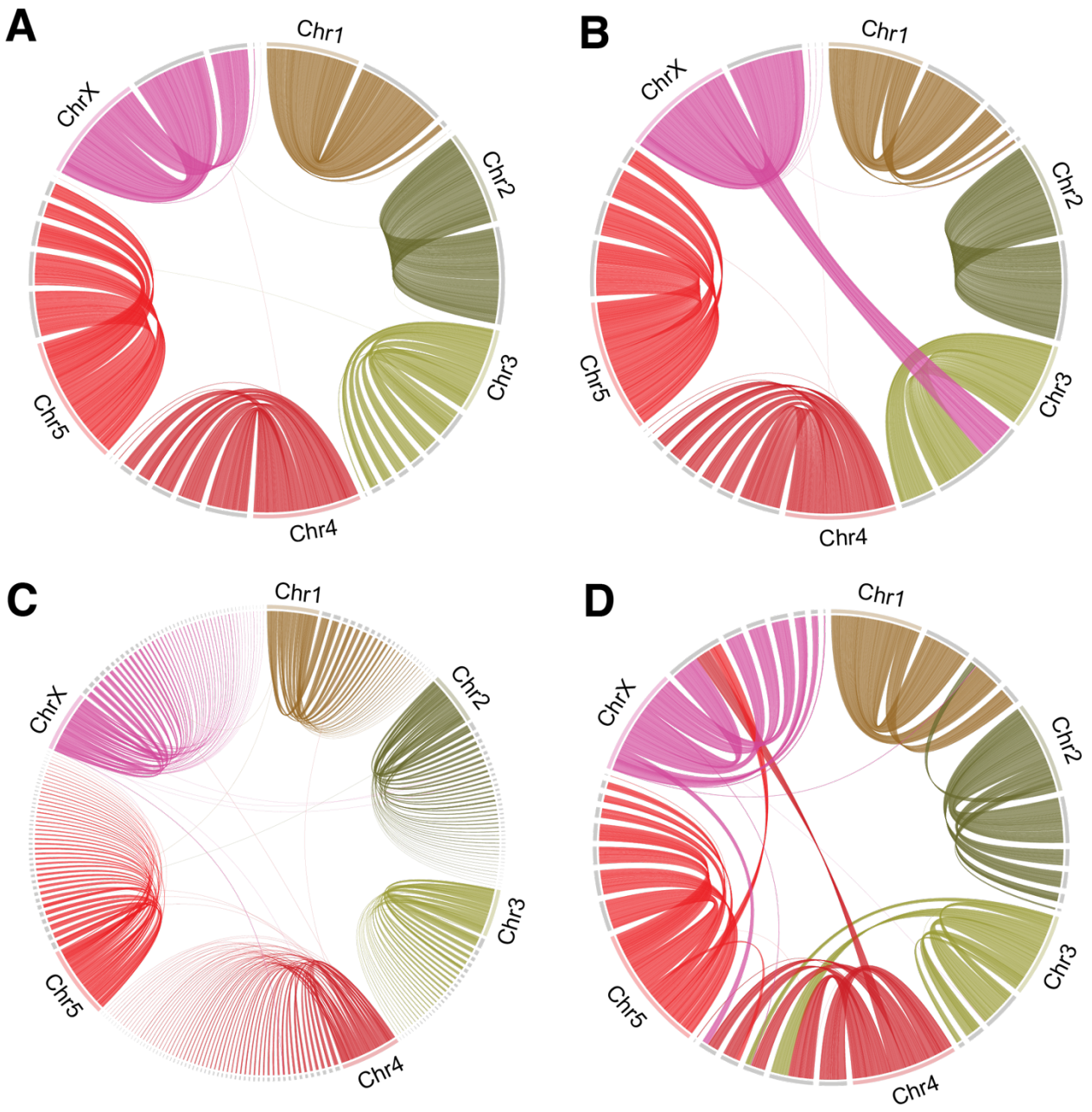

Figure 13. – Catégories de problématiques associées aux différents types d'assembleurs nanopore. Les alignements entre les assemblages nanopore de *C. elegans* et l'assemblage de référence ont été calculés avec NUCmer (104) et visualisés avec Circos (119). Les liens représentent des blocs ayant un alignement entre les chromosomes du génome de référence (en couleur) et les contigs des assemblages nanopore (en gris). (A) Assemblage présentant une bonne contiguïté et ne comportent aucune erreur structurelle majeur. (B) Assemblage présentant une bonne contiguïté et peu d'erreurs structurelles (sur cette figure un contig chimérique est

produit avec le chromosome 3 et le chromosome X). (C) Assemblage fragmentés mais ne présentent aucune erreur structurelle majeur. (D) Assemblage contenant un nombre important d'erreurs structurelles.

Comparaison des outils de polissage

Les lectures nanopore contiennent des erreurs systématiques (principalement dans les homopolymères), qui sont transmises à la séquence consensus des assemblages produits avec ces séquences (89,154). Lorsque ces erreurs sont situées dans une séquence codante, elles peuvent avoir un impact sur l'annotation en modifiant la séquence de la protéine ou en induisant un déphasage du cadre de lecture (81,155). Afin de corriger ces erreurs, des données Illumina, dont le taux d'erreur est significativement plus faible que les données nanopore, sont généralement utilisées pour corriger le génome lors du polissage (156). Plusieurs outils ont été développés pour effectuer cette étape et utilisent généralement l'alignement des lectures Illumina contre le génome pour effectuer la correction (82,84). Cette étape peut aussi être effectuée avec les lectures nanopore par les outils Medaka (développé par Oxford Nanopore Technologies), Nanopolish (157) et Racon (132) mais le polissage avec les lectures nanopore seules reste généralement insuffisant pour corriger toutes les erreurs dans l'assemblage (158).

Afin d'identifier les outils les plus performants, le taux d'erreurs a été déterminé dans le génome de *S. cerevisiae* assemblé avec Flye et corrigé avec les outils de polissage Apollo (134), HyPo (85), ntEdit (135), Pilon (84), POLCA (82) et Racon (132) en utilisant les données Illumina. De plus, les erreurs ont été séparées selon leur contexte génomique (région unique, répétitions dispersées ou région de faible complexité (*e.g.,* répétition en tandem ou homopolymère)). La majorité des outils diminuent sensiblement le nombre d'erreurs dans les assemblages (Figure 14). La performance des outils est similaire pour les régions uniques, mais Pilon semble légèrement plus performant. Pour les répétitions dispersées, les outils Pilon et ntEdit ont la meilleure performance alors que HyPo permet la correction du plus grand nombre d'erreurs pour les régions simples (microsatellites et homopolymères) (Figure 14). L'outils Apollo n'a cependant pas permis de réduction notable du nombre d'erreurs. Afin de tirer profil des avantages de chacun

des outils, les assemblages de *S. cerevisiae* ont été corrigé avec chaque combinaison de l'ensemble de 1 à 3 des outils suivants pour déterminer si une combinaison permet une meilleure correction du nombre d'erreurs. Aucun résultat clair n'a cependant été observé. Ainsi, puisque Pilon produit la meilleure correction dans les régions uniques (qui incluent la majorité des régions codantes) et les répétitions dispersées, cet outil sera utilisé pour effectuer le polissage des génomes. De plus, aucun des outils n'a permis de produire un assemblage ne contenant aucune différence avec l'assemblage de référence. Cette différence entre les assemblages n'est pas exclusivement causée par la présence d'erreur dans l'assemblage nanopore, mais peut aussi être imputable à la présence de variation biologique entre la souche d'où proviennent les données analysées, et la souche utilisée pour produire le génome de référence, ainsi qu'à la présence d'erreurs dans l'assemblage de référence.

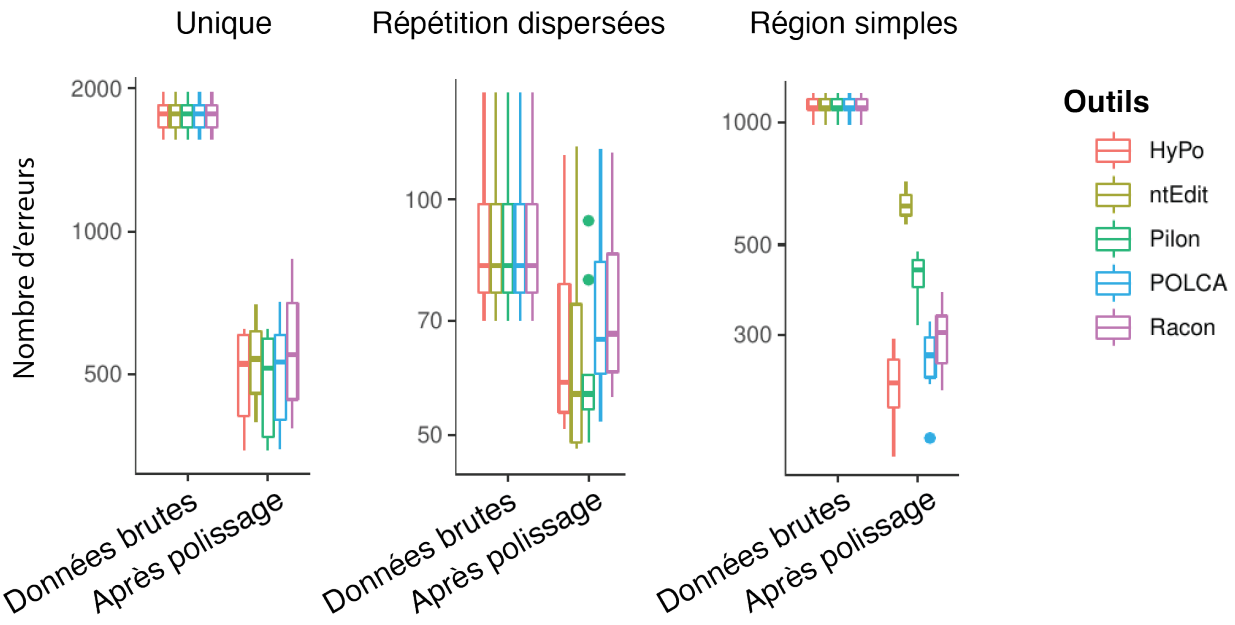

Figure 14. – Impact du polissage sur le nombre d'erreurs dans l'assemblage Flye de *S. cerevisiae*. Les assemblages du génome de *S. cerevisiae* ont été effectués avec Flye en utilisant des sous-échantillons de données nanopore (n=10) et ont été corrigés par polissage. Le nombre d'erreurs dans les régions uniques, les répétitions dispersées et les régions simples (microsatellites et homopolymères) a été évalué directement à partir de l'alignement de

NUCmer avec le programme show-snps de la collection MUMmer et de l'annotation de RepeatMasker.

Pipeline d'assemblage de génome avec les données nanopore

Les résultats de la comparaison des différents outils ont permis d'établir la suite d'outils qui seront utilisés dans le pipeline d'assemblage *de novo* des génomes de jakobides et malawimonades (Figure 15). Le *basecalling* est effectué avec Guppy, sélectionné pour sa rapidité et son faible taux d'erreur. Le génome est ensuite assemblé avec Flye et Shasta en parallèle. L'assembleur Flye a été choisi puisqu'il produit les assemblages les plus complets alors que Shasta est utilisé puisqu'il produit des assemblages assez complets avec un faible taux d'erreur. De plus, le graphe d'assemblage de Shasta est très informatif sur la structure du génome. Le génome est ensuite corrigé avec Medaka suivi de trois itérations de Pilon. Finalement, cette version préliminaire du génome est validée manuellement pour corriger les erreurs structurelles et pour ajouter les régions manquantes à partir de l'information des lectures nanopore.

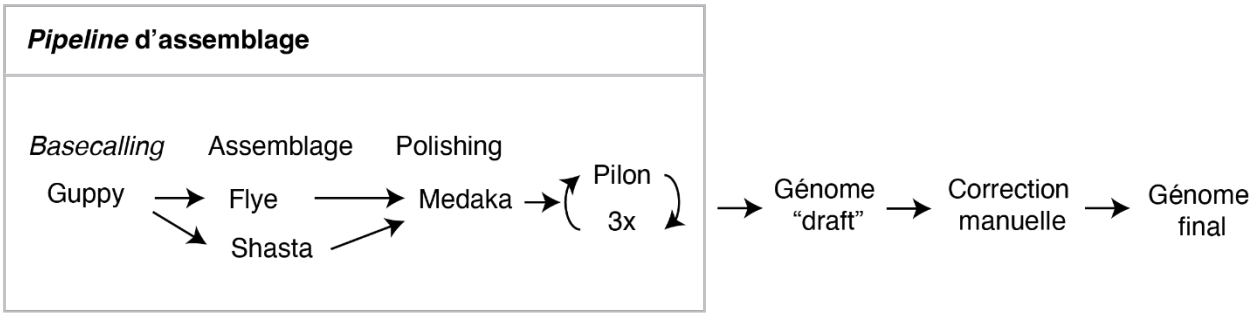

Figure 15. – Pipeline optimisé pour l'assemblage *de novo* des génomes des jakobides et malawimonades.

Assemblage des génomes des Jakobides

Assemblage de génomes haploïdes et diploïdes

Le nombre de copies des chromosomes (ploïdie) est un trait caractéristique des organismes qui varie fréquemment d'une espèce à l'autre mais aussi à l'intérieur d'une même espèce (159). Chez les organismes diploïdes ou polyploïdes, les différentes copies de chaque chromosome

présentent plusieurs variants qui peuvent avoir une importance biologique, notamment en affectant la régulation de l'expression génique (160). De plus, si la ploïdie n'est pas prise en compte, les variants structuraux hétérozygotes (ex. insertion ou délétion) peuvent être incorrectement assemblés. Il est donc utile de reconstruire correctement la séquence de chaque copie des chromosomes. Or, l'assemblage de génome diploïde (ou polyploïde) est un processus complexe et la majorité des assembleurs ne parviennent pas à générer des assemblages de bonne qualité. En effet, une certaine proportion de la variabilité peut parfois être capturée (dans les régions hautement hétérozygotes), mais pour la majorité du génome les haplotypes seront généralement combinés en une seule séquence. Cependant, les longues lectures nanopore contiennent l'information nécessaire pour reconstruire les haplotypes (*i.e.,* traversent plusieurs variants hétérozygotes), certains pipelines permettent donc de les utiliser pour produire un assemblage diploïde qui capture toute la variabilité génétique de l'organisme.

Déterminer la ploïdie

Avant de procéder à l'assemblage des génomes, la ploïdie des organismes a d'abord été estimée en observant l'alignement des lectures nanopore contre un assemblage préliminaire du génome. La présence d'indels et de SNPs hétérozygotes dans les lectures nanopore permet d'identifier l'organisme comme étant diploïde (Figure 16). Les protistes *A. godoyi* et *R. americana* ont ainsi été identifiés comme diploïdes et les protistes *S. ecuadoriensis* et *S. incarcerata* comme haploïdes. Cette identification a été confirmée par l'analyse des distributions de k-mers (Figure 17).

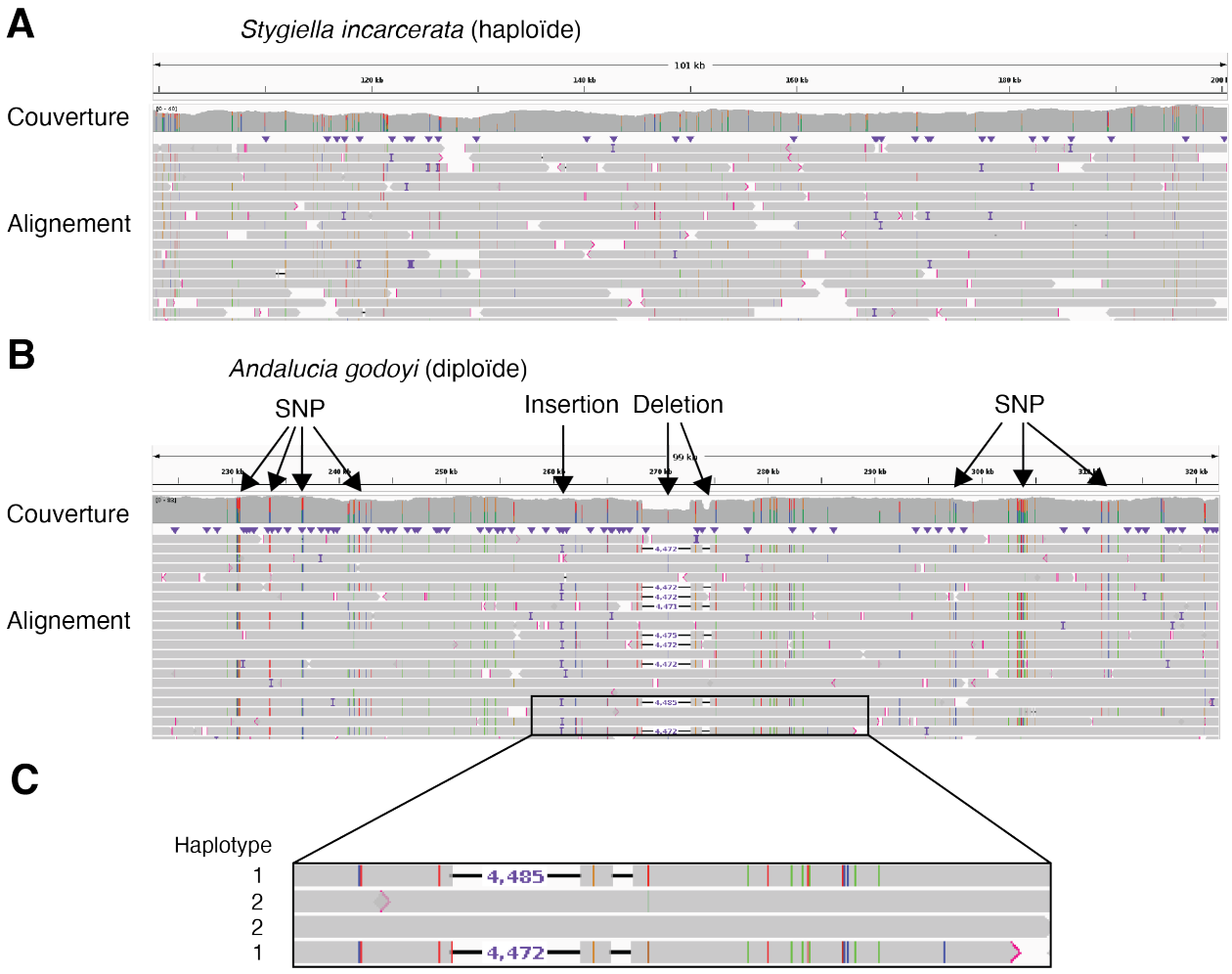

Figure 16. – La présence d'indels et de SNPs hétérozygotes dans l'alignement nanopore permet de déterminer la ploïdie de l'organisme. L'alignement des lectures nanopore contre un assemblage haploïde du génome a été visualisé dans IGV. Dans l'alignement, les SNPs sont représentées par une ligne verticale de couleur, les insertions par un « I » mauve et les délétions par un trait avec la taille. Les lectures ont été alignées contre une région contenant des variants hétérozygotes dans le génome du protiste diploïde *A. godoyi* (A) ainsi qu'une région représentative du génome du protiste haploïde *S. incarcerata* (B). Les haplotypes peuvent être distingués des erreurs de séquençage puisque les lectures présentent les variants d'un seul haplotype (C).

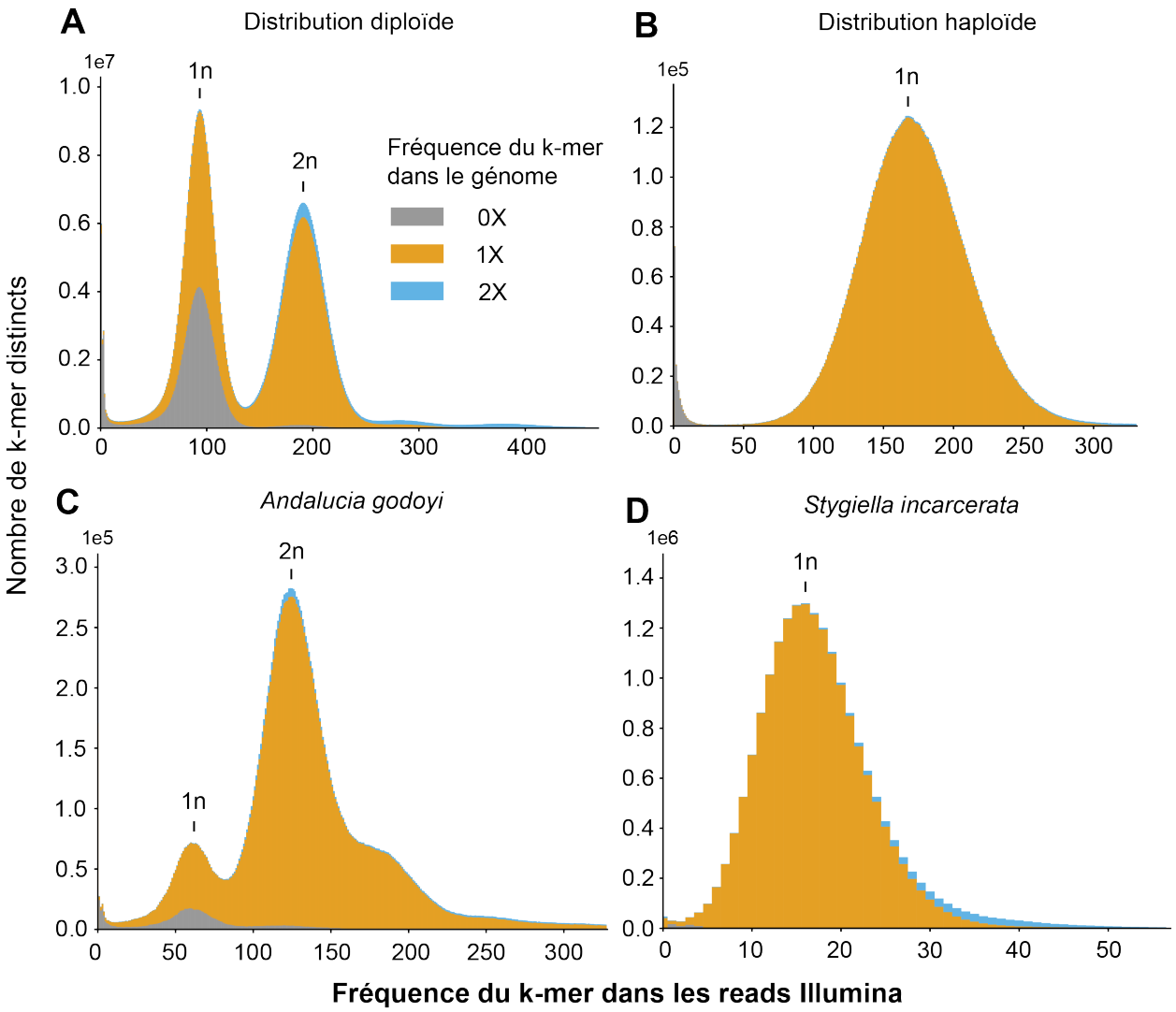

Figure 17. – Identification de la ploïdie de *A. godoyi* et *S. incarcerata* à partir de la distribution des k-mers. Les distributions de k-mers dans les lectures Illumina et le génome ont été comparées avec KAT (100). Les distributions de k-mers dans les lectures Illumina typique d'un organisme diploïde (A) et haploïde (B) ont été générées en utilisant des données du frêne diploïde (*Fraxinus excelsior*) et de la levure haploïde (*Saccharomyces cerevisiae*). L'organisme diploïde présente deux pics (1n : k-mers hétérozygotes et 2n : k-mers homozygote), alors que les organismes haploïdes présentent un seul pic. Les distributions de k-mer des jakobides *A. godoyi* et *S. incarcerata* ont été calculées pour confirmer leur ploïdie (C, D).

Assemblage des génomes des organismes diploïdes

Afin d'assembler correctement les génomes des organismes diploïdes, il est nécessaire d'assembler les deux haplotypes pour chaque chromosome. Deux techniques peuvent être employées pour effectuer cette tâche, soit l'utilisation d'un assembleur capable de séparer les haplotypes (*e.g.,* Shasta (88) ou Flye+HapDup (79,161)) ou l'assemblage de chaque haplotype à partir des lectures nanopore séparées pour chaque haplotypes (avec l'outils Whatshap (101) (Figure 18). La première technique a été employée pour l'assemblage du génome de *A. godoyi* avec Shasta et la deuxième pour celui de *R. americana*. Suite à l'assemblage diploïde, l'alignement des lectures nanopore contre les assemblages a permis de valider que les haplotypes ont été correctement assemblés (Figure 19A,B). De plus, pour valider que le polissage ne change pas les variants hétérozygotes pour ceux de l'autre haplotype, l'alignement des lectures nanopore contre le génome après polissage a aussi été visualisé. Ces résultats ont permis de confirmer que le polissage n'a induit aucun changement d'haplotype pour les variants hétérozygotes (Figure 19C).

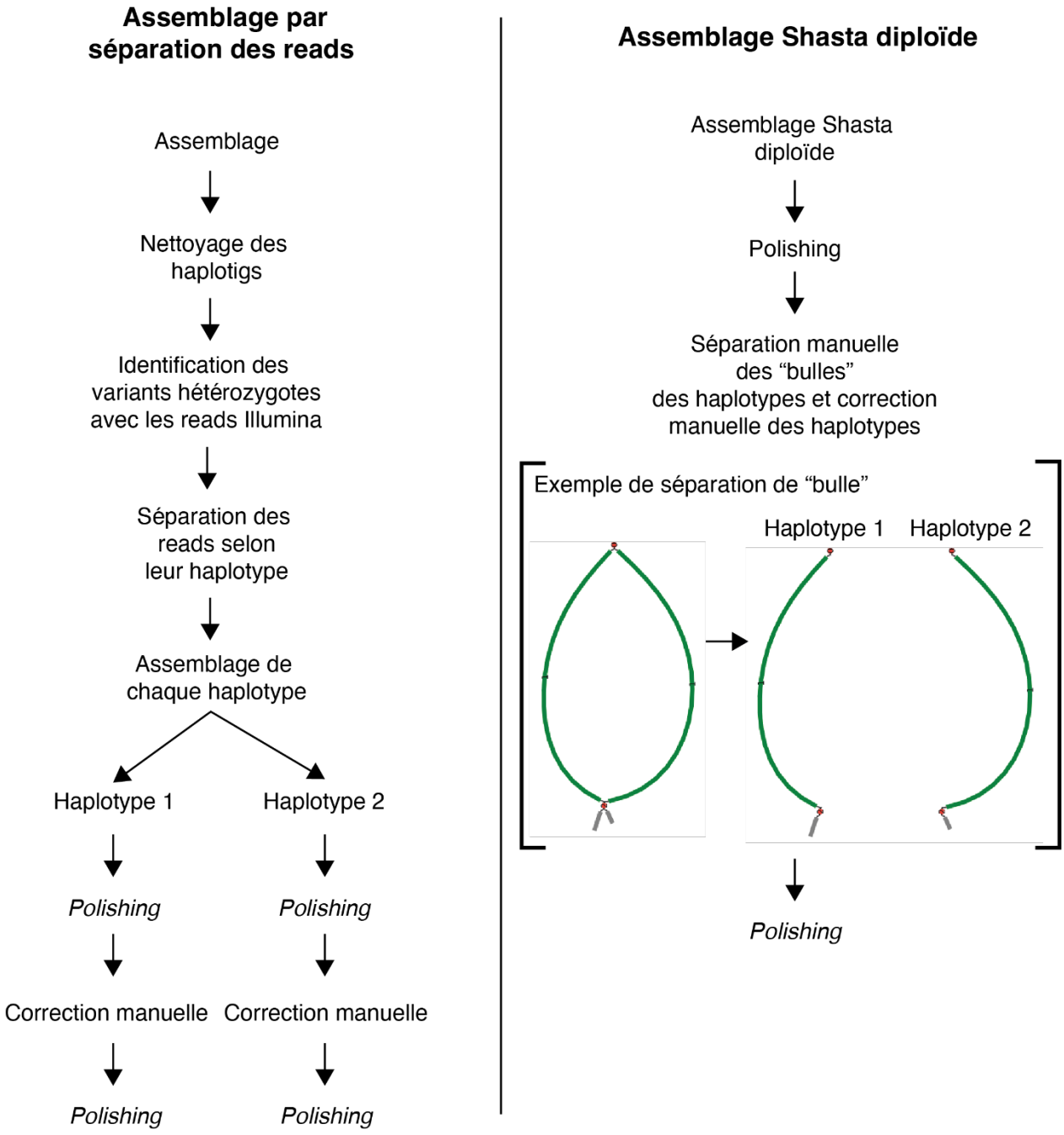

Figure 18. – *Pipeline* pour l'assemblage du génome d'organismes diploïdes.

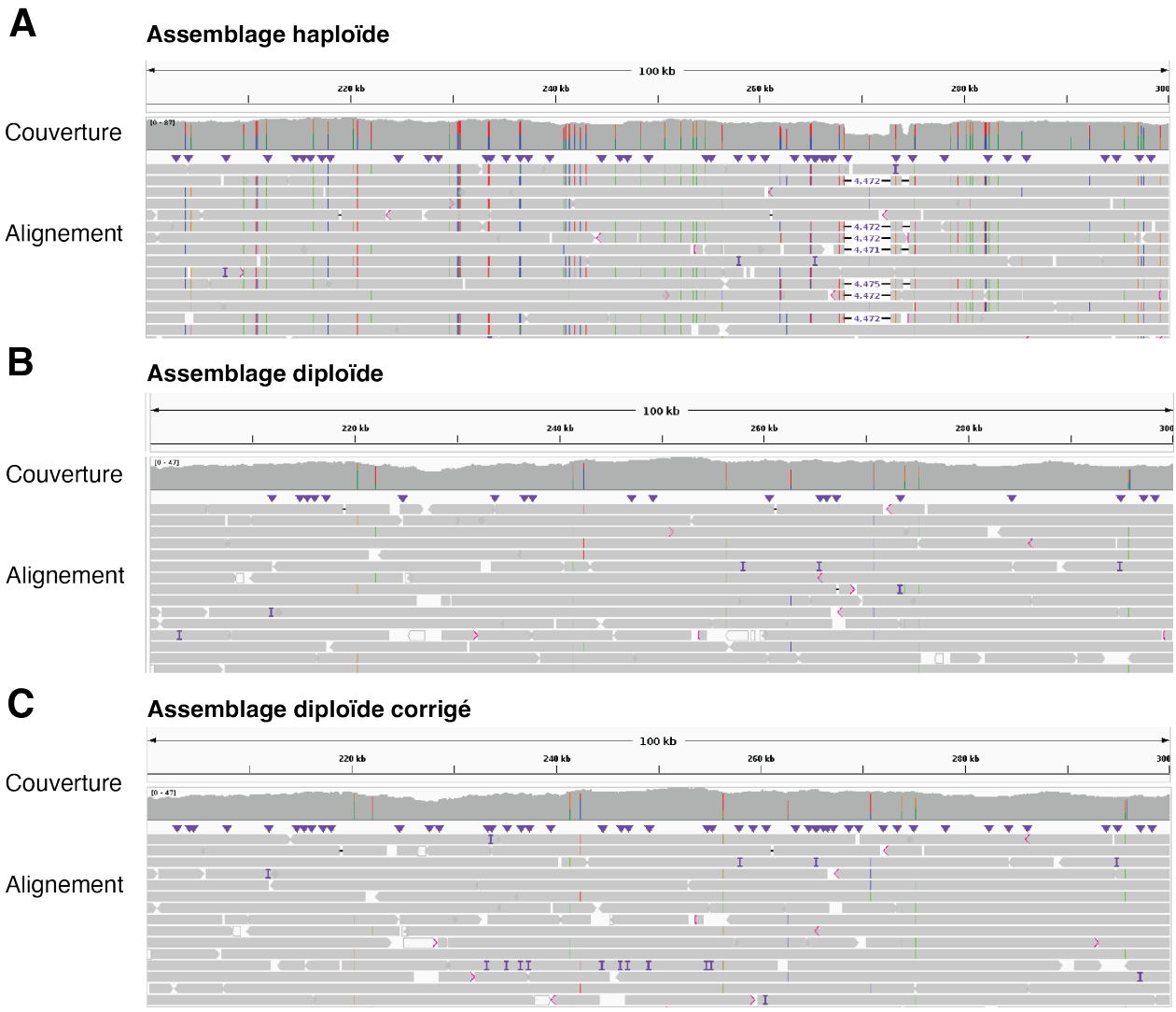

Figure 19. – L'alignement des séquences nanopore contre l'assemblage diploïde de *A. godoyi* permet de valider que les haplotypes ont été correctement assemblés et corrigés lors du polissage. L'alignement des lectures nanopore contre un assemblage haploïde du génome a été visualisé dans IGV. Dans l'alignement, les SNPs sont représentées par une ligne verticale de couleur, les insertions par un « I » mauve et les délétions par un trait avec la taille. Les lectures nanopore ont été alignées contre l'assemblage haploïde (A), l'assemblage diploïde (B) et l'assemblage diploïde après polissage (C).

Amélioration de la contiguïté des génomes dans l'assemblage nanopore

L'assemblage des génomes des protistes a permis d'améliorer de manière importante la contiguïté des génomes de *R. americana* et *S. ecuadoriensis*. (Tableau 12). En effet, les génomes assemblés avec les données nanopore sont beaucoup moins fragmentés que les génomes assemblés avec des données Illumina exclusivement (Tableau 12). De plus cette amélioration de la contiguïté se traduit par une amélioration de la qualité des modèles de gènes selon BUSCO, avec une diminution du nombre de gènes incomplets ou manquants (Tableau 12).

Tableau 12. – Métriques de l'assemblage des génomes des jakobides

Organisme	**Métrique**	**Assemblage Illumina**[†]	**Assemblage nanopore**
A. godoyi	Taille (Mb)	20.0	21.1
	N50 (Mb)	0.32	0.34
	Contigs	66	60[‡]
	BUSCO Complet (%)	79.6	80.4
	BUSCO Fragmenté (%)	5.9	5.9
	BUSCO Manquant (%)	14.5	13.7
R. americana	Taille (Mb)	52.4	61.6
	N50 (Mb)	0.006	1.5
	Contigs	16 374	172
	BUSCO Complet (%)	65.1	91.4
	BUSCO Fragmenté (%)	18.0	2.4
	BUSCO Manquant (%)	16.9	6.2
S. ecuadoriensis	Taille (Mb)	47.5	53.3
	N50 (Mb)	0.056	1.216
	Contigs	3739	64
	BUSCO Complet (%)	81.6	91.0
	BUSCO Fragmenté (%)	5.5	2.0
	BUSCO Manquant (%)	12.5	7.0
S. incarcerata	Taille (Mb)		21.7

N50 (Mb)	1.139
Contigs	19‡
BUSCO Complet (%)	75.7
BUSCO Fragmenté (%)	5.9
BUSCO Manquant (%)	18.4

† L'assemblage de *A. godoyi* a été produit avec des données de séquençage Illumina et 454.

‡ Nombre de chromosomes (contigs présentant des télomères aux deux extrémités).

Structure du génome de *Andalucia godoyi*

Les données de séquençage produites ont permis l'assemblage du génome de *A. godoyi* diploïde au niveau chromosomique. La structure de l'ensemble des chromosomes a été confirmée grâce à l'alignement des séquences nanopore et certaines corrections manuelles (*e.g.*, ajout des régions télomériques et sub-télomérique et correction d'indels; selon la méthode décrite à la Figure 15) ont été apportées afin de produire l'assemblage final. Cet assemblage est composé d'un total de 64 contigs dont la taille varie de 181 kb à 579 kb. De ces contigs, 60 ont été identifiés comme des chromosomes complets par la présence de séquence télomérique $[TTAGGG]_n$ aux deux extrémités des molécules assemblées.

La structure de l'ensemble de ces chromosomes est similaire à celle des chromosomes de la version publique (42). En effet, les 60 chromosomes identifiés dans l'assemblage nanopore sont aussi assemblés en un chromosome dans la version publique de l'assemblage. Cependant, les régions télomériques et sub-télomériques, généralement composées de transposons, d'une longueur d'environ 2-10 kb ne sont pas présentes dans l'assemblage publique (Figure 20). De plus, la séquence de certains transposons situés à l'intérieur des chromosomes est entièrement absente dans la version publique du génome (Figure 21). Afin de valider qu'il ne s'agit pas de différences biologiques, les séquences 454 utilisées pour produire l'assemblage public ont été alignées contre l'assemblage nanopore. L'alignement de ces séquences suggère que les transposons étaient présents dans le génome lors du séquençage par 454 et qu'il s'agit réellement d'erreurs d'assemblage. Ainsi, dans la majorité des cas, ces améliorations ne sont que structurelles et n'ont aucun impact sur les modèles de gènes obtenus à partir des assemblages.

Cependant, quelques différences importantes sont observées dans certains chromosomes où l'assemblage nanopore a permis d'assembler des séquences contenant des gènes qui ne sont pas présents dans la version publique de l'assemblage (Figure 22). L'ajout de ces séquences sur cinq chromosomes a permis d'identifier 30 régions codantes supplémentaires dans l'assemblage nanopore.

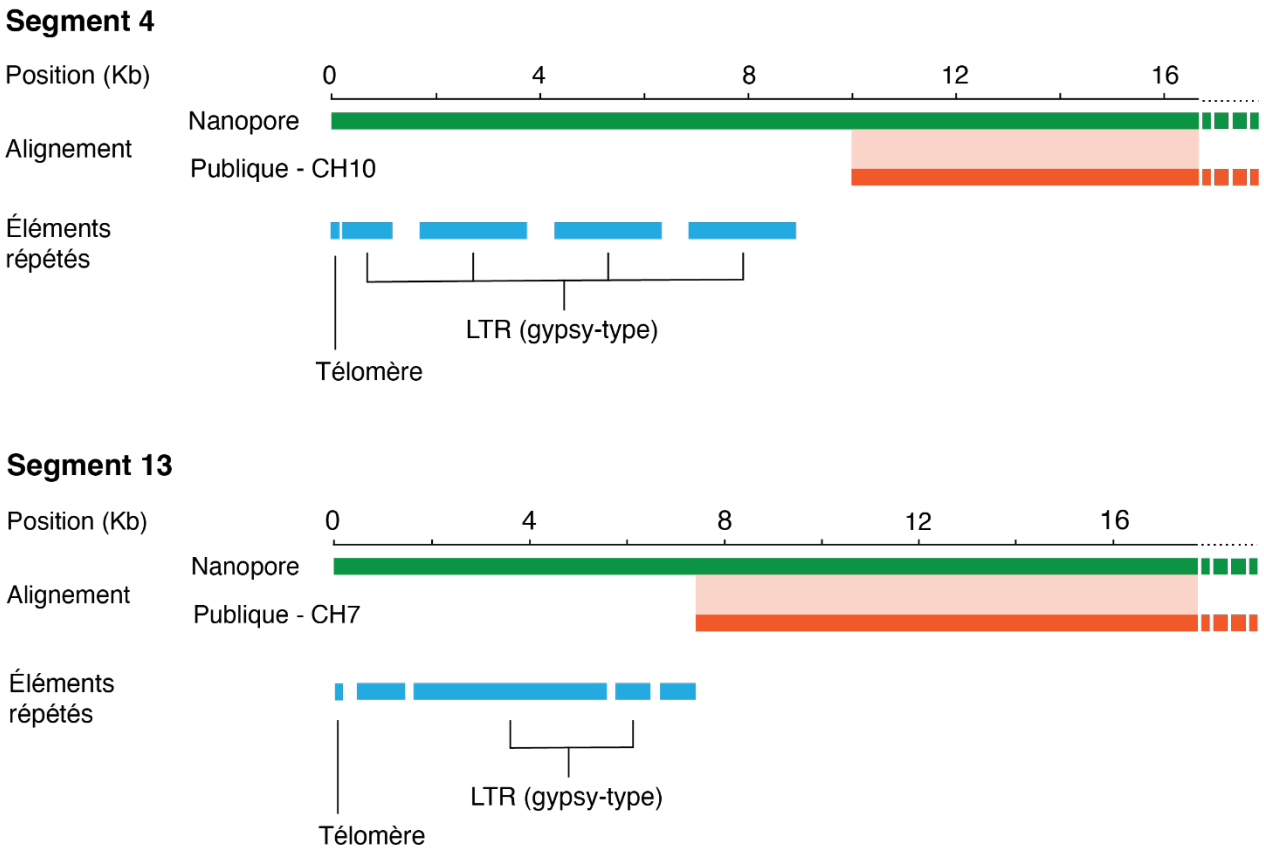

Figure 20. – Différences structurelles typiques dans les régions télomériques et sub-télomériques entre l'assemblage nanopore et l'assemblage publique. L'assemblage public du génome (en orange) a été aligné contre la version assemblée à partir des données nanopore (en vert) avec NUCmer. Les éléments répétés (en bleu) ont été annotés avec RepeatMasker. La figure a été générée à partir de la visualisation des données dans IGV.

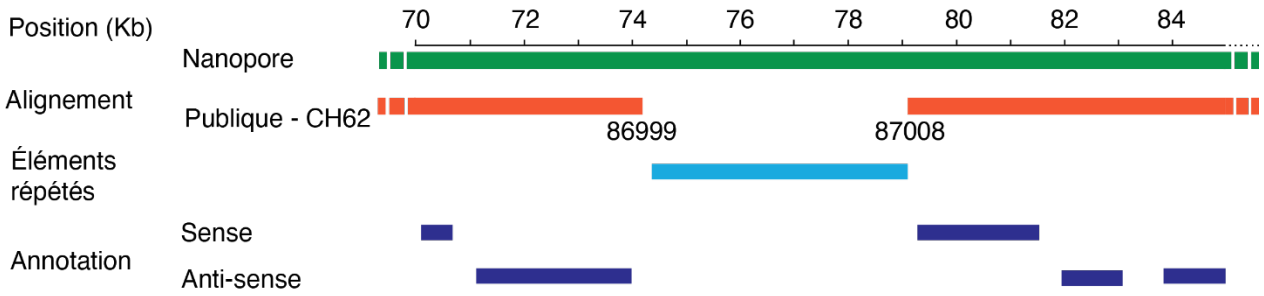

Figure 21. – Structure typique d'une région contenant un transposon qui n'est pas assemblée dans la version publique du génome. L'assemblage public du génome (en orange) a été alignée contre la version assemblée à partir des données nanopore (en vert) avec NUCmer. Les éléments répétés (en bleu pâle) ont été annotés avec RepeatMasker et les protéines (en bleu foncé) ont été annotés à l'aide d'une méthode interne. La figure a été générée à partir de la visualisation des données dans IGV.

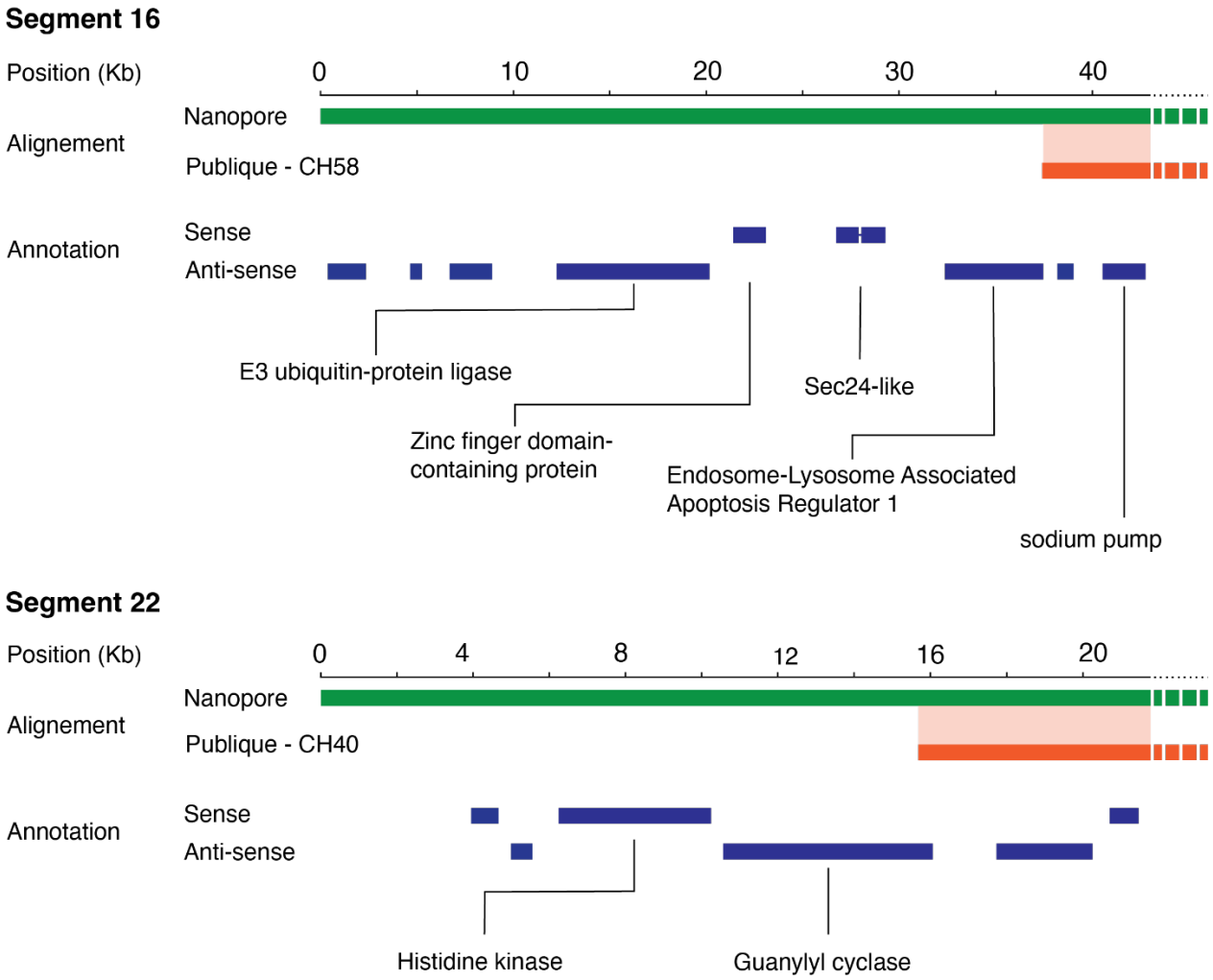

Figure 22. – Structure de régions manquantes dans l'assemblage public qui contienne des gènes codants. L'assemblage public du génome (en orange) a été aligné contre la version assemblée à partir des données nanopore (en vert) avec NUCmer. Les protéines (en bleu) ont été annotées à l'aide d'une méthode interne. La figure a été générée à partir de la visualisation des données dans IGV.

Les chromosomes circulaires de *A. godoyi*

En plus des 60 chromosomes linéaires identifiés dans l'assemblage nanopore de *A. godoyi*, quatre contigs d'une taille de 283 à 340 kb ne présentant pas de séquences télomériques ont été identifiés. L'analyse de l'alignement des lectures nanopore contre ces contigs suggère que les molécules sont circulaires puisque l'alignement des lectures situées aux extrémités de ces contigs continue sur l'autre extrémité. Cependant, ces molécules n'ont pas été identifiées comme des

plasmides bactériens lors de la filtration des séquences puisqu'elles ne possèdent pas de gènes qui codent pour des ARNr de type bactérien ou les protéines d'initiation de la réplication, les relaxases et les protéines de couplage de type IV associées aux plasmides connus (162–164), ni d'homologie avec d'autres gènes typiquement bactériens. De plus, les quatre chromosomes circulaires ont un contenu en GC similaire au restant du génome nucléaire de *A. godoyi* (environ 51%). Finalement, l'ensemble de ces molécules contiennent au moins un gène contenant un intron avec un site accepteur GT et donneur AG caractéristique des introns splicéosomaux (Tableau 13). L'ensemble de ces éléments, et surtout la présence de gènes avec des introns splicéosomaux, suggère que ces molécules font partie du génome nucléaire de *A. godoyi*.

Tableau 13. – Introns splicéosomaux présents sur les chromosomes circulaires de *A. godoyi*

Numéro du chromosome circulaire	Annotation	Nombre d'introns dans le gène
1	Réductase ferrique	1
2	Inconnu †	3
2	ADN topoisomérase de type II	2
2	Inconnu †	1
3	Catalase	1
4	Catalase	1
4	Oxidoréductase	2
4	Catalase	1

† Aucune annotation fonctionnelle n'a été obtenue pour ces gènes.

Les molécules d'ADN circulaires sont cependant rares chez les eucaryotes et n'ont été identifiées chez aucun autres jakobides et malawimonades. Deux types de molécules circulaires ont été décrites chez les eucaryotes, soit les plasmides des levures et champignons filamentaires et les ADN circulaires extrachromosomaux (ADNecc). Les plasmides ont été identifiés chez les champignons filamentaires (*e.g. Neurospora*, *Fusarium* et *Alternaria*) ainsi que chez les levures (165,166). Les plasmides ayant été identifiés sont tous des éléments de petite taille (<10 kb) contenant peu de gènes et (contrairement aux chromosomes circulaires d'*Andalucia*; Figure 25-

28) sont généralement présents en plusieurs copies dans les cellules (165,166). Les ADNecc quant à eux sont un type de molécule circulaire ayant été identifiés chez certaines espèces eucaryotes (167,168). Il s'agit également de courtes molécules (<1-40kb), dont le mécanisme de formation est inconnu, qui peuvent être produites à partir de régions uniques ou répétées du génome (169–171). Dans la majorité des cas, les ADNecc ne sont présents que dans certaines cellules d'une population , cependant, certains ADNecc peuvent être répliqués en mitose et être transmis (170).

Les molécules circulaires identifiées chez *A. godoyi* ne partagent cependant pas les caractéristiques de ces molécules puisqu'elles sont de grande taille et sont uniques dans le génome de *A. godoyi*, elles ne sont donc pas le produit de recombinaison. De plus, leur couverture nanopore (environ 35X pour les molécules circulaires [1n] et 80X pour les molécules linéaires [2n]) suggère qu'elles sont présentes en une seule copie par cellule. Cependant, afin de confirmer que ces molécules ne sont pas des molécules bactériennes, des phylogénies de protéines présentes sur ces molécules ont été construites. Pour ce faire, les protéines orthologues à l'ensemble des protéines présentes sur les molécules circulaires ont d'abord été identifiées dans une base de données de protéomes d'eucaryotes et de procaryotes grâce à un outil interne utilisant les profils HMM pour identifier les protéines homologues. Les séquences ont été alignées puis des phylogénies ont été inférées par *maximum likelihood*. La phylogénie de plusieurs protéines suggère que les gènes présents sur les molécules circulaires ont évolué avec les Discoba (exemple dans la Figure 22), ce qui permet de confirmer que les chromosomes circulaires sont des molécules d'ADN des protistes. La méthode utilisée pour identifier les protéines homologues n'a cependant pas permis d'obtenir de phylogénie de haute qualité pour la majorité des protéines analysées. En effet, cette recherche, basée sur un profil HMM, est très puissante pour identifier des protéines homologues chez des organismes divergents, mais ne permet pas toujours de savoir si les protéines identifiées sont de réelles orthologues (donc pas de paralogues), ce qui est nécessaire pour produire une phylogénie fiable (172,173). Puisque la majorité des gènes présents sur les chromosomes circulaires (Figure 25-28) proviennent de familles de gènes complexes (*e.g.,* Ca2+ ATPase, ADN topoisomérase de type II, histidine kinase ou encore transporteur ABC) présentes chez les trois domaines de la vie et chez les virus, mais l'identification d'orthologues avec cette technique reste difficile.

Le contenu génique des différents chromosomes circulaires 2 à 4 est majoritairement similaire (Figure 25-28). En effet, dans l'ensemble, les gènes présents sur ces chromosomes circulaires codent pour des protéines associées aux voies de signalisation (majoritairement adénylate cyclase et kinases) ainsi qu'au métabolisme. Le contenu génique du chromosome circulaire 1 diffère puisque la majorité des gènes codent pour des protéines de transport d'ions. Tous les chromosomes ne semblent cependant pas contenir d'ADN polymérase ou de transcriptase inverse, présents sur les plasmides des champignons filamenteux, ni de recombinase, présente sur les plasmides 2-micron des levures (165,166).

De plus, les chromosomes circulaires de *A. godoyi* ne présentent pas de signature caractéristique de la diploïdie (tel que décrit à la figure 15) dans l'alignement des lectures nanopore (Figure 24A). En effet, aucune région ou les lectures présentent deux patrons d'erreur distincts (caractéristique des organismes diploïdes) n'ont été identifiées et les erreurs semblent distribuées aléatoirement sur les lectures (Figure 24A). Par ailleurs, la couverture de séquençage Illumina des chromosomes circulaires est environ 50% de celle des chromosomes linéaires. Ces deux caractéristiques suggèrent que les chromosomes circulaires sont haploïdes contrairement au reste du génome qui est diploïde.

Ainsi, ces observations suggèrent que le génome de *A. godoyi* serait composé de 60 chromosomes linéaires et 4 chromosomes circulaires. La présence de chromosomes circulaires étant rare chez les eucaryotes et n'ayant pas déjà été observée dans les assemblages d'autres jakobides, cette observation soulève plusieurs questions quant à l'origine et à la biologie de ces molécules. Où sont-elles localisées dans les cellules ? Quel est leur rôle biologique pour les cellules de *A. godoyi* ? Quel est l'origine de ces molécules (*e.g.* plasmide bactérien, virale, chromosomes ancestraux du protiste) ? Comment ces molécules sont-elles transférées d'une génération à l'autre ? Est-ce que les autres espèces de jakobides partagent cette caractéristique ?

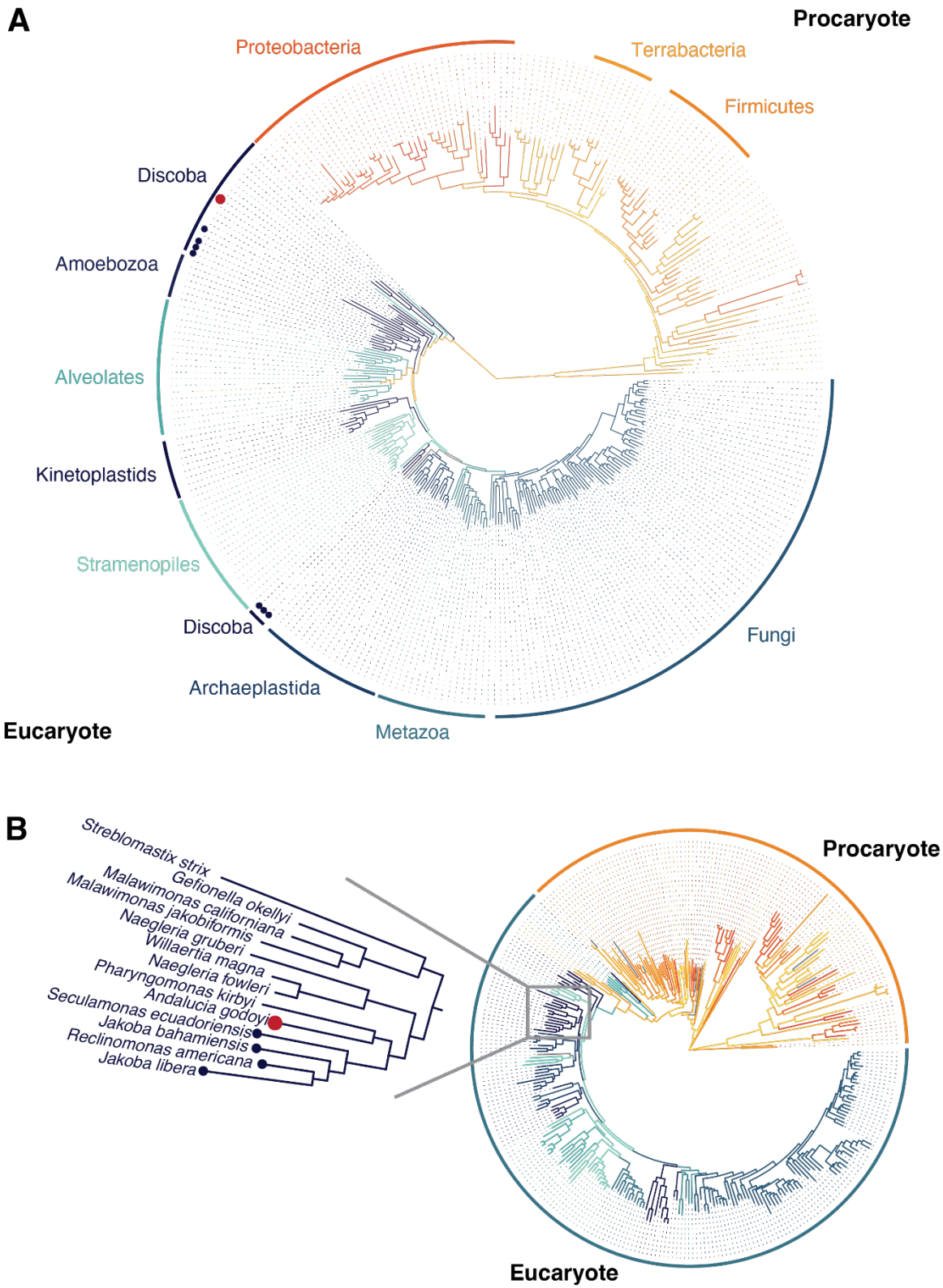

Figure 23. – Phylogénie représentative de protéine des chromosomes circulaires. Phylogénie (A) d'une topoisomérase de type II du chromosome circulaire 2 et (B) d'une sous-unité d'une Ca^{2+}

ATPase. Les séquences homologues à la séquence de *A. godoyi* ont été identifiées par recherche itérative à l'aide d'un profil HMM. La phylogénie *maximum likelihood* a été obtenue avec IQ-TREE et visualisée avec ggtree dans R. Les branches de procaryotes sont représentées en jaune et rouge et les branches des eucaryotes en bleu et vert. Les couleurs représentent les groupes taxonomiques. La position de *A. godoyi* est indiquée par le cercle rouge. La position des autres jakobides par les cercles bleu foncé.

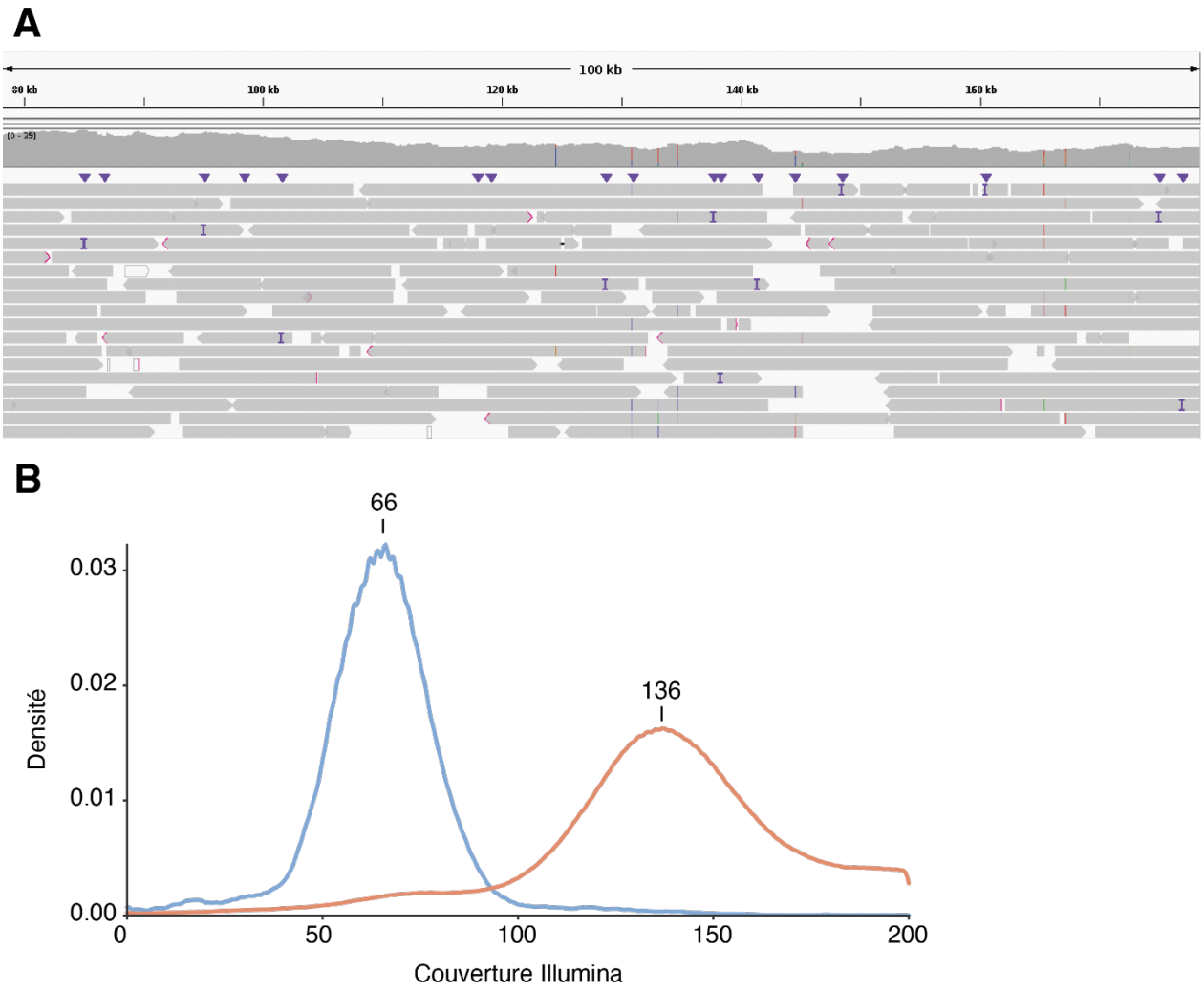

Figure 24. – Les chromosomes circulaires de *A. godoyi* sont haploïde. (A) Les séquences nanopore sont été alignées contre le génome de *A. godoyi* avec minimap2 et visualisée dans IGV. (B) La couverture des séquences Illumina a été évaluée avec bedtools selon l'alignement produit par Bowtie2. La distribution présente la couverture pour l'ensemble des chromosomes linéaires ou circulaires.

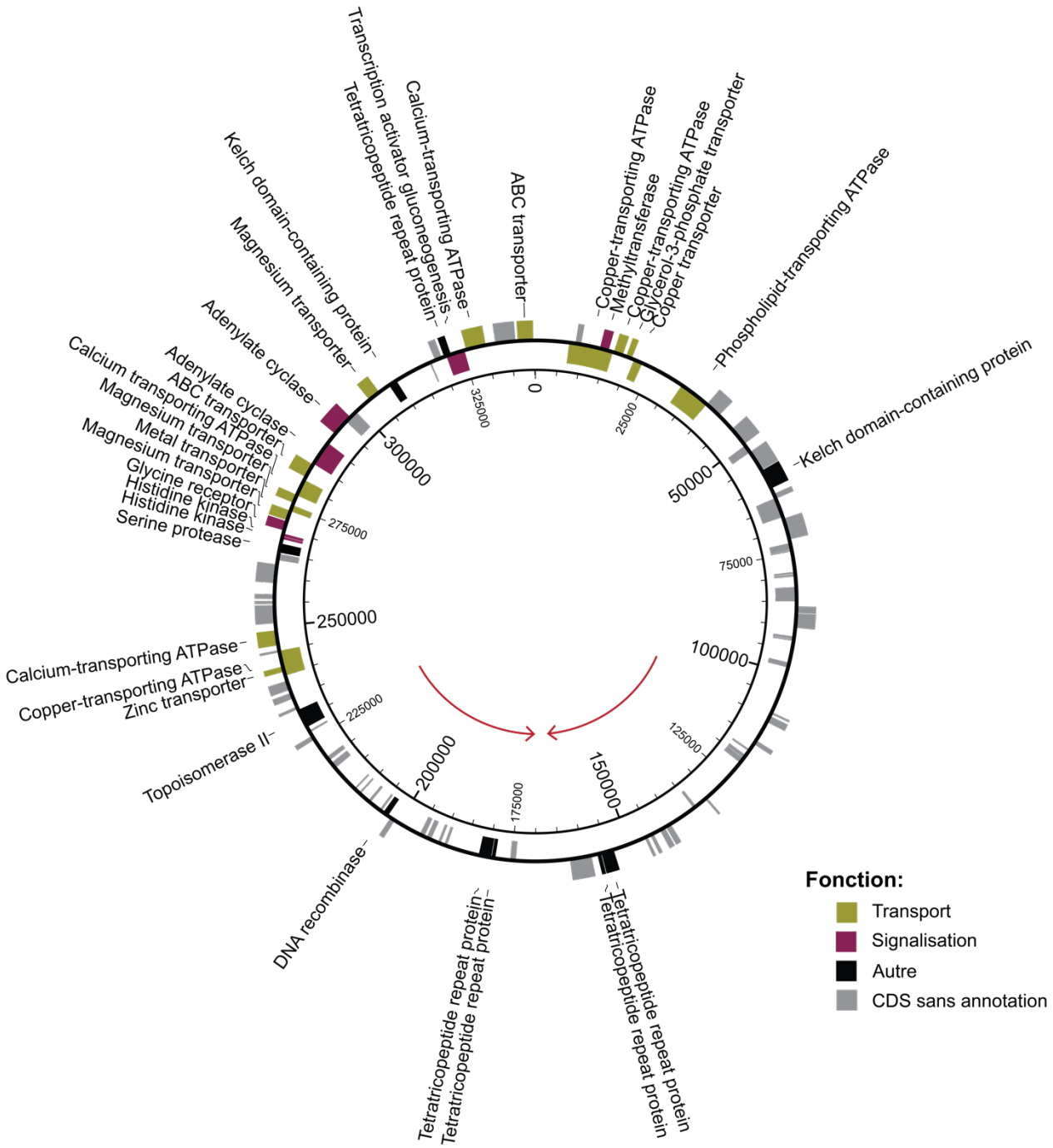

Figure 25. – Carte de gènes du chromosome circulaire 1 de *A. godoyi*. Les gènes ont été annotés selon les annotations des protéines de *A. thaliana*, *H. sapiens* et *S. cerevisiae* identifiées comme homologues à partir d'une recherche avec le profil HMM de la protéine de *A. godoyi*. Les

flèches rouges indiquent la position de répétitions inversées. La figure a été générée avec Circos (119).

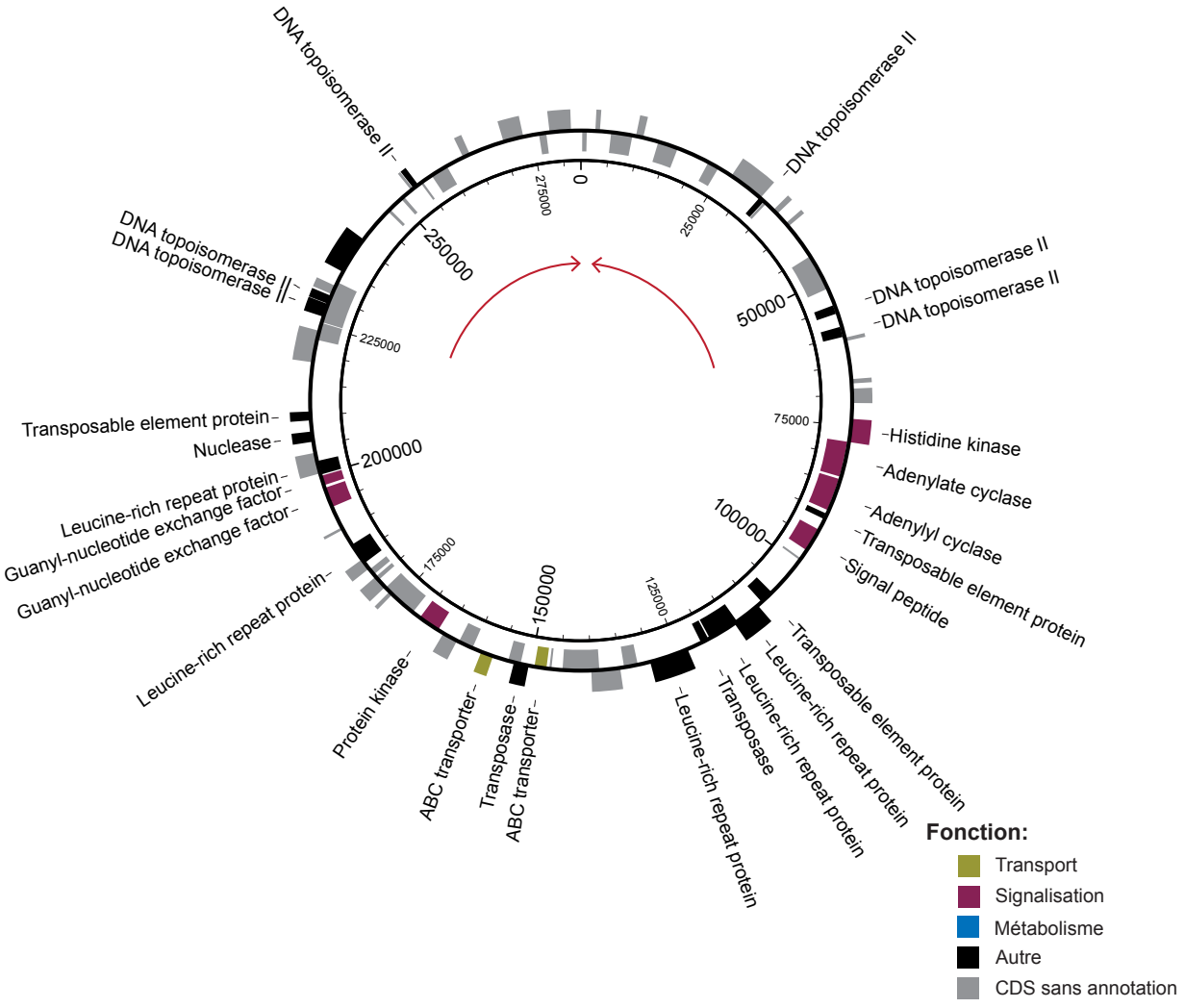

Figure 26. – Carte de gènes du chromosome circulaire 2 de *A. godoyi*. Les gènes ont été annotés selon les annotations des protéines de *A. thaliana*, *H. sapiens* et *S. cerevisiae* identifiées comme homologues à partir d'une recherche avec le profil HMM de la protéine de *A. godoyi*. Les flèches rouges indiquent la position d'une répétition inversée. La figure a été générée avec Circos (119).

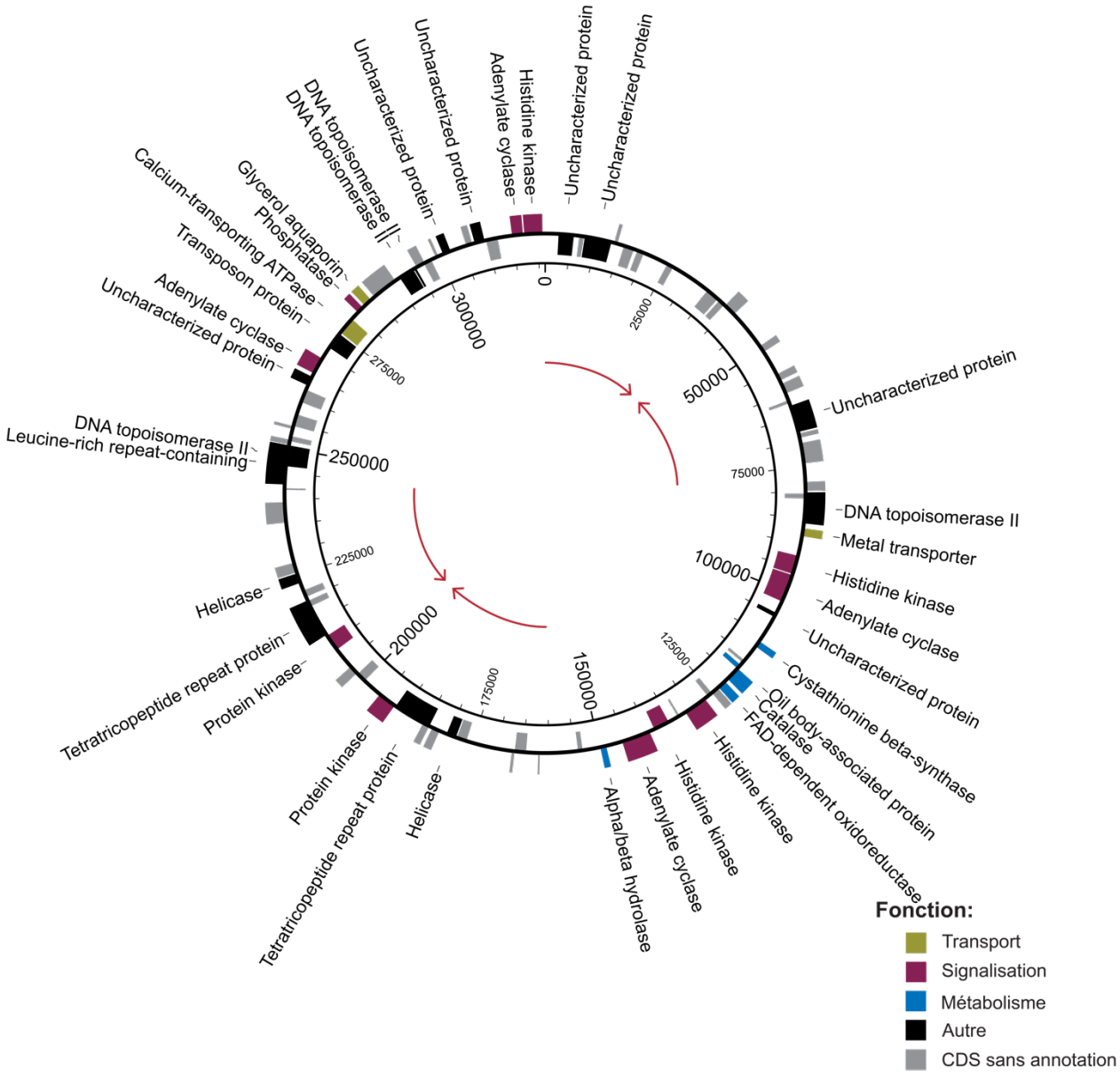

Figure 27. – Carte de gènes du chromosome circulaire 3 de *A. godoyi*. Les gènes ont été annotés selon les annotations des protéines de *A. thaliana*, *H. sapiens* et *S. cerevisiae* identifiées comme homologues à partir d'une recherche avec le profil HMM de la protéine de *A. godoyi*. Les flèches rouges indiquent la position de répétition inversée. La figure a été générée avec Circos (119).

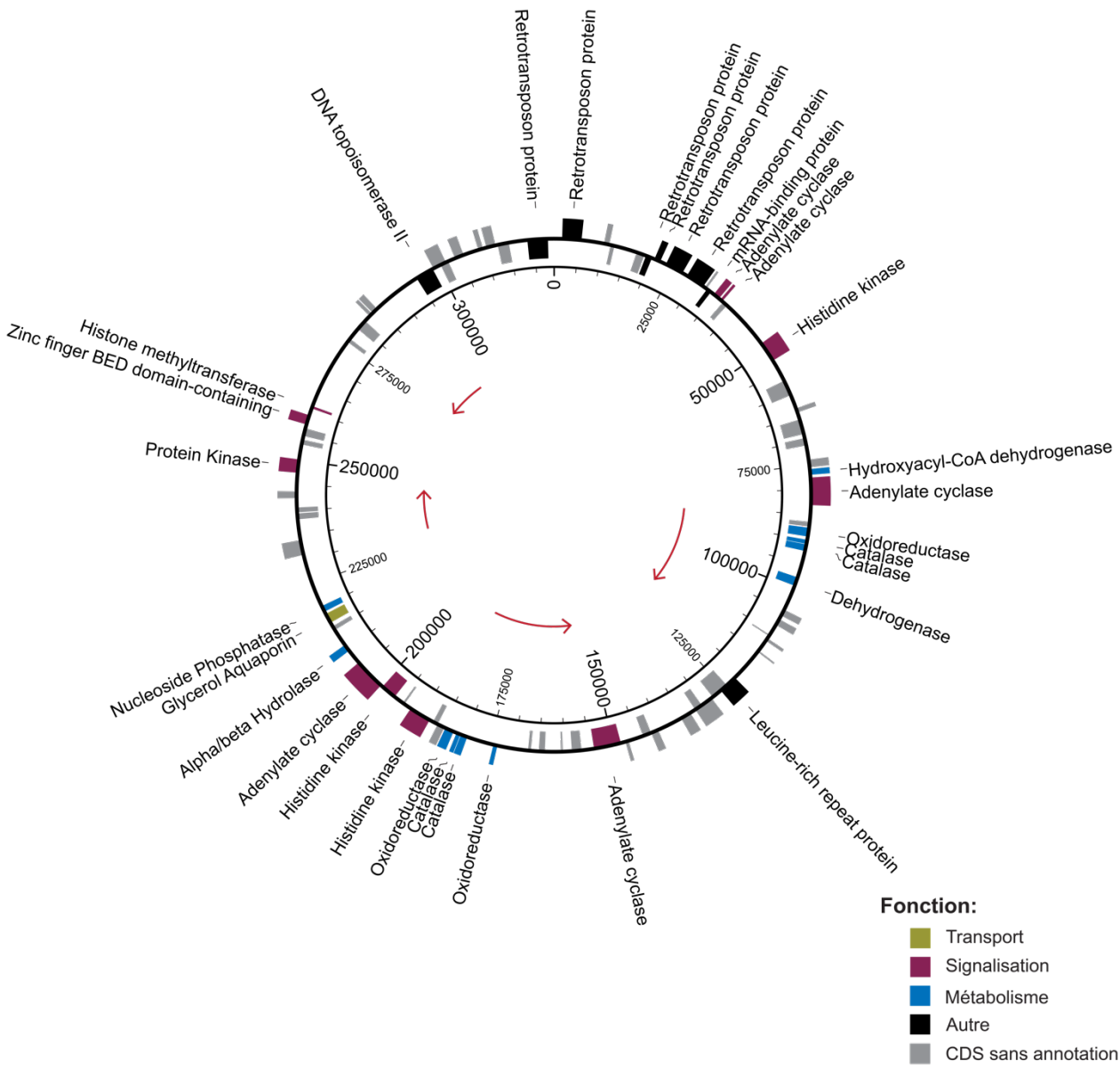

Figure 28. – Carte de gènes du chromosome circulaire 4 de *A. godoyi*. Les gènes ont été annotés selon les annotations des protéines de *A. thaliana*, *H. sapiens* et *S. cerevisiae* identifiées comme homologues à partir d'une recherche avec le profil HMM de la protéine de *A. godoyi*. Les flèches rouges indiquent la position de répétition inversée. La figure a été générée avec Circos (119).

Chapitre 4 – Conclusion

Le processus évolutif responsable de la transition évolutive majeure ayant produit les eucaryotes est loin d'être compris. En effet, l'identification de caractéristiques ancestrales et de processus évolutifs, chez un organisme ayant existé il y a presque deux milliards d'années est une tâche complexe, d'autant plus qu'il n'existe aucun organisme intermédiaire connu entre l'organisme résultant de l'endosymbiose entre une archéobactérie et une α-protéobactérie et le dernier ancêtre commun des eucaryotes. La recherche d'indice chez les organismes vivant actuellement est donc la seule manière de résoudre l'ensemble des questions qui entourent cette transition évolutive. Les connaissances actuelles ont permis d'identifier les caractéristiques potentielles de l'ancêtre des eucaryotes, qui peuvent être facilement déduites en observant les caractéristiques communes aux eucaryotes modernes. Cependant, la définition des étapes intermédiaires est plus complexe puisqu'elle nécessite l'identification de vestiges de ces étapes, qui ont généralement été effacées lors de la mise en place des mécanismes typiques des eucaryotes, plus efficace que ces mécanismes ancestraux. L'étude d'organismes « simples », dont le fonctionnement semble similaire à celui de l'ancêtre des eucaryotes et ayant une évolution lente tel que les jakobides et les malawimonades pourrait cependant permettre d'identifier certaines de ces traces.

La première étape pour ces études est la production d'une collection de génomes de haute qualité qui permet d'analyser et identifier les éléments conservés. Encore plus intéressant, elle pourrait permettre d'identifier des éléments qui ne le sont pas et qui peuvent fournir des indications quant à la mise en place des mécanismes cellulaires des eucaryotes. Ainsi, le développement et l'optimisation d'un pipeline d'assemblage, utilisant les outils Guppy, Flye, Medaka et Pilon suivit d'une correction manuelle a permis de terminer les assemblages de *A. godoyi* et *S. incarcerata* ainsi que de générer des versions préliminaires des assemblages de *R. americana* et *S. ecuadoriensis*. Cependant, le rendement de séquençage nanopore des autres protistes (notamment les malawimonades) est trop faible et du développement supplémentaire de la méthode d'extraction d'ADN pour permettre un meilleur retrait des polysaccharides contaminants sera nécessaire afin de continuer ce projet. Dans l'ensemble, les assemblages nanopore produits présentent une importante augmentation de leur contiguïté ainsi qu'un

meilleur assemblage des régions répétées, ce qui met en valeur l'importance de continuer ces développements.

Chromosomes circulaires de *A. godoyi*

Le génome nucléaire d'*A. godoyi* possède une caractéristique complètement inattendue, soit la présence de quatre chromosomes circulaires. Contrairement aux autres ADN circulaires très rarement présents chez les eucaryotes, ces molécules ne semblent être ni des éléments d'ADN égoïstes (165,166) ni des sous-produits de chromosomes linéaires (167), suggérant qu'il s'agirait d'une catégorie de molécule circulaire différente de celles identifiées précédemment. Cependant, l'absence de telles molécules circulaires chez le voisin phylogénétique d'Andalucia, S. *incarcerata*, indique qu'il ne s'agit pas d'une caractéristique commune à tous les jakobides et donc potentiellement d'une caractéristique spécifique de *A. godoyi*. Cela sera confirmé à la suite de validation manuelle des autres assemblages de jakobides qui permettra de déterminer si d'autres de ces organismes possèdent des chromosomes circulaires.

Quoi qu'il en soit, cette observation soulève cependant bon nombre de questions quant au mécanisme biologique entourant ces molécules. Comment ces molécules sont-elles répliquées et ségrégées lors de la division cellulaire ? Quel est le mécanisme évolutif ayant mené à la mise en place de ces molécules circulaires ? Est-ce que ces molécules proviennent de l'ancêtre eucaryote ou de bactéries ? Afin de répondre à ses questions, il sera nécessaire de mieux caractériser le contenu génique de ces molécules ainsi que l'origine évolutive de ces protéines. De plus, l'identification de mécanismes pouvant permettre la réplication et la ségrégation de ces molécules en mitose sera nécessaire afin de comprendre leur biologie. Finalement, la différence de ploïdie entre les chromosomes linéaires et les chromosomes circulaires

Annotation des génomes des jakobides et malawimonades

L'analyse de l'annotation du génome de *A. godoyi* (notamment présentée dans la section sur les chromosomes circulaires) a montré que l'annotation fonctionnelle des gènes d'un organisme aussi dérivé est un processus complexe. L'absence d'identification claire pour un nombre important de gènes (*e.g., Leucine-rich repeat protein*) limite notre capacité à analyser

rapidement ces annotations. Ainsi, les prochaines étapes de développement devront améliorer la qualité des annotations de l'ensemble du protéome, mais aussi des ARNnc de ces organismes.

Cette identification ne peut cependant pas être basée exclusivement sur le nom de la protéine, qui n'est pas standardisé et souvent peu informatif même chez l'humain. Elle devrait surtout être basée sur une inférence évolutive indiquant l'homologie de la protéine avec d'autres protéines pour lesquelles la fonction biologique est connue. De plus, l'identification de l'origine de chaque gène étudié est nécessaire pour identifier clairement l'histoire évolutive de ces organismes. Cette inférence nécessite donc une capacité à identifier une protéine sur une longue distance évolutive (qui sépare les jakobides, les malawimonades et les eucaryotes pour lesquels une annotation fonctionnelle de qualité est disponible), mais aussi de séparer les paralogues et d'identifier clairement les orthologues. Le développement de méthodes utilisant les profils HMM en cours au laboratoire pourra potentiellement être appliquée à cette problématique et permettre d'améliorer l'annotation de ces organismes.

Références bibliographiques

1. Maynard Smith J, Szathmáry E. The major transitions in evolution. Reprinted. Oxford: Oxford Univ. Press; 1997. 346 p.

2. Chatton E. Titres et travaux scientifiques. Impr. E. Sottano; 1938.

3. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proceedings of the National Academy of Sciences. 1977 Nov 1;74(11):5088–90.

4. Eme L, Spang A, Lombard J, Stairs CW, Ettema TJG. Archaea and the origin of eukaryotes. Nature Reviews Microbiology. 2017 Nov 10;15(12):711–23.

5. Lang BF, Burger G. Mitochondrial and Eukaryotic Origins. In: Advances in Botanical Research. Elsevier; 2012. p. 1–20.

6. Andersson JO. Gene Transfer and Diversification of Microbial Eukaryotes. Annu Rev Microbiol. 2009 Oct;63(1):177–93.

7. Cotton JA, McInerney JO. Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. Proc Natl Acad Sci USA. 2010 Oct 5;107(40):17252–5.

8. Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F, Gelius-Dietrich G, et al. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol Biol Evol. 2004 Sep;21(9):1643–60.

9. Thiergart T, Landan G, Schenk M, Dagan T, Martin WF. An Evolutionary Network of Genes Present in the Eukaryote Common Ancestor Polls Genomes on Eukaryotic and Mitochondrial Origin. Genome Biology and Evolution. 2012;4(4):466–85.

10. Keeling PJ, Palmer JD. Horizontal gene transfer in eukaryotic evolution. Nature Reviews Genetics. 2008 Aug;9(8):605–18.

11. Ford Doolittle W. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends in Genetics. 1998 Dec;14(8):307–11.

12. Rochette NC, Brochier-Armanet C, Gouy M. Phylogenomic Test of the Hypotheses for the Evolutionary Origin of Eukaryotes. Molecular Biology and Evolution. 2014 Apr;31(4):832–45.

13. Koumandou VL, Wickstead B, Ginger ML, van der Giezen M, Dacks JB, Field MC. Molecular paleontology and complexity in the last eukaryotic common ancestor. Critical Reviews in Biochemistry and Molecular Biology. 2013 Jul;48(4):373–96.

14. Dacks JB, Field MC. Evolution of the eukaryotic membrane-trafficking system: origin, tempo and mode. Journal of Cell Science. 2007 Aug 7;120(17):2977–85.

15. Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature. 2015 May 14;521(7551):173–9.

16. Nunoura T, Takaki Y, Kakuta J, Nishi S, Sugahara J, Kazama H, et al. Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. Nucleic Acids Res. 2011 Apr;39(8):3204–23.

17. Guy L, Ettema TJG. The archaeal "TACK" superphylum and the origin of eukaryotes. Trends Microbiol. 2011 Dec;19(12):580–7.

18. Cavalier-Smith T. The Neomuran Revolution and Phagotrophic Origin of Eukaryotes and Cilia in the Light of Intracellular Coevolution and a Revised Tree of Life. Cold Spring Harbor Perspectives in Biology. 2014 Sep 1;6(9):a016006–a016006.

19. Cavalier-Smith T. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. Biology Letters. 2010 Jun 23;6(3):342–5.

20. Derelle R, Torruella G, Klimeš V, Brinkmann H, Kim E, Vlček Č, et al. Bacterial proteins pinpoint a single eukaryotic root. Proc Natl Acad Sci USA. 2015 Feb 17;112(7):E693-699.

21. O'Kelly CJ. The Jakobid Flagellates: Structural Features of Jakoba, Reclinomonas and Histiona and Implications for the Early Diversification of Eukaryotes. The Journal of Eukaryotic Microbiology. 1993 Sep;40(5):627–36.

22. Koreny L, Field MC. Ancient Eukaryotic Origin and Evolutionary Plasticity of Nuclear Lamina. Genome Biol Evol. 2016 Sep;8(9):2663–71.

23. Schlacht A, Herman EK, Klute MJ, Field MC, Dacks JB. Missing Pieces of an Ancient Puzzle: Evolution of the Eukaryotic Membrane-Trafficking System. Cold Spring Harbor Perspectives in Biology. 2014 Oct 1;6(10):a016048–a016048.

24. Field MC, Dacks JB. First and last ancestors: reconstructing evolution of the endomembrane system with ESCRTs, vesicle coat proteins, and nuclear pore complexes. Current Opinion in Cell Biology. 2009 Feb;21(1):4–13.

25. Martin W, Koonin EV. Introns and the origin of nucleus–cytosol compartmentalization. Nature. 2006 Mar;440(7080):41–5.

26. Speijer D, Lukeš J, Eliáš M. Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. Proceedings of the National Academy of Sciences. 2015 Jul 21;112(29):8827–34.

27. Edgcomb VP, Roger AJ, Simpson AGB, Kysela DT, Sogin ML. Evolutionary Relationships Among "Jakobid" Flagellates as Indicated by Alpha- and Beta-Tubulin Phylogenies. Molecular Biology and Evolution. 2001 Apr 1;18(4):514–22.

28. Patterson DJ. *Jakoba libera* (Ruinen, 1938), a heterotrophic flagellate from deep oceanic sediments. J Mar Biol Ass. 1990 May;70(2):381–93.

29. Flavin M, Nerad TA. Reclinomonas americana N. G., N. Sp., a New Freshwater Heterotrophic Flagellate. J Eukaryotic Microbiology. 1993 Mar;40(2):172–9.

30. Lara E, Chatzinotas A, Simpson AGB. Andalucia (n. gen.)--the deepest branch within jakobids (Jakobida; Excavata), based on morphological and molecular study of a new flagellate from soil. J Eukaryot Microbiol. 2006 Apr;53(2):112–20.

31. Burger G, Gray MW, Forget L, Lang BF. Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. Genome Biol Evol. 2013;5(2):418–38.

32. Heiss AA, Kolisko M, Ekelund F, Brown MW, Roger AJ, Simpson AGB. Combined morphological and phylogenomic re-examination of malawimonads, a critical taxon for inferring the evolutionary history of eukaryotes. Royal Society Open Science. 2018 Apr;5(4):171707.

33. Rodríguez-Ezpeleta N, Brinkmann H, Burger G, Roger AJ, Gray MW, Philippe H, et al. Toward Resolving the Eukaryotic Tree: The Phylogenetic Positions of Jakobids and Cercozoans. Current Biology. 2007 Aug;17(16):1420–5.

34. O'Kelly CJ, Nerad TA. Malawimonas jakobiformis n. gen., n. sp. (Malawimonadidae n. fam.): A Jakoba-like Heterotrophic Nanoflagellate with Discoidal Mitochondrial Cristae. The Journal of Eukaryotic Microbiology. 1999 Sep;46(5):522–31.

35. Yabuki A, Gyaltshen Y, Heiss AA, Fujikura K, Kim E. Ophirina amphinema n. gen., n. sp., a New Deeply Branching Discobid with Phylogenetic Affinity to Jakobids. Scientific Reports [Internet]. 2018 Dec [cited 2019 Jun 7];8(1). Available from: http://www.nature.com/articles/s41598-018-34504-6

36. Derelle R, Lang BF. Rooting the Eukaryotic Tree with Mitochondrial and Bacterial Proteins. Molecular Biology and Evolution. 2012 Apr 1;29(4):1277–89.

37. Pánek T, Táborský P, Pachiadaki MG, Hroudová M, Vlček Č, Edgcomb VP, et al. Combined Culture-Based and Culture-Independent Approaches Provide Insights into Diversity of Jakobids, an Extremely Plesiomorphic Eukaryotic Lineage. Front Microbiol [Internet]. 2015 Nov 18 [cited 2022 Aug 29];6. Available from: http://journal.frontiersin.org/Article/10.3389/fmicb.2015.01288/abstract

38. Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB, Lemieux C, et al. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature. 1997 May;387(6632):493–7.

39. Burger G, Jackson CJ, Waller RF. Unusual Mitochondrial Genomes and Genes. In: Bullerwell CE, editor. Organelle Genetics [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012 [cited 2020 Oct 19]. p. 41–77. Available from: http://link.springer.com/10.1007/978-3-642-22380-8_3

40. Shutt TE, Gray MW. Homologs of mitochondrial transcription factor B, sparsely distributed within the eukaryotic radiation, are likely derived from the dimethyladenosine methyltransferase of the mitochondrial endosymbiont. Mol Biol Evol. 2006 Jun;23(6):1169–79.

41. Nishimura Y, Tanifuji G, Kamikawa R, Yabuki A, Hashimoto T, Inagaki Y. Mitochondrial Genome of *Palpitomonas bilix* : Derived Genome Structure and Ancestral System for Cytochrome *c* Maturation. Genome Biol Evol. 2016 Oct;8(10):3090–8.

42. Gray MW, Burger G, Derelle R, Klimeš V, Leger MM, Sarrasin M, et al. The draft nuclear genome sequence and predicted mitochondrial proteome of Andalucia godoyi, a protist with the most gene-rich and bacteria-like mitochondrial genome. BMC Biol. 2020 Dec;18(1):22.

43. Roger AJ, Muñoz-Gómez SA, Kamikawa R. The Origin and Diversification of Mitochondria. Current Biology. 2017 Nov;27(21):R1177–92.

44. Calvo SE, Clauser KR, Mootha VK. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. Nucleic Acids Res. 2016 Jan 4;44(D1):D1251–7.

45. Morgenstern M, Stiller SB, Lübbert P, Peikert CD, Dannenmaier S, Drepper F, et al. Definition of a High-Confidence Mitochondrial Proteome at Quantitative Scale. Cell Reports. 2017 Jun;19(13):2836–52.

46. Desmond E, Brochier-Armanet C, Forterre P, Gribaldo S. On the last common ancestor and early evolution of eukaryotes: reconstructing the history of mitochondrial ribosomes. Research in Microbiology. 2011 Jan;162(1):53–70.

47. Palmer T, Berks BC. The twin-arginine translocation (Tat) protein export pathway. Nat Rev Microbiol. 2012 Jul;10(7):483–96.

48. Fulnečková J, Ševčíková T, Fajkus J, Lukešová A, Lukeš M, Vlček Č, et al. A Broad Phylogenetic Survey Unveils the Diversity and Evolution of Telomeres in Eukaryotes. Genome Biology and Evolution. 2013 Mar;5(3):468–83.

49. Fleischmann R, Adams M, White O, Clayton R, Kirkness E, Kerlavage A, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995 Jul 28;269(5223):496–512.

50. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics. 2012 May;13(5):329–42.

51. Teng H, Cao MD, Hall MB, Duarte T, Wang S, Coin LJM. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. GigaScience. 2018 May 1;7(5):giy037.

52. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2 [Internet]. Bioinformatics; 2019 Jan [cited 2019 Sep 19]. Available from: http://biorxiv.org/lookup/doi/10.1101/530972

53. Chen TW, Gan RC, Fang YK, Chien KY, Liao WC, Chen CC, et al. FunctionAnnotator, a versatile and efficient web tool for non-model organism annotation. Sci Rep. 2017 Dec;7(1):10430.

54. Haas LW, Webb KL. Nutritional mode of several non-pigmented microflagellates from the York River estuary, Virginia. Journal of Experimental Marine Biology and Ecology. 1979 Jun;39(2):125–34.

55. Metzker ML. Sequencing technologies — the next generation. Nature Reviews Genetics. 2010 Jan;11(1):31–46.

56. Steinbock LJ, Radenovic A. The emergence of nanopores in next-generation sequencing. Nanotechnology. 2015 Feb 20;26(7):074003.

57. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends in Genetics. 2014 Sep;30(9):418–26.

58. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016 Jun;17(6):333–51.

59. Ip CLC, Loose M, Tyson JR, de Cesare M, Brown BL, Jain M, et al. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Research. 2015 Oct 15;4:1075.

60. Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Research. 2018 Mar 16;46(5):2159–68.

61. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods. 2017 Apr;14(4):407–10.

62. Kasianowicz JJ, Brandin E, Branton D, Deamer DW. Characterization of individual polynucleotide molecules using a membrane channel. Proceedings of the National Academy of Sciences. 1996 Nov 26;93(24):13770–3.

63. Brown CG, Clarke J. Nanopore development at Oxford Nanopore. Nature Biotechnology. 2016 Aug;34(8):810–1.

64. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, et al. The potential and challenges of nanopore sequencing. Nature Biotechnology. 2008 Oct;26(10):1146–53.

65. Stoddart D, Heron AJ, Mikhailova E, Maglia G, Bayley H. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. Proceedings of the National Academy of Sciences. 2009 May 12;106(19):7702–7.

66. Teng H, Cao MD, Hall MB, Duarte T, Wang S, Coin LJM. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. GigaScience. 2018 May 1;7(5):giy037.

67. Boža V, Perešíni P, Brejová B, Vinař T. DeepNano-blitz: a fast base caller for MinION nanopore sequencers. Birol I, editor. Bioinformatics. 2020 Aug 15;36(14):4191–2.

68. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 2019 Dec;20(1):129.

69. Xu L, Seki M. Recent advances in the detection of base modifications using the Nanopore sequencer. J Hum Genet. 2020 Jan;65(1):25–33.

70. Rice ES, Green RE. New Approaches for Genome Assembly and Scaffolding. Annual Review of Animal Biosciences. 2019 Feb 15;7(1):17–40.

71. Pevzner PA, Tang H. Fragment assembly with double-barreled data. Bioinformatics. 2001 Jun 1;17(Suppl 1):S225–33.

72. Myers EW. Toward Simplifying and Accurately Formulating Fragment Assembly. Journal of Computational Biology. 1995 Jan;2(2):275–90.

73. Cherukuri Y, Janga SC. Benchmarking of de novo assembly algorithms for Nanopore data reveals optimal performance of OLC approaches. BMC Genomics. 2016 Aug;17(S7):507.

74. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature Biotechnology. 2018 Jan 29;36(4):338–45.

75. Tyson JR, O'Neil NJ, Jain M, Olsen HE, Hieter P, Snutch TP. Whole genome sequencing and assembly of a Caenorhabditis elegans genome with complex genomic rearrangements using the MinION sequencing device. bioRxiv [Internet]. 2017 Jan 30 [cited 2019 May 16]; Available from: http://biorxiv.org/lookup/doi/10.1101/099143

76. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics. 2016 Apr 1;32(7):1009–15.

77. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. Phillippy AM, editor. PLoS Comput Biol. 2017 Jun 8;13(6):e1005595.

78. Zimin AV, Puiu D, Luo MC, Zhu T, Koren S, Marçais G, et al. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii* , a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res. 2017 May;27(5):787–92.

79. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019 May;37(5):540–6.

80. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive *k* -mer weighting and repeat separation. Genome Res. 2017 May;27(5):722–36.

81. Baeza JA, García-De León FJ. Are we there yet? Benchmarking low-coverage nanopore long-read sequencing for the assembling of mitochondrial genomes using the vulnerable silky shark Carcharhinus falciformis. BMC Genomics. 2022 Dec;23(1):320.

82. Zimin AV, Salzberg SL. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. Ouzounis CA, editor. PLoS Comput Biol. 2020 Jun 26;16(6):e1007981.

83. Chen Z, Erickson DL, Meng J. Benchmarking Long-Read Assemblers for Genomic Analyses of Bacterial Pathogens Using Oxford Nanopore Sequencing. Int J Mol Sci. 2020 Dec 1;21(23):E9161.

84. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. Wang J, editor. PLoS ONE. 2014 Nov 19;9(11):e112963.

85. Kundu R, Casey J, Sung WK. HyPo: Super Fast & Accurate Polisher for Long Read Genome Assemblies [Internet]. Bioinformatics; 2019 Dec [cited 2022 Aug 1]. Available from: http://biorxiv.org/lookup/doi/10.1101/2019.12.19.882506

86. Haghshenas E, Asghari H, Stoye J, Chauve C, Hach F. HASLR: Fast Hybrid Assembly of Long Reads. iScience. 2020 Aug;23(8):101389.

87. Vaser R, Šikić M. Raven: a de novo genome assembler for long reads [Internet]. Bioinformatics; 2020 Aug [cited 2020 Oct 21]. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.08.07.242461

88. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat Biotechnol. 2020 Sep;38(9):1044–53.

89. Wick RR, Holt KE. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. F1000Res. 2020 Sep 17;8:2138.

90. Fu S, Wang A, Au KF. A comparative evaluation of hybrid error correction methods for error-prone long reads. Genome Biol. 2019 Dec;20(1):26.

91. Zhang H, Jain C, Aluru S. A comprehensive evaluation of long read error correction methods. BMC Genomics. 2020 Dec 21;21(Suppl 6):889.

92. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Research. 2012 Mar 1;22(3):557–67.

93. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013 Apr 15;29(8):1072–5.

94. Parra G, Bradnam K, Ning Z, Keane T, Korf I. Assessing the gene space in draft genomes. Nucleic Acids Research. 2009 Jan;37(1):289–97.

95. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 Oct 1;31(19):3210–2.

96. Leger MM, Eme L, Hug LA, Roger AJ. Novel Hydrogenosomes in the Microaerophilic Jakobid *Stygiella incarcerata*. Mol Biol Evol. 2016 Sep;33(9):2318–36.

97. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018 Sep 1;34(17):i884–90.

98. Song L, Florea L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. GigaSci. 2015 Dec;4(1):48.

99. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2016 Jul 15;32(14):2103–10.

100. Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics. 2016 Oct 22;btw663.

101. Patterson M, Marschall T, Pisanti N, van Iersel L, Stougie L, Klau GW, et al. W HATS H AP : Weighted Haplotype Assembly for Future-Generation Sequencing Reads. Journal of Computational Biology. 2015 Jun;22(6):498–509.

102. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094–100.

103. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics and Bioinformatics. 2021 Jan 6;3(1):lqaa108.

104. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.

105. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21.

106. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011 May 15;29(7):644–52.

107. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003 Oct 1;31(19):5654–66.

108. Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 2006 Feb 9;7:62.

109. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004 May 14;5:59.

110. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 2014 Sep;42(15):e119.

111. Testa AC, Hane JK, Ellwood SR, Oliver RP. CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC Genomics. 2015 Mar 11;16:170.

112. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008 Jan 11;9(1):R7.

113. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019 Dec;20(1):275.

114. Smit AFA, Hubley R, Green P. RepeatMasker [Internet]. 2013. Available from: http://www.repeatmasker.org

115. Chan PP, Lowe TM. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. In: Kollmar M, editor. Gene Prediction [Internet]. New York, NY: Springer New York; 2019 [cited 2022 Aug 18]. p. 1–14. (Methods in Molecular Biology; vol. 1962). Available from: http://link.springer.com/10.1007/978-1-4939-9173-0_1

116. Eddy SR. Accelerated Profile HMM Searches. Pearson WR, editor. PLoS Computational Biology. 2011 Oct 20;7(10):e1002195.

117. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution. 2013 Apr 1;30(4):772–80.

118. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution. 2015 Jan;32(1):268–74.

119. Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res. 2009 Sep;19(9):1639–45.

120. Sholes SL, Karimian K, Gershman A, Kelly TJ, Timp W, Greider CW. Chromosome-specific telomere lengths and the minimal functional telomere revealed by nanopore sequencing. Genome Res. 2022 Apr;32(4):616–28.

121. Song G, Dickins BJA, Demeter J, Engel S, Dunn B, Cherry JM. AGAPE (Automated Genome Analysis PipelinE) for Pan-Genome Analysis of Saccharomyces cerevisiae. Schacherer J, editor. PLoS ONE. 2015 Mar 17;10(3):e0120671.

122. Ding Q, Li R, Ren X, Chan L yan, Ho VWS, Xie D, et al. Genomic architecture of 5S rDNA cluster and its variations within and between species. BMC Genomics. 2022 Dec;23(1):238.

123. Cook DE, Zdraljevic S, Roberts JP, Andersen EC. CeNDR, the *Caenorhabditis elegans* natural diversity resource. Nucleic Acids Res. 2017 Jan 4;45(D1):D650–7.

124. Konishi H, Yamaguchi R, Yamaguchi K, Furukawa Y, Imoto S. Halcyon: an accurate basecaller exploiting an encoder–decoder model with monotonic attention. Inanc B, editor. Bioinformatics. 2021 Jun 9;37(9):1211–7.

125. Chen Y, Nie F, Xie SQ, Zheng YF, Dai Q, Bray T, et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat Commun. 2021 Dec;12(1):60.

126. Lee H, Gurtowski J, Yoo S, Marcus S, McCombie WR, Schatz M. Error correction and assembly complexity of single molecule sequencing reads. [Internet]. Bioinformatics; 2014 Jun [cited 2022 Jul 23]. Available from: http://biorxiv.org/lookup/doi/10.1101/006395

127. Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction. Bioinformatics. 2014 Dec 15;30(24):3506–14.

128. Miclotte G, Heydari M, Demeester P, Rombauts S, Van de Peer Y, Audenaert P, et al. Jabba: hybrid error correction for long sequencing reads. Algorithms Mol Biol. 2016 Dec;11(1):10.

129. Wang JR, Holt J, McMillan L, Jones CD. FMLRC: Hybrid long read error correction using an FM-index. BMC Bioinformatics. 2018 Dec;19(1):50.

130. Morisse P, Lecroq T, Lefebvre A. Hybrid correction of highly noisy long reads using a variable-order de Bruijn graph. Berger B, editor. Bioinformatics. 2018 Dec 15;34(24):4213–22.

131. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol. 2012 Jul;30(7):693–700.

132. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017 May;27(5):737–46.

133. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Apr;9(4):357–9.

134. Firtina C, Kim JS, Alser M, Senol Cali D, Cicek AE, Alkan C, et al. Apollo: a sequencing-technology-independent, scalable and accurate assembly polishing algorithm. Cowen L, editor. Bioinformatics. 2020 Jun 1;36(12):3669–79.

135. Warren RL, Coombe L, Mohamadi H, Zhang J, Jaquish B, Isabel N, et al. ntEdit: scalable genome sequence polishing. Berger B, editor. Bioinformatics. 2019 Nov 1;35(21):4430–2.

136. Howison M, Zapata F, Dunn CW. Toward a statistically explicit understanding of de novo sequence assembly. Bioinformatics. 2013 Dec 1;29(23):2959–63.

137. Wetzel J, Kingsford C, Pop M. Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies. BMC Bioinformatics. 2011 Dec;12(1):95.

138. Ricker N, Qian H, Fulthorpe RR. The limitations of draft assemblies for understanding prokaryotic adaptation and evolution. Genomics. 2012 Sep;100(3):167–75.

139. Delahaye C, Nicolas J. Sequencing DNA with nanopores: Troubles and biases. Andrés-León E, editor. PLoS ONE. 2021 Oct 1;16(10):e0257521.

140. Dokos R, Stoddart D, Reid S. Blocking, Unblocking and Flow Cell Output [Internet]. Blocking, Unblocking and Flow Cell Output. 2019 [cited 2022 Jul 14]. Available from: https://community.nanoporetech.com/posts/blocking-unblocking-and-f

141. Remaut H, Van Der Verren S, Van Gerven N, Nishantha Jayasinghe L, Jayne Wallace E, Raj Singh P, et al. Pore. 20220056517.

142. Heron AJ, Graham JE, Gutierrez RA, Rebecca Victoria B, White J, Brown CG, et al. Method for nucleic acid detection by guiding through a nanopore. 20220064723, 2022.

143. Schmidt M. Nanopore Sequencing in Plants: From Greenhouse to Genome. 2020 Feb 18.

144. Doyle J, Doyle J. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. 1987;19:11–5.

145. Healey A, Furtado A, Cooper T, Henry RJ. Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. Plant Methods. 2014;10(1):21.

146. Haghshenas E, Hach F, Sahinalp SC, Chauve C. CoLoRMap: Correcting Long Reads by Mapping short reads. Bioinformatics. 2016 Sep 1;32(17):i545–51.

147. Morisse P, Marchet C, Limasset A, Lecroq T, Lefebvre A. Scalable long read self-correction and assembly polishing with multiple sequence alignment. Sci Rep. 2021 Dec;11(1):761.

148. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods. 2011 Jan;8(1):61–5.

149. Green P. Whole-genome disassembly. Proc Natl Acad Sci USA. 2002 Apr 2;99(7):4143–4.

150. Schrader L, Schmitz J. The impact of transposable elements in adaptive evolution. Mol Ecol. 2019 Mar;28(6):1537–49.

151. Faino L, Seidl MF, Shi-Kunne X, Pauper M, van den Berg GCM, Wittenberg AHJ, et al. Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. Genome Res. 2016 Aug;26(8):1091–100.

152. Sun J, Li R, Chen C, Sigwart JD, Kocot KM. Benchmarking Oxford Nanopore read assemblers for high-quality molluscan genomes. Phil Trans R Soc B. 2021 May 24;376(1825):20200160.

153. Zhang X, Liu CG, Yang SH, Wang X, Bai FW, Wang Z. Benchmarking of long-read sequencing, assemblers and polishers for yeast genome. Briefings in Bioinformatics. 2022 May 13;23(3):bbac146.

154. Wick RR, Judd LM, Cerdeira LT, Hawkey J, Méric G, Vezina B, et al. Trycycler: consensus long-read assemblies for bacterial genomes. Genome Biol. 2021 Dec;22(1):266.

155. Watson M, Warr A. Errors in long-read assemblies can critically affect protein prediction. Nat Biotechnol. 2019 Feb;37(2):124–6.

156. Wick RR, Holt KE. Polypolish: Short-read polishing of long-read bacterial genome assemblies. Schneidman-Duhovny D, editor. PLoS Comput Biol. 2022 Jan 24;18(1):e1009802.

157. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods. 2015 Aug;12(8):733–5.

158. Lee JY, Kong M, Oh J, Lim J, Chung SH, Kim JM, et al. Comparative evaluation of Nanopore polishing tools for microbial genome assembly and polishing strategies for downstream analysis. Sci Rep. 2021 Dec;11(1):20740.

159. Lewis WH. Polyploidy: Biological Relevance [Internet]. Boston, MA: Springer US; 1980 [cited 2022 Jun 22]. Available from: https://doi.org/10.1007/978-1-4613-3069-1

160. Galitski T, Saldanha AJ, Styles CA, Lander ES, Fink GR. Ploidy Regulation of Gene Expression. Science. 1999 Jul 9;285(5425):251–4.

161. Shafin K, Pesout T, Chang PC, Nattestad M, Kolesnikov A, Goel S, et al. Haplotype-aware variant calling enables high accuracy in nanopore long-reads using deep neural networks [Internet]. Bioinformatics; 2021 Mar [cited 2022 Jul 27]. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.03.04.433952

162. Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, de la Cruz F. Mobility of Plasmids. Microbiol Mol Biol Rev. 2010 Sep;74(3):434–52.

163. Garcillán-Barcia MP, Francia MV, de La Cruz F. The diversity of conjugative relaxases and its application in plasmid classification. FEMS Microbiol Rev. 2009 May;33(3):657–87.

164. Shintani M, Sanchez ZK, Kimbara K. Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. Front Microbiol [Internet]. 2015 Mar 31 [cited 2022 Jul 28];6. Available from: http://www.frontiersin.org/Evolutionary_and_Genomic_Microbiology/10.3389/fmicb.2015.00242/abstract

165. Griffith AJF. Natural Plasmids of Filamentous Fungi. Microbiological Reviews. 1995;59(4):673–85.

166. Chan KM, Liu YT, Ma CH, Jayaram M, Sau S. The 2 micron plasmid of Saccharomyces cerevisiae: A miniaturized selfish genome with optimized functional competence. Plasmid. 2013 Jul;70(1):2–17.

167. Cohen S, Segal D. Extrachromosomal Circular DNA in Eukaryotes: Possible Involvement in the Plasticity of Tandem Repeats. Cytogenet Genome Res. 2009;124(3–4):327–38.

168. Gaubatz JW. Extrachromosomal circular DNAs and genomic sequence plasticity in eukaryotic cells. Mutation Research/DNAging. 1990 Sep;237(5–6):271–92.

169. Shoura MJ, Gabdank I, Hansen L, Merker J, Gotlib J, Levene SD, et al. Intricate and Cell Type-Specific Populations of Endogenous Circular DNA (eccDNA) in *Caenorhabditis elegans* and *Homo sapiens*. G3 Genes|Genomes|Genetics. 2017 Oct 1;7(10):3295–303.

170. Møller HD, Parsons L, Jørgensen TS, Botstein D, Regenberg B. Extrachromosomal circular DNA is common in yeast. Proc Natl Acad Sci USA [Internet]. 2015 Jun 16 [cited 2022 Jul 28];112(24). Available from: https://pnas.org/doi/full/10.1073/pnas.1508825112

171. Cohen S, Yacobi K, Segal D. Extrachromosomal Circular DNA of Tandemly Repeated Genomic Sequences in *Drosophila*. Genome Res. 2003 Jun;13(6a):1133–45.

172. Warnow T. Computational Phylogenetics: An Introduction to Designing Methods for Phylogeny Estimation [Internet]. 1st ed. Cambridge University Press; 2017 [cited 2020 Jan 19]. Available from: https://www.cambridge.org/core/product/identifier/9781316882313/type/book

173. Bern M, Goldberg D. Automatic selection of representative proteins for bacterial phylogeny. BMC Evol Biol. 2005;5(1):34.