

THESE DE DOCTORAT DE

L'UNIVERSITE DE NANTES
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 598
Sciences de la Mer et du littoral
Spécialité : Biologie des organismes

Par

Jérémy BERTHELIER

**Étude de l'activité des éléments mobiles chez
Tisochrysis lutea dans un contexte d'amélioration
des espèces de microalgues.**

Thèse présentée et soutenue à Nantes le 11 décembre 2018
Unité de recherche : Laboratoire Physiologie et Biotechnologie des Algues, Ifremer
Laboratoire Mer Molécule Santé, Université du Mans

Thèse N° :

Rapporteurs avant soutenance :

Hervé Moreau
François Sabot

Directeur de recherche HDR, CNRS
Directeur de recherche HDR, IRD

Composition du Jury :

Examineurs : Jérémie Bourdon
Emmanuelle Lerat
Florian Maumus

Professeur HDR, Université de Nantes
Chargée de recherche HDR, CNRS
Chargé de recherche HDR, INRA

Directrice de thèse : Nathalie Casse
Co-enc. de thèse : Grégory Carrier

Maitre de conférences HDR, Université du Mans
Chargé de recherche, Ifremer

Table des matières

Table des matières	1
Liste des Figures	6
Liste des Tableaux	9
Abréviations	10
Liste des publications et communications	12
Remerciements	16
Préambule	18
Introduction générale	21
1. Les microalgues	22
1.1. Rôles écologiques du phytoplancton	23
1.2. La diversité des microalgues	26
1.3. Origine des microalgues	28
1.4. Les taxons de microalgues les plus étudiés	30
1.5. Le phylum des Haptophytes	32
2. La génomique	38
2.1. Historique de la génomique	39
2.2. La génomique chez les microalgues	43
3. Les éléments transposables (ETs)	45
3.1. Découverte et rôle des ETs	46
3.2. Définition, ubiquité et proportions des ETs	48
3.3. La classification des ETs:	51
3.4. Le cycle de vie des ETs	63
3.5. Les mécanismes de régulation des ETs	64
3.6. Impacts des ETs sur les génomes	66
3.7. Les ETs : moteurs de l'évolution adaptative des espèces	73
3.8. L'activité des ETs chez les microalgues	76
Contexte de l'étude	81

1. Les microalgues : ressource d’avenir pour les biotechnologies	83
1.1. Alimentation	83
1.2. Remédiation	84
1.3. Energie	84
1.4. Molécules actives	85
1.5. Des verrous à lever pour l’utilisation des microalgues	86
2. La domestication des microalgues	88
3. Présentation de l’espèce d’étude : <i>Tisochrysis lutea</i>	92
Objectifs de la thèse	96
Chapitre 1 Annotation et étude de l’activité des ETs de <i>T. lutea</i>	99
1. Un nouvel assemblage du génome de <i>T. lutea</i>	102
1.1. Contexte de l’étude	102
1.2. Matériel génétique	104
1.3. Stratégie de séquençage	104
1.4. Assemblage du génome de <i>T. lutea</i>	106
2. Elaboration d’un pipeline d’annotation des ETs de <i>T. lutea</i>	108
2.1. Introduction	108
2.2. Détection de séquences d’ETs	109
2.3. Réduction du nombre de séquences détectées	127
2.4. Classification des séquences détectées	127
2.5. Création des bibliothèques de séquences d’ETs référents	130
2.6. Annotation des ETs de <i>T. lutea</i>	133
3. Activité des ETs autonomes chez <i>T. lutea</i>	134
4. A transposable element annotation pipeline and expression analysis reveal potentially active elements in the microalga <i>T. lutea</i>	135
5. Résultats complémentaires	151
5.1. Identification de séquences similaire à des IS chez <i>T. lutea</i>	151
5.2. Etude de l’expression des ETs à partir d’un nouvel assemblage transcriptomique de <i>T. lutea</i>	153
5.3. Distribution et amélioration du pipeline PiRATE	154
6. Conclusion	157

Chapitre 2 Etude de la dynamique des ETs chez deux souches domestiquées de <i>T. lutea</i>	161
1. Obtention de données génomiques pour des souches domestiquées et sauvage de <i>T. lutea</i>	164
1.1. Origine des souches	164
1.2. Séquençage des souches	168
1.3. Alignement des données de séquençage sur le génome de référence	168
2. Etat de l'art des approches visant la prédiction d'indels	172
2.1. Approches permettant la prédiction d'indels à l'aide de données <i>short reads</i>	172
2.2. Outils sélectionnés pour prédire des indels associées à des ETs à l'aide de données <i>short reads</i>	176
2.3. Outil sélectionné pour prédire des indels associées à des ETs à l'aide de données <i>long reads</i>	177
3. Prédiction d'indels associées aux ETs et spécifiques aux souches domestiquées de <i>T. lutea</i>	180
3.1. Prédiction d'indels associées à des ETs avec des outils utilisant des données <i>short reads</i>	180
3.2. Prédiction d'indels associées à des ETs avec un outil utilisant des données <i>long reads</i>	183
4. Identification des ETs les plus dynamiques dans les génomes des souches domestiquées de <i>T. lutea</i>	186
4.1. Proportion des familles d'ETs associées à des indels	186
4.2. Activité prépondérante de trois familles d'ETs	188
4.3. Caractérisation des familles d'ETs les plus dynamiques	189
5. Existe-t-il un lien entre les indels associées aux ETs et les phénotypes des souches domestiquées ?	194
5.1. Recherche de gènes potentiellement impactés par les indels	194
5.2. Analyse des gènes candidats associés au métabolisme des lipides neutres	197
6. Conclusion	200
Conclusion générale et perspectives	203

Références	213
Annexes	241
Annexe 1	242
Annexe 2	247
Annexe 3	251
Annexe 4	254
Annexe 5	256
Annexe 6	258
Annexe 7	260
Annexe 8	262
Annexe 9	267
Annexe 10	269
Annexe 11	271

Liste des Figures

Figure 1 : Représentation d'organismes aquatiques de l'antiquité au 19 ^{ème} siècle.	19
Figure 2 : Illustration de la pompe biologique et de la pompe physique du carbone.	25
Figure 3 : Arbre phylogénétique représentant la distribution des taxons de microalgues.	27
Figure 4 : Représentation schématique de l'origine endosymbiotique des microalgues.	29
Figure 5 : Aperçu de la diversité phénotypique des microalgues.	31
Figure 6 : Représentation schématique de la composition d'une cellule Haptophyte.	33
Figure 7 : Contribution des Haptophytes, Diatomées, et Cyanobactéries à la biomasse photosynthétique de l'année 2000.	33
Figure 8 : Visualisation par microscopie de cellules d'Haptophytes.	35
Figure 9 : Phases de vie des différentes familles d'Haptophytes.	35
Figure 10 : Photographie satellite d'une efflorescence de la microalgue Haptophyte <i>Emiliania huxleyi</i> à proximité des côtes de la Grande-Bretagne.	36
Figure 11 : Représentation des séquences de l'ADN en bio-informatique.	38
Figure 12 : Aperçu visuel de l'évolution des technologies de séquençage de l'ADN.	42
Figure 13 : Barbara McClintock.	45
Figure 14 : Grains de maïs présentant des phénotypes instables.	47
Figure 15 : Proportion d'ETs dans le génome de différents organismes.	49
Figure 16 : Corrélation entre la taille des génomes (ADN, picogramme) et la proportion en ETs qu'ils possèdent.	50
Figure 17 : Système de classification des ETs proposé par Wicker et al. en 2007.	53
Figure 18 : Représentation schématique du mécanisme de transposition d'un rétrotransposon LTR Copia.	57
Figure 19 : Illustration de la relation entre les LTR, LARD et TRIM.	57
Figure 20 : Illustration des mécanismes de transposition des LINE et des SINE.	59
Figure 21 : Illustration du mécanisme de transposition d'un TIR Mariner.	61
Figure 22 : Illustration de deux mécanismes permettant la duplication d'un élément TIR.	62
Figure 23 : Illustration de la relation entre des TIR autonomes et non-autonomes.	62
Figure 24 : Schéma représentatif du cycle de vie théorique des ETs.	65
Figure 25 : Exemple de différents impacts pouvant être causés par l'insertion de novo d'un ET dans un gène ou à sa proximité.	68

Figure 26 : Impacts d'un rétrotransposon LTR sur la couleur de différentes variétés de vigne.	69
Figure 27 : Impact de l'inactivation de l'ET domestiqué <i>Mustang</i> sur la physiologie d' <i>Arabidopsis thaliana</i> .	72
Figure 28 : Photographie par microscopie illustrant l'impact de l'insertion d'un ET dans le gène <i>InvA</i> de la microalgue multicellulaire <i>V. carteri</i> .	77
Figure 29 : Impact de l'insertion d'un ET dans le gène <i>MetE</i> dans le génome d'une souche clonale de <i>C. reinhardtii</i> .	77
Figure 30 : Facteurs les plus influents sur le coût de production d'aliments à base de microalgues.	87
Figure 31 : Photographie au microscope de plusieurs cellules de <i>T. lutea</i> .	94
Figure 32 : Position phylogénique de l'espèce <i>T. lutea</i> chez les Haptophytes.	94
Figure 33 : Distribution géographique des différentes souches isolées de <i>T. lutea</i> .	95
Figure 34 : Illustration schématique de l'assemblage d'une région d'un génome présentant deux copies d'un ET à l'aide de données <i>short reads</i> ou <i>long reads</i> .	103
Figure 35 : Schéma illustrant la notion de profondeur de séquençage d'une région d'un génome.	105
Figure 36 : Représentation schématique des différentes étapes du pipeline PiRATE.	112
Figure 37 : Méthode de construction de profils HMMs (Hidden Markov Models).	115
Figure 38 : Représentation schématique des paramètres recherchés par LTRharvest.	117
Figure 39 : Représentation des motifs et structures recherchés par SINE-Finder.	117
Figure 40 : Représentation des motifs et structures recherchés par HelSearch.	119
Figure 41 : Illustration de la méthodologie de RepeatScout.	121
Figure 42 : Illustration de la méthodologie de RepeatExplorer.	123
Figure 43 : Représentation schématique de la méthodologie de RepARK.	125
Figure 44 : Représentation schématique des paramètres de CD-HIT-est.	128
Figure 45 : Représentation schématique de la méthode de construction des trois bibliothèques en « poupées russes ».	131
Figure 46 : Visualisation de l'interface du PiRATE-Galaxy (version 1).	155
Figure 47 : Illustration des étapes des programmes d'amélioration utilisées pour obtenir les deux souches domestiquées de <i>T. lutea</i> .	166
Figure 48 : Illustration de l'alignement de <i>short reads</i> en paires (<i>paired-end</i>) et de <i>long reads</i> sur un génome de référence.	169

Figure 49 : Illustration des approches d'identification d'indels à l'aide de données de séquençage <i>short reads</i> . _____	173
Figure 50 : Illustration du pipeline employé pour prédire des indels associées à des ETs qui sont spécifiques aux souches domestiquées. _____	181
Figure 51 : Pourcentage d'indels associées aux familles d'ETs dans le génome des deux souches domestiquées de <i>T. lutea</i> . _____	187
Figure 52 : Représentation schématique des séquences autonomes des familles <i>Luffy</i> , <i>Ace</i> et <i>Shanks</i> . _____	193

Liste des Tableaux

Tableau 1 : Comparaison de critères entre des instruments de séquençage Illumina et Oxford Nanopore. _____	41
Tableau 2 : Aperçu des corrections réalisées par les outils Quiver et Pilon. _____	107
Tableau 3 : Liste des outils disponibles visant la détection de séquences d’ETs. _____	110
Tableau 4 : Nombre de séquences potentielles d’ETs détectées par les outils, sans et avec filtres. _____	126
Tableau 5 : ETs identifiés dans un transcriptome de <i>T. lutea</i> , obtenu sous condition de forte concentration en iode. _____	153
Tableau 6 : Résultats d’alignements des short reads et long reads. _____	170
Tableau 7 : Synthèse des caractéristiques des outils choisies pour conduire les prédictions d’indels associées aux ETs. _____	178
Tableau 8 : Résultats obtenus avec Pindel et Breakdancer pour les génomes des souches Tiso-Arg, Tiso-LipidePlus et TisoLipideMoins. _____	180
Tableau 9 : Prédictions d’événements de délétions spécifiques à la souche Tiso-LipidePlus. _____	182
Tableau 10 : Prédictions d’événements d’insertions consensus entre LoRTE et Mobster pour la souche Tiso-LipidePlus. _____	185
Tableau 11 : Liste des gènes candidats pour la souche Tiso-LipidePlus. _____	195
Tableau 12 : Liste des gènes candidats pour la souche Tiso-LipideMoins. _____	196

Abréviations

ABI : Applied Biosystems

ADN : acide désoxyribonucléique

ADNc : acide désoxyribonucléique complémentaire

ARN : acide ribonucléique

ANSES : Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail.

AP : aspartique

BLAST : *Basic Local Alignment Search Tool*

CEA : Commissariat à l'Énergie Atomique et aux Énergies Alternatives

DIRS : *Dictyostelium-like element*

DHA : acide docosahexaénoïque

EMS : méthanesulfonate d'éthyle

EMBL-EBI : *European Molecular Biology Laboratory - European Bioinformatics Institute*

EPA : acide eicosapentaénoïque

ERV : *Endogenous retrovirus*

ET : élément transposable

Ifremer : Institut Français de Recherche pour l'Exploitation de la Mer

IN : intégrase

indels : insertions et délétions de séquences nucléotidiques

INRA : Institut national de la recherche agronomique

IRHS : l'Institut de Recherche en Horticulture et Semences

IS : *Insertional Sequences*

GO : ontologie de gène

HMMs : *Hidden Markov Models*

HSP : *Heat-Shock Proteins*

HTT : *Horizontal Transfer of Transposable elements*

LARD : *L*Arge *R*etrotransposon *D*erivative

LINE : *L*ong *I*nterspersed *N*uclear *E*lement

LTR : *L*ong *T*erminal *R*epeats

MITE : *M*iniature *I*nverted-repeat *T*ransposable *E*lements

MNNG : méthylnitronitrosoguanidine

NCBI : *N*ational *C*enter for *B*iotechnology *I*nformation

NF1 : neurofibromatose 1

ORF : *O*pen *R*eadin*F*rame

PASTEC : Pseudo Agent System for Transposables Elements Classification

PiRATE : *P*ipeline to *R*etrieve and *A*nnotate *T*ransposable *E*lements

PLE : *P*enelope-*L*ike *E*lements

POL : polyprotein

OGM : organismes génétiquement modifiés

pb : paire de bases

RH : Ribonucléase H

RT : rétro-transcriptase

SINE : *S*hort *I*nterspersed *N*uclear *E*lement

SNP : polymorphismes nucléotidiques

TAG : triacylglycérol

TASE : transposase

TIR : *T*erminal *I*nverted *R*epeats

TRIM : *T*erminal-repeat *R*etrotransposons *I*n *M*iniature

TSD : *T*erminal *S*ide *D*uplication

UTR : *U*ntranslated *T*ranscribed *R*egion

Liste des publications et communications

Publication concernant le sujet de thèse

- Berthelie, J., Casse, N., Daccord, N., Jamilloux, V., Saint-Jean, B., & Carrier, G. (2018). A transposable element annotation pipeline and expression analysis reveal potentially active elements in the microalga *Tisochrysis lutea*. *BMC genomics*, 19(1), 378.

Manuscrit en préparation concernant le sujet de thèse

- PiRATE v2: detection, classification and annotation of transposable elements of non-model organisms.
- Transposable elements dynamics in domesticated strains of the microalga *Tisochrysis lutea*.

Autre publication

- Berthelie, J., Schnitzler, C. E., Wood-Charlson, E. M., Poole, A. Z., Weis, V. M., & Detournay, O. (2017). Implication of the host TGF β pathway in the onset of symbiosis between larvae of the coral *Fungia scutaria* and the dinoflagellate *Symbiodinium* sp. (clade C1f). *Coral Reefs*, 36(4), 1263-1268.

Communications lors de congrès nationales ou internationales

1. Gen2bio 2016 « Annotation of TEs in the draft genome of the marine microalga *Tisochrysis lutea* », 31 mars 2016, Saint-Brieuc, France. Poster.
2. International Congress on Transposable Elements (ICTE) « Annotation of TEs in the draft genome of the marine microalga *Tisochrysis lutea* », 16-19 avril 2016, Saint-Malo, France. Poster.
3. Réunion du groupement de recherche éléments mobiles (GDR 3546), « De novo annotation of TEs from the genome of the microalga *Tisochrysis lutea* », 8-9 décembre 2016, Paris, France. Présentation orale.
4. Colloque annuel de la Société Phycologique de France, « Annotation of Transposable Elements in the genome of *Tisochrysis lutea* », 28 mars 2017, Paris, France. Présentation orale.
5. Colloque du réseau EFOR, « PiRATE: a Pipeline to Retrieve and Annotate Transposable Elements in *Tisochrysis lutea* genome », 3 mai 2017, Paris, France. Présentation orale.

6. International Society for Applied Phycology, « Transposable elements as markers of the genetic stability of domesticated microalgae after cryopreservation process », 18-23 juin 2017, Nantes, France. Poster.
7. International Congress of Phycology, « PiRATE: a Pipeline to Retrieve and Annotate Transposable Elements in *Tisochrysis lutea* genome », 13-19 août 2017, Szczecin, Pologne. Présentation orale.
8. EMBL Symposium The Mobile Genome: Genetic and Physiological Impacts of Transposable Elements, « PiRATE: a Pipeline to Retrieve and Annotate TEs in microalgal genomes », 11-14 octobre 2017, Heidelberg, Allemagne. Poster.
9. Réunion du groupement de recherche éléments mobiles (GDR 3546), «PiRATE: a Pipeline to Retrieve and Annotate TEs in microalgal genomes», 4-5 décembre 2017, Paris, France. Poster.
10. Plant and Animal Genome XXVI, «PiRATE: a Pipeline to Retrieve and Annotate TEs in non-model organisms», 13-17 janvier 2018, San Diego, USA. Présentation orale et poster.
11. Atlantic Microalgae Consortium Conference, «Etude de l'activité des éléments transposables chez *Tisochrysis lutea* dans un context d'amélioration des espèces de microalgues», 1^{er} juin 2018, Nantes, France. Présentation orale et poster.
12. JOBIM, «PiRATE v2: detection, classification and annotation of transposable elements of non-model organisms», 3-6 juillet 2018, Marseille, France. Poster.
13. Plant epi/genetics Symposium, «PiRATE v2: detection, classification and annotation of transposable elements of non-model organisms», 29-31 octobre 2018, Angers, France. Présentation orale.

Autres communications

1. Journée scientifique de l'école doctoral VENAM, « Annotation of TEs in the draft genome of the marine microalga *Tisochrysis lutea* », 3-4 novembre 2016, Nantes, France. Premier prix poster.
2. Journée des doctorants du centre Atlantique de l'IFREMER, «PiRATE: a Pipeline to Retrieve and Annotate Transposable Elements in microalgal genomes», 17 octobre 2017, Nantes, France. Présentation orale et poster.
3. Journées de l'ED SML, « Annotation des éléments transposables dans le génome de la microalga *Tisochrysis lutea* », 13-14 février 2018, Brest, France. Présentation orale et poster.

Publication en ligne d'outils bio-informatiques

- PiRATE : <http://doi.org/10.17882/51795>
- pirateClassif : <https://github.com/francetr/pirateClassif>

Remerciements

Je remercie François Sabot et Hervé Moreau pour avoir accepté d'être rapporteurs à cette thèse.

De même, je remercie Emmanuelle Lerat, Florian Maumus et Jérémie Bourdon d'avoir accepté de l'examiner.

Je remercie également Jonathan Filée et Etienne Bucher, pour leurs conseils lors des comités de suivi de thèse.

Merci à Nicolas Daccord et Véronique Jamilloux pour leur aide dans ce travail de thèse.

Merci à la Région Pays de la Loire (Programme Atlantic Microalgae) et Ifremer pour avoir financé ce travail de thèse.

Je tiens, bien sûr, à remercier mes encadrants Nathalie Casse, Grégory Carrier et Bruno Saint-Jean.

MERCI à mon maître Jedi, Grégory, qui m'aura permis de m'initier à la bio-informatique durant cette thèse. Comme disait ton propre encadrant « Toujours par deux, ils vont. Ni plus, ni moins. Le maître et son apprenti. » Je pense, Grégory, que tu dois figurer parmi les meilleurs encadrants de la « Galaxy » !

MERCI à ma directrice de thèse, Nathalie, l'amoureuse des animaux. Même si nous n'étions pas dans le même laboratoire, j'ai toujours eu ton aide et un fort soutien. Je t'en remercie.

MERCI à Bruno, le légendaire Targaryen ~~à la timide chevelure argentée~~. Merci pour ton aide, ton soutien et ta rigueur lors de cette thèse. Je saurai me souvenir de ton humour et me montrer à la hauteur de ton enseignement !

Votre pédagogie et votre disponibilité à tous m'auront permis de réaliser cette thèse sereinement et avec beaucoup de plaisir.

Si cette thèse s'est très bien déroulée, c'est également grâce aux autres personnes formidables qui sont embarquées au laboratoire PBA !

Merci au capitaine Gaël, qui retient tant bien que mal la barque PBA, pour ne pas qu'elle parte trop à la dérive. Merci à Isabelle, l'amiral de PBA, qui heureusement adore couvrir ses thésards. Merci à Elodie qui m'a accueilli dans son bureau lors de mon arrivé. Merci à Ewa, pour m'avoir fait visiter un petit bout de Pologne à Szczecin lors de l'IPC 11 (et Bigos !). Merci à Matthieu « le jazzman pêcheur » et Jean-Baptiste « le métaleux tout en finesse ». Merci à Aurélie et Nathalie pour leur aide dans les manip ! Merci à Camila pour ses conseils. Merci également à Liet, pour nos discussions sur le pays du soleil levant. Merci aux anciens stagiaires Julien, Simon, Anne-Cécile et Marie. Enfin, merci aux jeunes recrues, Virginie, Francis et Thomas, pour continuer à faire vivre la bonne ambiance au labo.

Merci également à ceux qui ont quitté la barque PBA : Hansy, Nicolas, Caroline et Elodie, partis pour d'autres aventures professionnelles. Loïc, déposé en Nouvelle-Calédonie. Catherine et Raymond, nos joyeux retraités aux grands cœurs !

Merci bien évidemment à Stan et Judith, le duo de thésards de la génération antérieure, pour votre accueil, votre aide et vos conseils lors de cette thèse.

Merci à ma « compère » Charlotte, qui forma avec moi la 3^{ème} génération de binômes de thésards *aka* : les « poussins ». Nous avons certainement formé la paire la plus « geek » que n'ait jamais vu le laboratoire PBA ! Merci à toi et à Erwann pour votre amitié, et votre aide !

Je remercie bien sûr tous mes autres compatriotes thésards : Norman, Valentin, Arthur, Francesco, Nour, Katy, Marin, Floriane, Widya et Audric.

Merci également Myriam et Tristan pour votre implication et votre intérêt dans la réalisation de la seconde version de PiRATE.

Merci également à la première personne m'ayant transmis le virus de la recherche : Olivier Detournay. Merci à toi de m'avoir donné l'opportunité, lors de mon cursus universitaire, de conduire une très belle étude sur la symbiose entre les coraux et les *Symbiodinium*.

Pour finir, je remercie l'ensemble de ma famille et amis.

Je remercie en particulier mes parents, Pierre et Catherine, mon frère Yann, et ma chérie Lucie, pour leur soutien et leur amour.

Merci à vous d'avoir éveillé en moi la passion pour l'océan et ses habitants lors de notre précédente vie dans l'océan Indien.

Enfin, merci de m'avoir soutenu dans mes études et de m'avoir fait confiance.

Préambule

Comme en témoignent les illustrations laissées par les anciennes civilisations, les océans et leurs habitants ont depuis toujours été une source de fascination pour l'homme (Figure 1). Durant l'antiquité, l'ouvrage « *Historia Animalium* » écrit par Aristote ouvre la voie vers le recensement et la description des organismes marins, principalement des poissons et mollusques (Aristote, 343 AEC). Bien plus tard, au XVI^{ème} siècle, Pierre Belon et Guillaume Rondelet recenseront en Europe les poissons et invertébrés observés lors de voyages dans de nombreuses citées méditerranéennes (Figure 1c) (Belon, 1551; Rondelet, 1554, 1555). Ce n'est cependant que dans les années 1670, qu'Antoni van Leeuwenhoek commence à attirer l'attention sur des habitants marins encore insoupçonnés : les microorganismes. Dans ses lettres destinées à la Royal Society, Leeuwenhoek rapporte l'observation de « petits animalcules » présentant « diverses couleurs et formes » dans un échantillon d'eau prélevé dans un lac aux Pays-Bas (Corliss, 1975). Il fut ainsi probablement le premier à observer et décrire des organismes microscopiques et des espèces phytoplanctoniques (Figure 1d) (Finlay, 2001).

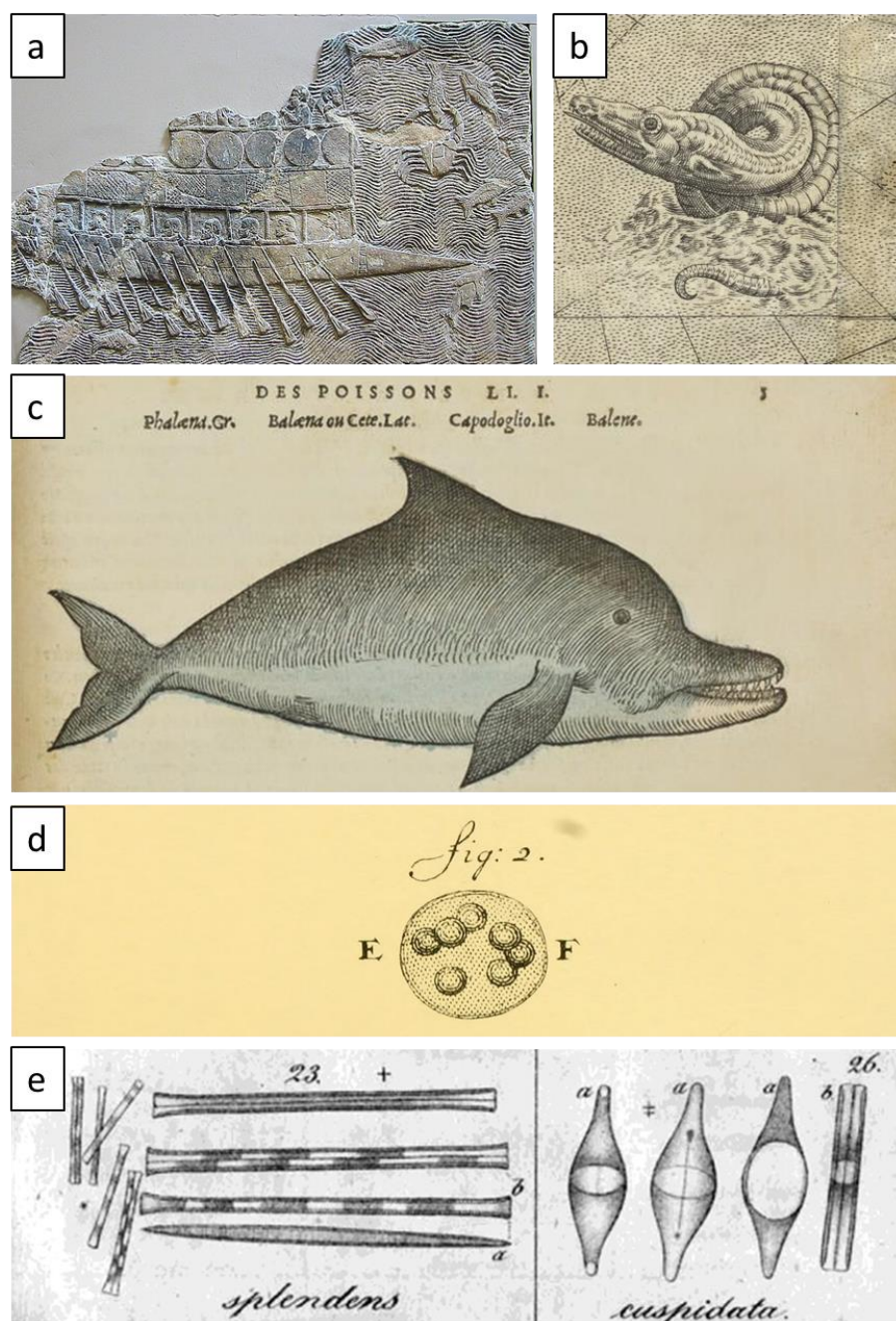
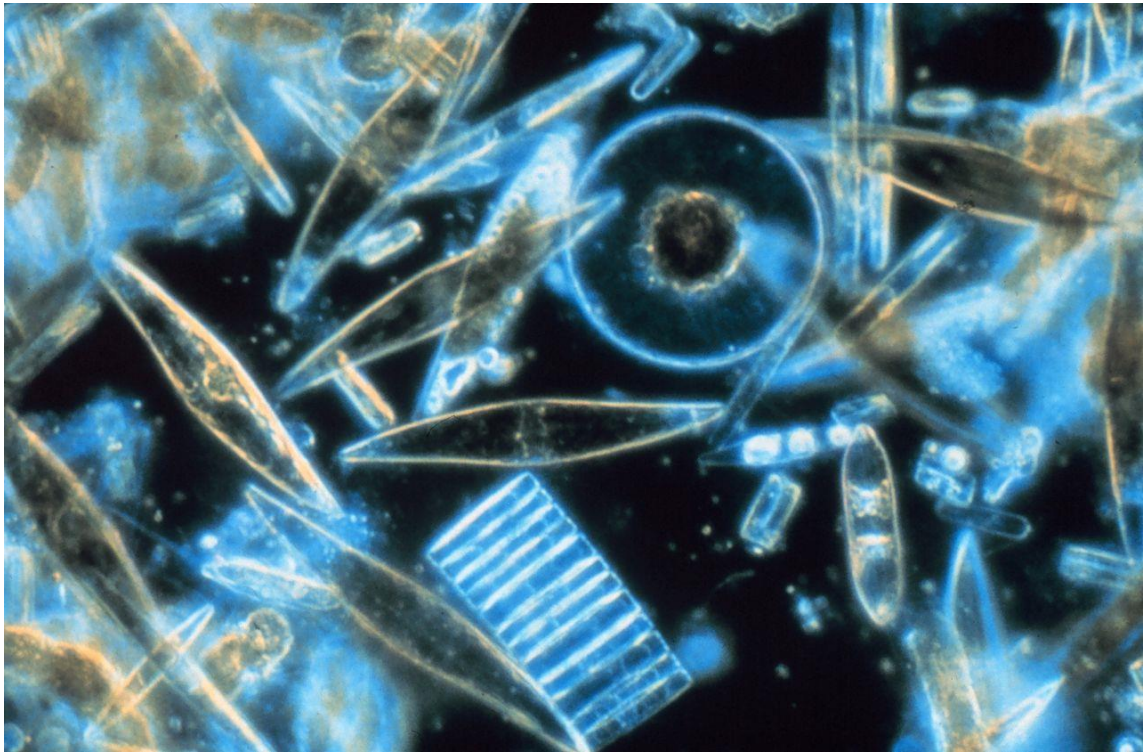


Figure 1 : Représentation d’organismes aquatiques de l’antiquité au 19^{ème} siècle.

a) Relief représentant des poissons et crabes, œuvre datée de 700 AEC provenant d’Assyrie et exposée au British Museum. b) Illustration d’un monstre marin sur la carte du monde de Mercator en 1569. c) Illustration d’un dauphin par Pierre Belon au 16^{ème} siècle (Belon, 1551). d) Croquis de la microalgue *Volvox* sp. par Antoni van Leeuwenhoek, qu’il considéra à l’époque comme un « animalcule », dessin provenant d’une lettre datée du 2 janvier, 1700. e) Croquis de microalgues Diatomées par Kützing (1833).

Introduction générale

1. Les microalgues



Photographie réalisée par le Pr. Gordon T. Taylor, Stony Brook University.

1.1. Rôles écologiques du phytoplancton

Le phytoplancton constitue l'ensemble des microorganismes photosynthétiques vivant dans un milieu aquatique. Ces organismes sont subdivisés en deux groupes selon leur structuration cellulaire interne : celui des Cyanobactéries, composé d'organismes procaryotes et celui des microalgues, formé par des organismes eucaryotes. La photosynthèse permet à ces organismes microscopiques de transformer le CO₂ atmosphérique solubilisé dans l'eau en carbone organique et de produire de l'O₂. Ce processus est réalisé par la conversion de l'énergie solaire en énergie chimique, grâce à l'intervention de pigments chlorophylliens contenus dans les chloroplastes. Dans les milieux aquatiques, la photosynthèse est réalisée par le phytoplancton dans la zone où la lumière solaire traverse la masse d'eau. Celle-ci est nommée la « zone euphotique ».

A l'échelle globale de notre planète, le phytoplancton est présent sur les trois quarts de la surface de la Terre. Pourtant, il représente moins de 1% des 600 milliards de tonnes de carbone que représente la biomasse photosynthétique globale (Falkowski, 2002). Cependant, la contribution de ces organismes à la photosynthèse est estimée à 40-45% à l'échelle de la planète et à 80-90% au niveau des océans (Duarte et Cebrian, 1996; Field et al., 1998; Geider et al., 2001). Par son rôle de producteur primaire, le phytoplancton constitue le premier compartiment des réseaux trophiques des océans. Il représente la source principale de carbone organique nécessaire à la vie des organismes hétérotrophes marins, tels que des petits crustacés, bivalves ou poissons. Par conséquent, le phytoplancton possède un rôle clef dans le cycle du carbone des océans, en participant à la « pompe biologique du carbone » (Geider et al., 2001) (Figure 2).

En effet, le carbone organique généré par le phytoplancton à la surface est par la suite exporté vers les fonds marins. Par exemple sous forme de cellules mortes ou de pelotes fécales (Figure 2) (Falkowski, 2002). Le carbone organique qui atteint le fond des océans peut ensuite être reminéralisé par des bactéries et être séquestré durant plusieurs milliers d'années (Figure 2) (Falkowski, 2002). Cette source de carbone participe à la formation de puits de pétrole au cours des temps géologiques. Les organismes phytoplanctoniques permettent ainsi l'export et la séquestration du carbone atmosphérique dissous de la couche euphotique vers les fonds océaniques (Figure 2) (Volk et Hoffert, 1985). Par cette contribution à limiter l'accumulation de CO₂ dans l'atmosphère à l'échelle de la Terre, le

phytoplancton limite la proportion de ce gaz à effet de serre et a ainsi une influence majeure sur le climat (Falkowski, 2002). Les organismes phytoplancton participent également à d'autres cycles biogéochimiques qui sont indispensables à leur survie, tels que l'azote et le phosphore (Corner et Davies, 1971).

Parmi les communautés phytoplanctoniques, le groupe des Cyanobactéries est composé d'organismes divers. Le nombre d'espèces de ce groupe est estimé à environ 8 000 espèces, bien qu'environ 3 000 seulement aient été décrites (Guiry, 2012; Nabout et al., 2013). Celui-ci est majoritairement représenté par les genres, *Prochlorococcus* et *Synechococcus* (Waterbury, 1986; Partensky et al., 1999; 2012). Le nombre des espèces de microalgues est quant à lui estimé entre 70 000 et 150 000, bien qu'il pourrait être proche de 1 million (Guiry, 2012; de Vargas et al., 2015).

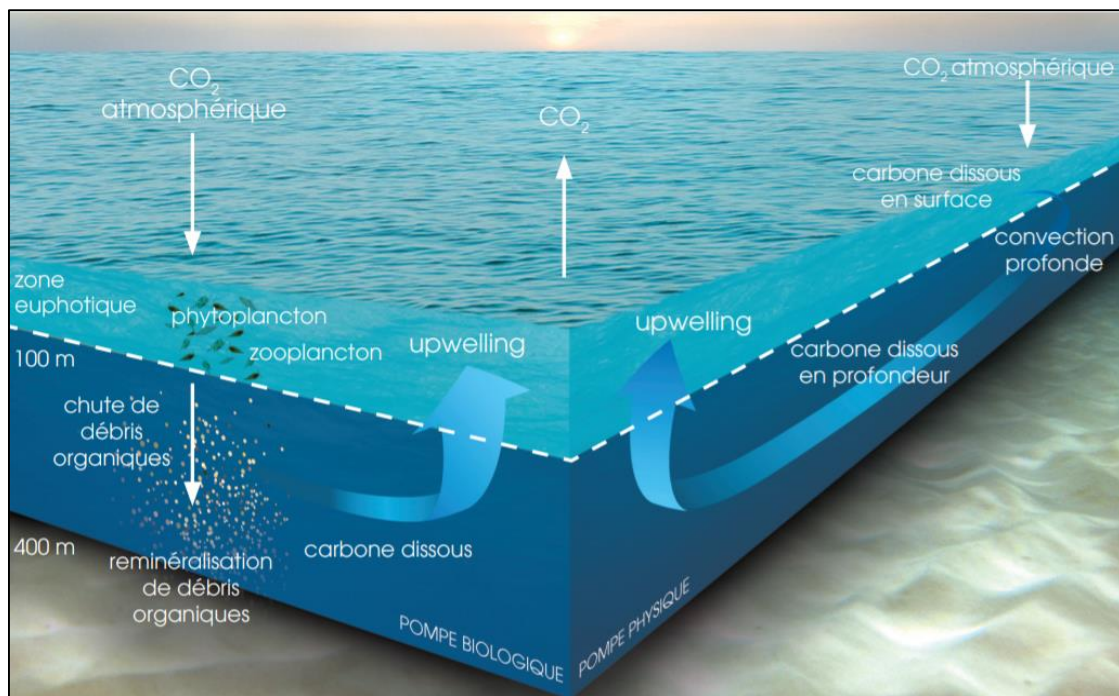


Figure 2 : Illustration de la pompe biologique et de la pompe physique du carbone.

Pompe biologique : le carbone atmosphérique est transformé en carbone organique par la photosynthèse du phytoplancton. Le carbone organique fixé par les organismes phytoplanctoniques est ensuite consommé par les prédateurs hétérotrophes. Une fraction du carbone organique est transférée vers le fond, où elle est reminéralisée par des bactéries. Au cours des échelles de temps géologiques, ce carbone séquestré dans le fond conduit à la formation de pétrole. Pompe physique : les courants qui plongent vers le fond exportent le CO_2 présent en surface vers les profondeurs et ne sera plus en contact avec l'atmosphère. Des phénomènes de remontées des eaux (*upwelling*), peuvent transporter le CO_2 dissous vers la surface. Image provenant de Bopp et al. (2002).

1.2. La diversité des microalgues

Les microalgues forment un groupe d'organismes polyphylétiques (Figure 3). Le terme « microalgue » n'a ainsi pas de sens taxonomique. Il permet par commodité de regrouper les microorganismes photosynthétiques eucaryotes, unicellulaires ou multicellulaires, vivant dans un milieu aquatique. Il existe en effet des espèces unicellulaires telle que *Chlamydomonas reinhardtii* et multicellulaires telle que *Volvox carteri*. Les microalgues sont des organismes aquatiques cosmopolites, recensés dans la majorité des milieux aquatiques du globe terrestre. Elles sont présentes dans des eaux douces, saumâtres ou salées, tels que les fleuves, les marécages et les océans. Ces organismes ont également été identifiés dans des habitats extrêmes tels que des croûtes biologiques de sable dans les déserts, des plaques de glace, des cours d'eau à très haute altitude, des sources d'eaux chaudes et dans l'atmosphère (Sakai et al., 1995; Rott et al., 2006; Pawlowski et al., 2011; Anesio et Laybourn-Parry, 2012; Treves et al., 2013; Tesson et al., 2016). Au cours de leur vie, ces organismes vivent de façon solitaire ou en colonie, mais peuvent également se développer à l'intérieur des tissus d'un autre organisme. Par exemple, certaines microalgues parasitent des poissons ou des bivalves (Gray et al., 1999). D'autres espèces vivent en symbiose, par exemple avec des coraux (Bertheliet et al., 2017). Cette symbiose permet par exemple un échange de nutriments entre ces organismes et leur permet de vivre dans des régions océaniques oligotrophes.

Il existe donc de très nombreuses espèces de microalgues. Elles ont pu s'adapter à la plupart des milieux aquatiques et présentent une forte diversité génétique et morphologique. Ainsi, au sein d'une même espèce, différentes souches provenant d'écotypes très différents peuvent coexister. Bien que peu documenté dans la littérature, il existe une grande diversité génétique au sein d'une même espèce (Flowers et al., 2015; Gallaher et al., 2015; Blanc-Mathieu et al., 2017; Koester et al., 2018). De plus, cette diversité génétique peut également être observable pour des souches d'origine non-clonale suggérant ainsi qu'une souche correspond davantage à une population d'individus génétiquement différents (Carrier et al., 2018). L'origine de cette diversité génétique est notamment liée au cycle de vie des microalgues qui est en général court (en moyenne 1 division par jour) et du fait de la dérive génétique. La reproduction sexuée permet également d'amplifier ce phénomène. Les mécanismes de reproduction et les cycles de vie sont cependant encore très mal connus pour la plupart des espèces de microalgues.

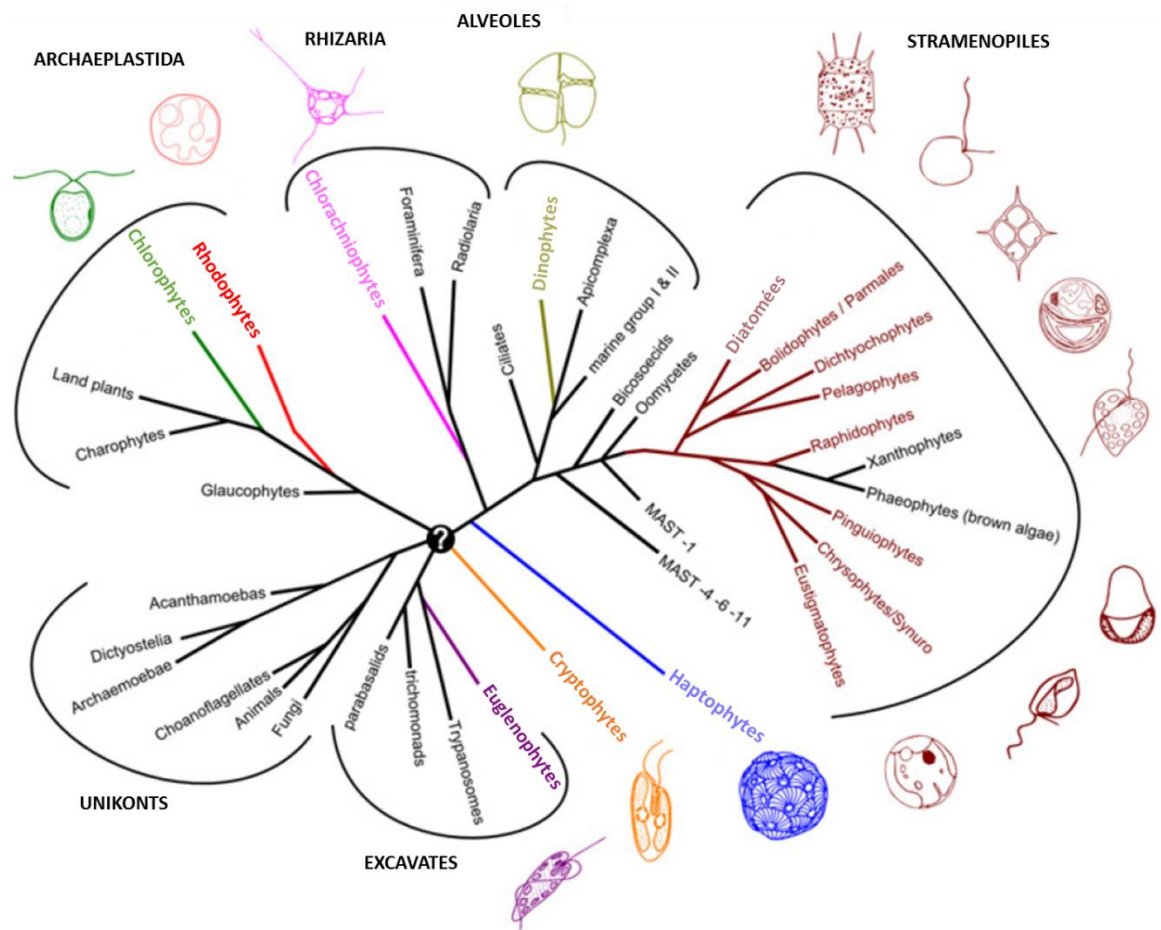


Figure 3 : Arbre phylogénétique représentant la distribution des taxons de microalgues.
Image adaptée de Not et al. (2012).

1.3. Origine des microalgues

L'origine commune des microalgues s'appuie sur la théorie de l'endosymbiose (Mereschkowsky, 1910; Margulis, 1975). Celle-ci est basée sur l'étude des chloroplastes et explique l'origine commune de l'ensemble des eucaryotes photosynthétiques. Il y a 1,8 milliards d'années, une Cyanobactérie ancestrale aurait été phagocytée par une cellule eucaryote hétérotrophe (Figure 4). La cellule eucaryote aurait ensuite pu intégrer la Cyanobactérie ancestrale dans sa machinerie cellulaire. Celle-ci donnant naissance au cours du temps au premier chloroplaste eucaryote (Gould et al., 2008; Poole et Neumann, 2011). Cet évènement « d'endosymbiose primaire » est supporté par des études génétiques démontrant des similitudes entre les chloroplastes des eucaryotes et des Cyanobactéries (Not et al., 2012). Au cours de l'évolution, cette endosymbiose primaire aurait permis l'émergence des Chlorophytes (lignée verte), Rhodophytes (lignée rouge) et des Glaucophytes (Figure 4). De nouveaux évènements d'endosymbioses, nommés « endosymbioses secondaires » auraient ensuite eu lieu (Keeling, 2010). Des évènements indépendants de phagocytoses de Chlorophytes ancestraux par des eucaryotes hétérotrophes seraient à l'origine de l'émergence des Chlorarachniophytes et des Euglénophytes (Figure 4). De même, des évènements indépendants de phagocytoses de Rhodophytes ancestraux par des eucaryotes hétérotrophes auraient conduit à l'apparition des Stramenopiles, Dinophytes, Cryptophytes et des Haptophytes (Lang-Unnasch et al., 1998) (Figure 4). D'autres séries d'endosymbioses, nommées « endosymbioses tertiaires », auraient eu lieu par la suite. L'ensemble de ces évènements d'endosymbiose serait à l'origine de la diversité taxonomique actuelle des microalgues chez les eucaryotes (Figure 3).

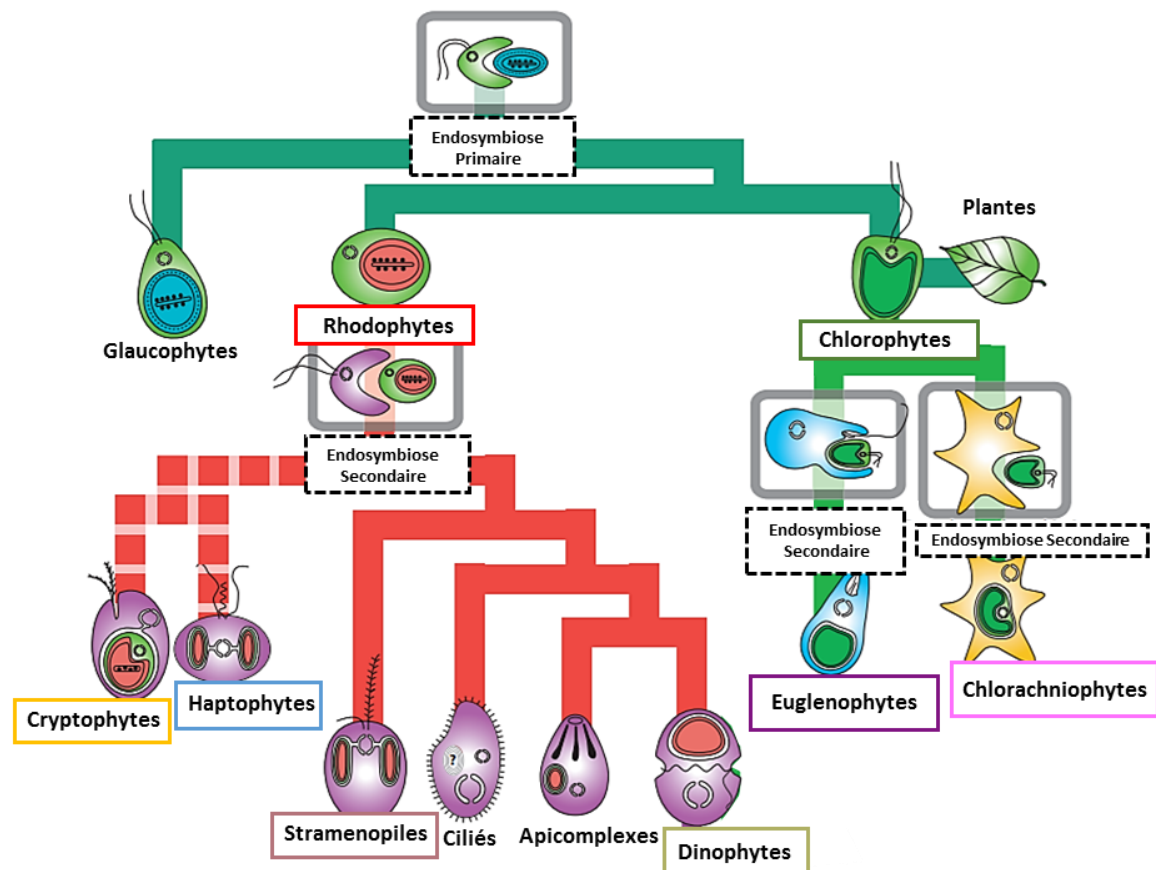


Figure 4 : Représentation schématique de l'origine endosymbiotique des microalgues.

Image simplifiée et adaptée de Keeling (2013).

1.4. Les taxons de microalgues les plus étudiés

Parmi les microalgues, les Chlorophytes constituent le taxon le plus examiné. Celles-ci sont fortement étudiées dans la compréhension de l'évolution des eucaryotes photosynthétiques. Les microalgues d'eau douce, *Chlamydomonas reinhardtii*, *Chlorella* sp. et *Volvox carteri* (Figure 5) sont les principales espèces modèles de ce taxon. Les microalgues marines des genres *Micromonas*, *Bathycoccus* et *Ostreococcus* sont également très étudiés. Ces organismes sont fortement abondants dans les zones euphotiques des océans (Worden et al., 2009). Chez les Straménopiles, les Diatomées sont considérées comme les microalgues dominantes au niveau des côtes des océans. Elles sont ainsi fortement étudiées pour leur participation aux grands cycles biogéochimiques et notamment celui du carbone. Leur contribution annuelle dans la production primaire globale est estimée à 20% (Tréguer et Pondaven, 2000). Certaines espèces effectuent des efflorescences massives et sont examinées pour leurs intérêts écologiques, géologiques et paléontologiques. Les espèces *Phaeodactylum tricornutum* (Figure 5), *Thalassiosira pseudonana* et *Nannochloropsis gaditana* sont parmi les plus analysées. Chez les Alvéolées, les Dinophytes sont également considérés comme abondants dans les océans (Field et al., 1998). Ces organismes sont analysés pour leur capacité à réaliser des efflorescences pouvant être toxiques tel que les genres *Alexandrium* ou *Karenia*. Certaines espèces sont aussi examinées pour leur aptitude à vivre en symbiose avec des coraux, telles que les espèces du genre *Symbiodinium* (Figure 5), permettant ainsi la formation des récifs coralliens. Enfin, les Haptophytes sont quant à eux de plus en plus considérés comme un groupe majeur de l'écologie des milieux océaniques (Liu et al., 2009a).

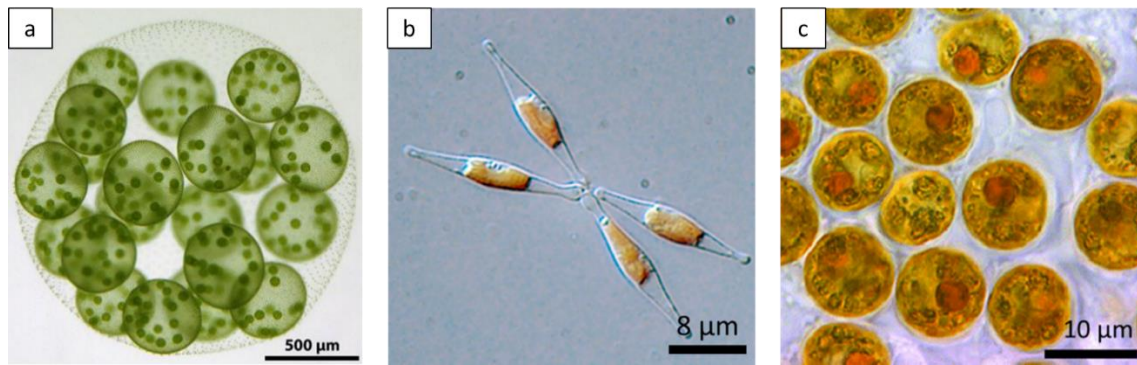


Figure 5 : Aperçu de la diversité phénotypique des microalgues.

Photographie par microscopie de a) *Volvox carteri* (Chlorophyte), b) *Phaeodactylum tricornutum* (Diatomée) et c) *Symbiodinium* sp. (Dinophyte). Photographies prises par D. Kirk, Washington University ; A. de Martino et C. Bowler, Stazione Zoologica et Ecole Normale Supérieure ; T. C. LaJeunesse, Pennsylvania State University.

1.5. Le phylum des Haptophytes

Les Haptophytes sont aujourd'hui considérés parmi les microalgues les plus dominantes du phytoplancton marin, que ce soit au niveau des écosystèmes côtiers ou des océans ouverts (Liu et al., 2009a; Eikrem et al., 2016). Ces organismes présentent une taille généralement comprise entre 2 et 30 μm (Eikrem et al., 2016), bien que de nombreuses espèces possèdent une taille inférieure située entre 0.2 et 2 μm (Jardillier et al., 2010; Kirkham et al., 2011). En raison d'un grand nombre d'espèces de petite taille, la diversité des Haptophytes a tout d'abord été fortement sous-estimée lors de leur identification par microscopie (Edvardsen et al., 2016). Ce sont finalement les méthodes moléculaires basées sur l'analyse des séquences d'ARNr 18S, qui ont permis de prendre conscience de leur forte diversité (Moon-van der Staay et al., 2001). La classification phylogénétique de ces organismes repose aujourd'hui sur cette méthode d'analyse. Les Haptophytes sont ainsi représentés par deux groupes taxonomiques majoritaires. Celui des Pavlovophycées, formé par quatre genres et comprenant 13 espèces (Bendif et al., 2011, 2013) et celui des Prymnesiophycées constitué de 318 espèces regroupées dans 76 genres (Jordan et al., 2004).

La plupart des Haptophytes vivent de façon solitaire, même si certaines forment des colonies (Green et Parke, 1975; Andersen et al., 2014) ou vivent en symbiose avec des radiolaires ou des foraminifères (Gast et al., 2000; Decelle et al., 2012). Leurs cellules possèdent généralement deux flagelles qui leur permettent un déplacement de type elliptique par la réalisation de pulsations (Eikrem et al., 2016). Un appendice filiforme nommé haptonème, situé entre ces deux flagelles (Figure 6) est impliqué dans la détection d'obstacles et la capture de proies (Kawachi et al., 1991). Bien que la plupart des Haptophytes soient photo-autotrophes, certaines espèces sont mixotrophes (Tillmann 1999) et dans de rares cas hétérotrophes stricts (Marchant et Thomsen, 1994). Des auteurs estiment que leur capacité de mixotrophie, leur fort taux de croissance et leur taille relativement grande expliqueraient leur forte abondance dans l'océan par rapport aux autres populations de microalgues (Cuvelier et al., 2010). En effet, les Haptophytes représenteraient entre 30 et 50% de la biomasse photosynthétique de l'océan mondial. Cette proportion serait ainsi près de deux fois supérieure à celle des Diatomées et des Cyanobactéries (Figure 7) (Liu et al., 2009a).

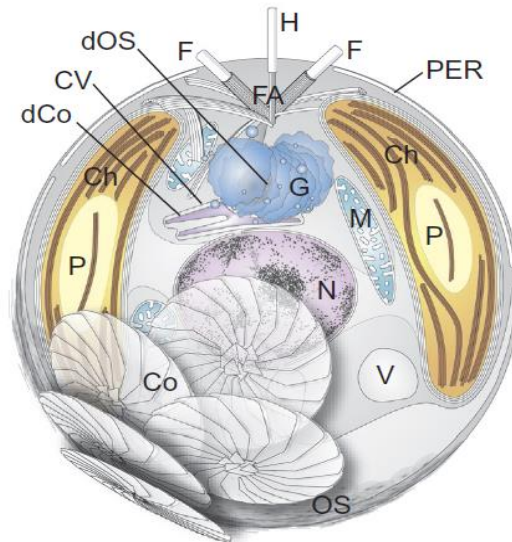


Figure 6 : Représentation schématique de la composition d'une cellule Haptophyte.

Ch : chloroplaste, Co : coccolithes, CV : vésicule formant les coccolithes, dCo : développement d'hétérococcolithe, dOS : écailles organiques en développement, F : flagelle, FA : appareil flagellaire, G : appareil de Golgi, H : haptonème, M : mitochondrie, N : noyau, OS : écailles organiques, P : pyrénoïde, PER : réticulum endoplasmique périphérique, V : vacuole. Image provenant de Tsuji et Yoshida (2017).

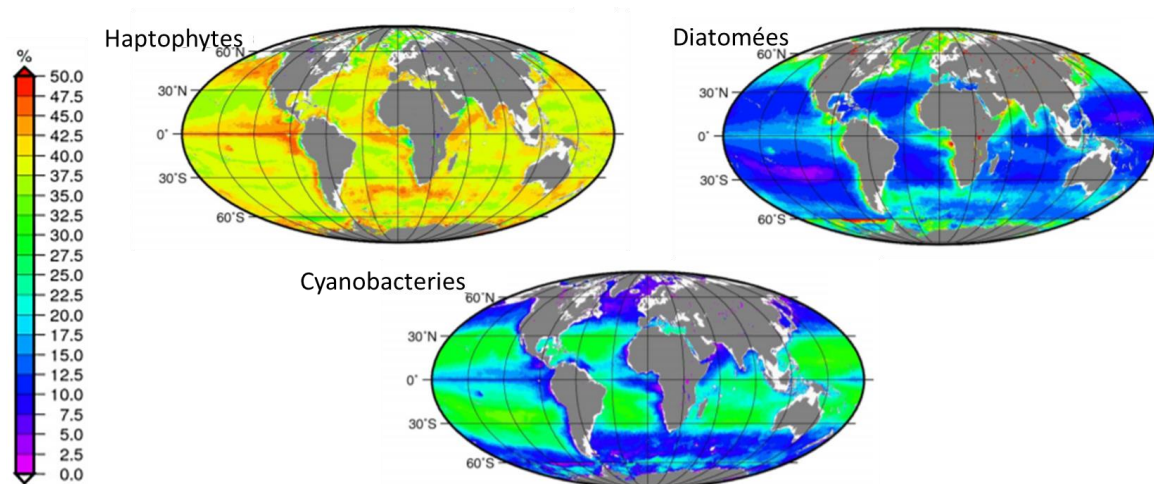


Figure 7 : Contribution des Haptophytes, Diatomées, et Cyanobactéries à la biomasse photosynthétique de l'année 2000.

Image provenant de Liu et al. (2009a).

De couleur jaune-marron, la cellule des Haptophytes est généralement couverte par une ou plusieurs couches d'écailles organiques (Figures 6 et 8). Certaines espèces nommées coccolithophores génèrent également des écailles calcifiées nommées «coccolithes». Par la formation de ces écailles, les coccolithophores possèdent un rôle important dans le cycle biogéochimique du cycle du carbone (Iglesias-Rodríguez et al., 2002). La formation de leurs coccolithes varie selon le cycle de vie de l'espèce (Figure 9) (Brownlee et al., 2015). De manière générale, le cycle de vie de la plupart des Haptophytes est haplodiplophasique (Figure 9). Néanmoins, le cycle de vie de certaines espèces comme celles constituant l'Ordre des Isochrysidales (non coccolithophores) est encore mal connu (Figure 9).

Les Haptophytes coccolithophores sont fortement étudiées pour leur intérêt en géologie et paléoclimatologie. En effet, les Haptophytes représentent une part importante de la craie et leur présence est connue depuis le Trias (-252 à -201 millions d'années). Elle est fréquente aussi au Crétacé (-145 à -65 millions d'années). Au cours du temps, l'accumulation des coccolithes de ces organismes forme les importantes couches de sédiments qui composent les falaises de Douvres ou d'Étretat en France. De plus, les Haptophytes appartenant à l'Ordre des Isochrysidales produisent des lipides neutres uniques en leur genre, nommés alcénones, qui sont conservés dans les sédiments anciens (Marlowe et al., 1984). La production de ces lipides est connue pour être influencée par la température de l'eau. En effet, des températures élevées favorisent la production d'alcénones di-insaturés par rapport aux alcénones tri-insaturés (Conte et al., 2006). Ainsi, les alcénones sont utilisés comme biomarqueurs en paléoclimatologie (Rechka et Maxwell, 1988).

La coccolithophore *Emiliania huxleyi* constitue l'organisme modèle des Haptophytes. Cette microalgue est notamment étudiée en raison des efflorescences étendues et massives de bloom qu'elle est capable de réaliser (Figure 10). Les premières données transcriptomiques ont été obtenues pour cette espèce (von Dassow et al., 2009) et son génome a ensuite été séquencé (Read et al., 2013). Depuis, de nouvelles données génomiques et transcriptomiques commencent à être accumulées pour d'autres espèces telles que *Tisochrysis lutea* et *Chrysochromulina tobin* (Annexe 1) (Carrier et al., 2014; Hovde et al., 2015; Carrier et al., 2018).

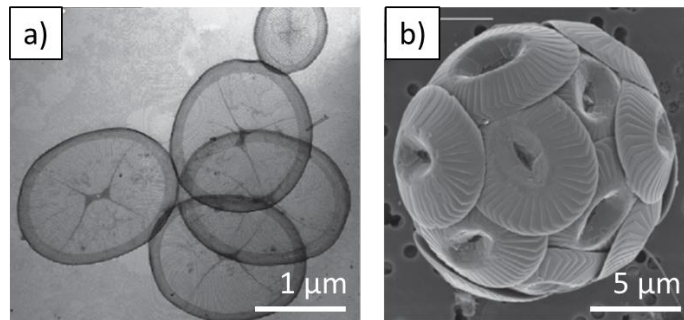


Figure 8 : Visualisation par microscopie de cellules d'Haptophytes.

a) Ecailles organiques de *Chrysochromulina* sp. b) Ecailles calcifiées de *Coccolithus pelagicus*. Images provenant de Tsuji et Yoshida (2017).

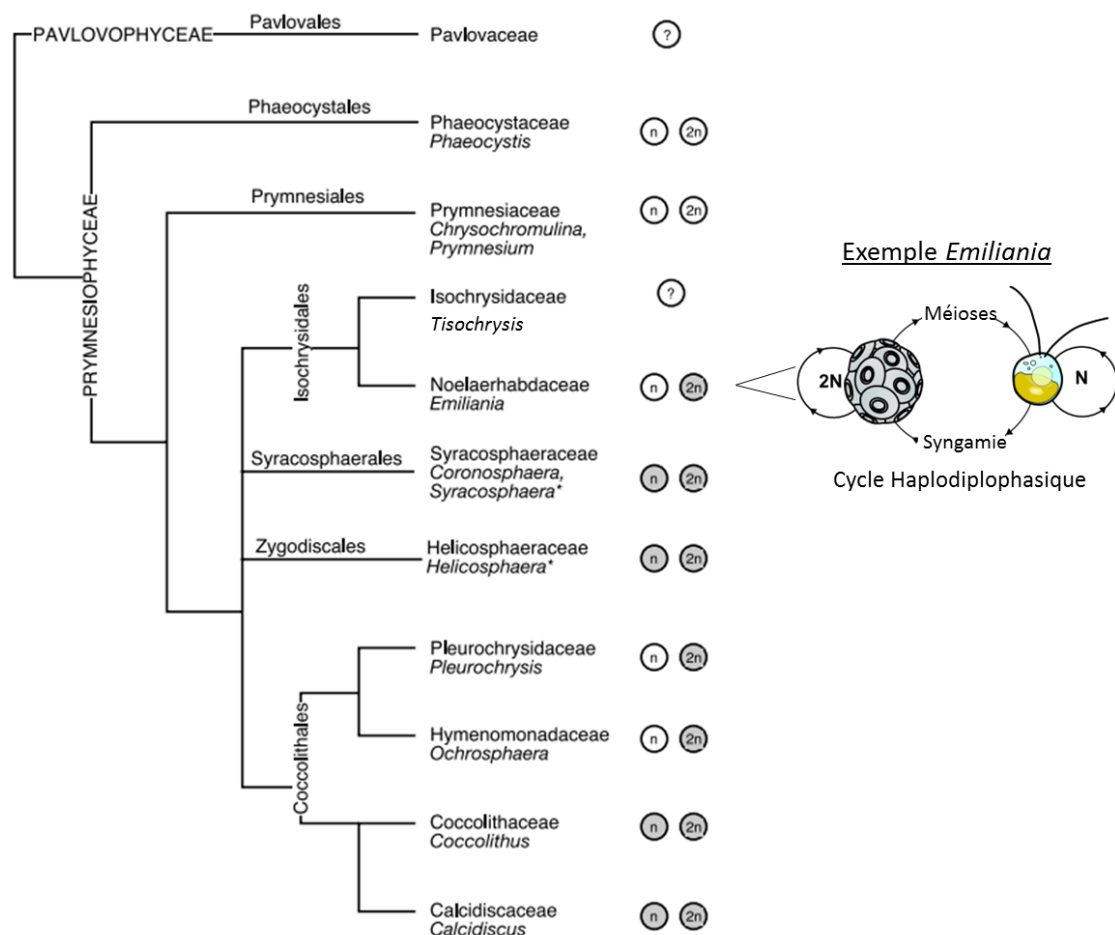


Figure 9 : Phases de vie des différentes familles d'Haptophytes.

Le cycle de vie des Isochrysidaceae est encore inconnu. Les cercles gris indiquent les phases produisant des coccolites. L'exemple du cycle de vie d'*Emiliana* sp. est proposé avec une alternance d'une phase sexuée et d'une phase asexuée. Arbre phylogénique de Houdan et al. (2004) et illustration de M. Bendif, University of Oxford.



Figure 10 : Photographie satellite d'une efflorescence de la microalgue Haptophyte *Emiliania huxleyi* à proximité des côtes de la Grande-Bretagne.

Photographie de S. Groom, Plymouth Marine Laboratory.

2. La génomique



Figure 11 : Représentation des séquences de l'ADN en bio-informatique.

Image provenant du National Human Genome Research Institute (NHGRI).

2.1. Historique de la génomique

Le séquençage de l'acide désoxyribonucléique (ADN) est une méthode permettant de déterminer l'enchaînement de nucléotides constituant un fragment ou l'intégralité d'une molécule d'ADN. En génétique et bio-informatique, les bases azotées nucléiques sont représentées par les lettres A, G, T et C (Figure 11). Celles-ci représentent les bases puriques : adénine (A) et guanine (G) et les bases pyrimidiques : thymine (T) et cytosine (C).

Les techniques de séquençage de l'ADN ont été initiées parallèlement par les équipes de Sanger et Maxam (Sanger et Coulson, 1975; Maxam et Gilbert, 1977). Frederick Sanger et ses collègues furent les premiers à séquencer le génome complet d'un organisme, le virus ϕ X174, d'une taille de 5 375 paires de bases (pb) (Sanger et al., 1977). Des appareils utilisant en partie la méthode de séquençage Sanger ont ensuite été développés par ABI (Applied Biosystems) en 1987 (Shendure et al., 2017). Celle-ci deviendra par la suite la technologie de référence de séquençage de l'ADN. Elle possède un taux d'erreur inférieur à 0.01% mais reste coûteuse, pour un faible débit de séquençage (environ 1 heure pour séquencer 1 000 pb). Elle est donc onéreuse pour séquencer de grands génomes complexes d'organismes eucaryotes composés de millions, voire de milliards de pb. Pour exemple, le projet de séquençage du génome humain, utilisant en partie cette technologie, a duré 10 ans (1991-2001) et a coûté autour de 2,7 milliards de dollars US (<https://www.genome.gov/11006943/>).

Dans les années 2005, les techniques de séquençages d'ADN de seconde génération ont été développées afin de permettre un débit plus élevé et moins onéreux : les technologies 454 (Roche Diagnostic) et Illumina (Tableau 1 et Figure 12). En raison de son coût peu élevé et de son faible taux d'erreur, la technologie Illumina est actuellement la plus utilisée pour séquencer un génome entier. Pour exemple, le taux d'erreur de l'Illumina HiSeq 4000 est aujourd'hui inférieur à 1% et elle figure parmi les technologies les moins coûteuses du marché (Tableau 1). Cette technologie réalise l'amplification et le séquençage de courts fragments d'ADN nommés *short reads* de 100-150 pb qui peuvent être obtenus « en paires » (*paired-end*). Les données générées par le séquençage peuvent ensuite être traitées par bio-informatique. Des logiciels ont ainsi été conçus pour assembler ces *short reads* et reconstruire les séquences d'ADN du génome.

Le séquençage d'un génome entier nécessite que les *short reads* représentent l'ensemble de l'ADN du génome, avec une profondeur permettant de s'affranchir des erreurs de séquençage et de limiter les variations génétiques. Cependant, l'assemblage de ces *short reads* par les outils bio-informatiques ne permet pas d'obtenir la reconstruction des séquences des chromosomes, mais de morceaux de chromosomes nommés *contigs*. Cette limite est en partie due aux longues régions répétées pouvant représenter une forte proportion des génomes. L'assemblage des *short reads* d'ADN ne permet généralement pas de résoudre la construction des portions répétées (Treangen et Salzberg, 2011).

Depuis les années 2010, la troisième génération de séquençage est apparue à travers les technologies Pacific Biosciences et Oxford Nanopore Technology (Tableau 1 et Figure 12). Ces procédés réalisent le séquençage direct de longues molécules d'ADN. Ces séquences sont nommées *long reads* et ayant une taille en moyenne supérieure à 10 000 pb (Tableau 1). L'assemblage de *long reads* peut permettre la construction d'un génome peu fragmenté. En effet, contrairement aux *short reads* obtenus par les technologies de seconde génération, ces *long reads* peuvent résoudre l'assemblage des parties répétées du génome. Cependant, ces nouvelles techniques présentent, à l'heure actuelle, un taux d'erreur d'environ 10% (Tableau 1). Le taux d'erreur de ces assemblages peut être diminué en les associant avec des données *short reads* de seconde génération (Phillippy, 2017; Zimin et al., 2017).

Pour recenser et traiter les premières données de séquençage, le développement de la bio-informatique a été réalisé en parallèle de l'amélioration des différentes technologies de séquençage de l'ADN. Le recensement de séquences nucléiques ou protéiques, dont la fonction avait été identifiée par des études fonctionnelles, a pu être facilité à partir des années 1980 par la création de banques de données internationales telles que « *EMBL Nucleotide Sequence Data Library* » de EMBL-EBI (*European Molecular Biology Laboratory - European Bioinformatics Institute*) ou GenBank de NCBI (*National Center for Biotechnology Information*). De plus, la création d'outils d'alignements de séquences, tels que le FASTA (Lipman et Pearson, 1985), puis le BLAST (*Basic Local Alignment Search Tool*) (Altschul et al., 1990), a permis la recherche et la comparaison des séquences nucléiques présentes dans les banques (Shendure et al., 2017).

En parallèle, les premiers séquençages de génomes entiers d'organismes complexes ont débuté dans les années 90 avec le séquençage du génome du nématode *Caenorhabditis elegans*, de la plante Arabette des dames (*Arabidopsis thaliana*), de la drosophile

(*Drosophila melanogaster*) et de l'Homme (*C. elegans* Sequencing Consortium, 1998; Adams, 2000; Lander et al., 2001; The Arabidopsis Genome Initiative, 2000; Venter et al., 2001). Une fois le séquençage réalisé, ces génomes ont commencé à être annotés. L'annotation correspond à l'identification des différentes régions du génome (e.g. gènes, régions régulatrices et régions répétées). Dans le même temps, l'annotation fonctionnelle par méthode expérimentale s'est amplifiée pour connaître la fonction des différentes régions des génomes, et permette en même temps d'enrichir les banques de données.

Aujourd'hui, l'annotation fonctionnelle des nouveaux génomes est principalement réalisée de façon automatique par des approches de prédictions. Pour cela, les outils bio-informatiques utilisent des séquences de gènes présentes dans les banques de données et pour lesquelles la fonction est connue. Les limites de ces outils dépendent généralement du faible nombre de données recensées ainsi que du manque d'études fonctionnelles expérimentales. En général, seulement 30% des gènes d'un génome ont une prédiction correcte. Cette limite est fréquente lors de l'annotation des génomes de microalgues (Hovde et al., 2015; Carrier et al., 2018).

Tableau 1 : Comparaison de critères entre des instruments de séquençage Illumina et Oxford Nanopore.

Tableau provenant d'une présentation réalisée par la plateforme de séquençage Genoscope du CEA (Commissariat à l'Énergie Atomique et aux Énergies Alternatives).

Instrument	Illumina MiSeq	Illumina NextSeq	Illumina HiSeq 4000	Illumina NovaSeq 6000	Oxford Nanopore MinION	Oxford Nanopore GridION	Oxford Nanopore Prometh-ION
Total bases/run	5 à 15 Gbases	120 Gbase	1,5 Tbases	6 Tbases	5 Gbases	25 Gbases	2,4 Tbases
Temps pour 1 run (max)	3 jours	29 heures	3,5 jours	40 heures	48 heures	48 heures	64 heures
Longueur des reads	2 x 300 bases	2 x 150 bases	2 x 150 bases	2 x 150 bases	>10 Kilobases	>10 Kilobases	>10 Kilobases
Taux d'erreur	0,1 à 1%	0,1 à 1%	0,1 à 1%	0,1 à 1%	<10%	<10%	<10%
Coût de l'instrument	\$ 99 K	\$ 250 K	\$ 740 K	\$ 1000 K	\$ 1 K	/	/
Coût du séquençage par Gb	\$ 106	\$ 36	\$ 21	\$ 11	\$ 120 à 240	\$ 120 à 240	\$ 13 à 40

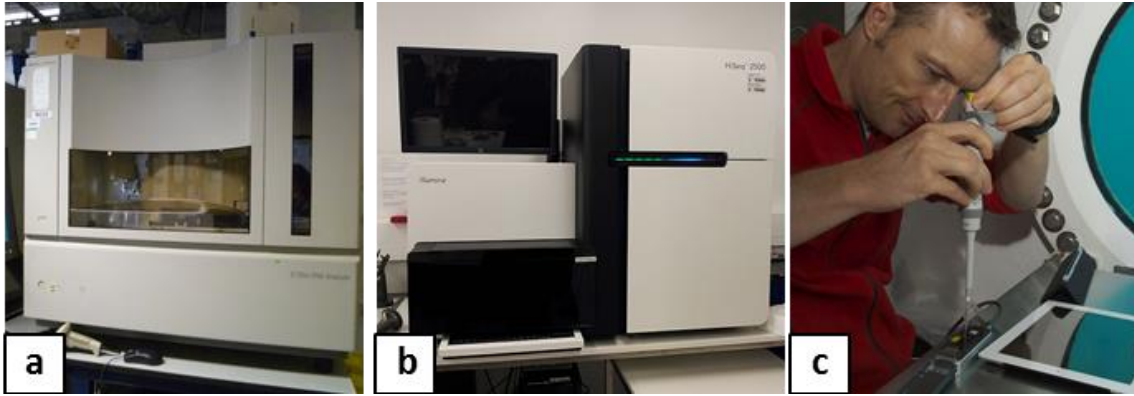


Figure 12 : Aperçu visuel de l'évolution des technologies de séquençage de l'ADN. Applied Biosystems Genetic Analysis systems (1^{er} génération), b) Illumina HiSeq (2nd génération), c) Oxford Nanopore MinION (3^{ème} génération). Photographies provenant de : [https://commons.wikimedia.org/wiki/File:ABI_3730_\(2\).JPG](https://commons.wikimedia.org/wiki/File:ABI_3730_(2).JPG);
https://fr.m.wikipedia.org/wiki/Fichier:Illumina_HiSeq_2500.jpg;
https://www.nasa.gov/mission_pages/station/research/news/dna_sequencing

2.2. La génomique chez les microalgues

L'émergence de la génomique chez les microalgues a débuté pendant les années 2000 par le séquençage des trois chromosomes nucléo-morphes de la Cryptophyte *Guillardia theta* (Douglas et al., 2001). C'est cependant à partir de 2004 que les premiers génomes de microalgues ont été publiés, ceux de la Rhodophyte *Cyanidioschyzon merolae* (Matsuzaki et al., 2004) et de la Diatomée *Thalassiosira pseudonana* (Armbrust, 2004). Quelques années plus tard, les séquençages des génomes des espèces *Ostreococcus tauri*, *Ostreococcus lucimarinus*, *Chlamydomonas reinhardtii* (Chlorophytes) et *Phaeodactylum tricornutum* (Diatomées) ont été réalisés (Derelle et al., 2006; Palenik et al., 2007; Merchant et al., 2007; Bowler et al., 2008). Actuellement, une centaine de génomes d'espèces différentes ont été séquencés (Annexe 1). La moitié de ces espèces appartiennent au taxon des Chlorophytes. Cette forte proportion s'explique par leur promiscuité taxonomique avec les plantes, d'où l'intérêt de mieux comprendre l'origine évolutive de ces derniers. Chez les Straménopiles, les génomes d'une vingtaine de microalgues appartenant principalement au groupe des Diatomées et Eustigmatophytes ont été séquencés. Chez les Alvéolées, Huit génomes d'espèces différentes sont disponibles. Un plus grand nombre appartient aux Dinophytes et au genre *Symbiodinium*. Les assemblages génomiques obtenus pour les Dinophytes sont généralement incomplets en raison de la taille importante de leurs génomes. Chez les Haptophytes, cinq génomes d'espèces différentes ont été séquencés. Le génome du coccolithophore *Emiliania huxleyii* (Read et al., 2013) et les génomes des espèces non-coccolithophores *Chrysochromulina tobin* (Hovde et al., 2015), *Chrysochromulina parva* (non publié), *Tisochrysis lutea* (Carrier et al., 2018) et Pavlova sp. (non publié). Enfin, peu de génomes ont été séquencés chez les autres taxons de microalgues. A notre connaissance, seuls les génomes de trois Rhodophytes, deux Cryptophytes, une Chlorarachniophyte et une Glaucophyte ont été séquencés (Curtis et al., 2012; Price et al., 2012) (Annexe 1). Ce recensement permet de mettre en évidence que peu de génomes de microalgues ont été séquencés par rapport à leur abondance et à leur diversité. La proportion d'espèces séquencées est très hétérogène selon les taxons. Les taxons les plus étudiés au niveau génomique sont les Chlorophytes et les Diatomées. Les assemblages de génomes de certaines espèces ont été caractérisés à l'échelle des chromosomes (exemple : *C. reinhardtii*), mais la majorité des assemblages

des génomes des autres microalgues sont majoritairement fragmentés en *contigs* (Annexe 1).

De manière générale, il existe une faible connaissance des gènes de microalgues ainsi que des autres constituants des génomes tels que les éléments répétés. Parmi les portions répétées des génomes, les éléments transposables (ETs) ont été très peu étudiés chez les microalgues, malgré leur forte proportion dans les génomes eucaryotes (Kazazian, 2004).

3. Les éléments transposables (ETs)



Figure 13 : Barbara McClintock.

B. McClintock examinant les grains de maïs dans son laboratoire à Cold Spring Harbor en 1963. Elle reçoit le prix Nobel de physiologie ou médecine en 1983 pour sa découverte des éléments transposables.

Photographie provenant de : <https://profiles.nlm.nih.gov/ps/retrieve/ResourceMetadata/LLBBPQ>

3.1. Découverte et rôle des ETs

Les travaux de Mendel (1865) sur la transmission des caractères héréditaires, puis ceux de Morgan (1910) sur le rôle des chromosomes dans cette transmission ont longtemps conduit les scientifiques à considérer les génomes des organismes comme une suite de gènes positionnés les uns après les autres (Mendel, 1865; Morgan, 1910). Le génome était ainsi considéré comme une entité stable, où seuls les gènes avaient un intérêt.

Cependant, dans les années 1950, Barbara McClintock (Figure 13) remet en question cette vision, par la découverte de séquences d'ADN pouvant changer de position (=transposer) dans le génome du maïs. Elle découvre que la mobilité de ces séquences provoque des cassures de l'ADN et peut avoir des impacts sur le phénotype des grains de maïs (McClintock, 1950) (Figure 14). Ces séquences mobiles sont aujourd'hui nommées éléments transposables (ETs). A l'époque, cette découverte remet en question le caractère stable de l'ADN. La présence d'ETs a ensuite été identifiée chez des bactéries et où ils sont nommées IS (*Insertional Sequences*) (Shapiro, 1969). A l'époque, l'existence des ETs a été en grande partie contestée par la communauté scientifique qui ne concevait pas que les génomes puissent être dynamiques (Biemont, 2010). Dans les années 1980, la découverte du rôle des ETs dans la pathologie de dysgénésie des hybrides chez *D. melanogaster* est décisive dans l'acceptation de leur importance (Picard et al., 1978; Kidwell, 1979). Les effets délétères des ETs conduisent les chercheurs Doolittle et Sapienza en 1980 à développer l'idée d'éléments « égoïstes » (Doolittle et Sapienza, 1980).

Pour la découverte de ces éléments, B. McClintock reçoit un prix Nobel en 1983, et contrairement à l'idée de l'élément « égoïste », elle propose que les ETs constituent des composants majeurs des génomes, pouvant jouer un rôle important dans l'évolution des espèces (McClintock, 1984; Biemont, 2010). Ce rôle est aujourd'hui reconnu et étudié chez la plupart des organismes (Kazazian, 2004).



Figure 14 : Grains de maïs présentant des phénotypes instables.

Les phénotypes de différentes couleurs sont causés par l'insertion ou non d'un ET dans un gène codant une enzyme liée à la synthèse du pigment anthocyanine. Image adaptée de Feschotte et al. (2002).

3.2. Définition, ubiquité et proportions des ETs

Les éléments transposables (ETs) se définissent comme des séquences d'ADN pouvant se déplacer (ou transposer) d'une région à une autre d'un chromosome, ou d'un chromosome à un autre. Les ETs sont présents chez les génomes de tous les organismes : eucaryotes, bactéries, archées et virus. Selon leur mécanisme de déplacement, les ETs peuvent générer des copies lors de leur transposition. Ainsi, ils constituent généralement une part importante des régions répétées des génomes. Chez les eucaryotes, ils représentent une fraction variable : 3% du génome de la levure, 45% chez l'homme et jusqu'à 80% chez le maïs (Lander et al., 2001; Schnable et al., 2009; Carr et al., 2012) (Figure 15). La proportion d'ETs peut être similaire pour certaines espèces évolutivement proches, tels que l'homme et le chimpanzé (The Chimpanzee Sequencing and Analysis Consortium, 2005), ou fortement varier entre des espèces du même genre, par exemple entre des espèces de drosophiles. La proportion d'ETs est près de trois fois supérieure chez *Drosophila melanogaster* par rapport à *Drosophila simulans* (Lerat et al., 2011; Vieira et al., 2012). De plus, la proportion des ETs influe sur la taille du génome de l'hôte (Figure 16). En effet, le génome du maïs possède une taille importante de 2,3 Gpb et est constitué à 80% d'ETs. D'autres mécanismes comme la polyploïdie chez les plantes expliquent aussi une taille importante de génomes. Chez les microalgues, les Dinophytes présentent également des tailles de génomes hors normes pouvant être estimées à 110 Gpb pour *Alexandrium tamarense* (LaJeunesse et al., 2005; Jaekisch et al., 2011). Cette taille importante n'est pas attribuée aux ETs mais à des événements de duplications de gènes et du génome entier (Van Dolah et al., 2009).

L'ère du séquençage haut débit a permis de mettre en évidence une grande diversité d'ETs chez des organismes eucaryotes de phylum variés. Le nombre croissant de génomes séquencés et la découverte exponentielle de nouveaux éléments ont conduit à la proposition de méthodologies de classification des ETs chez les organismes eucaryotes.

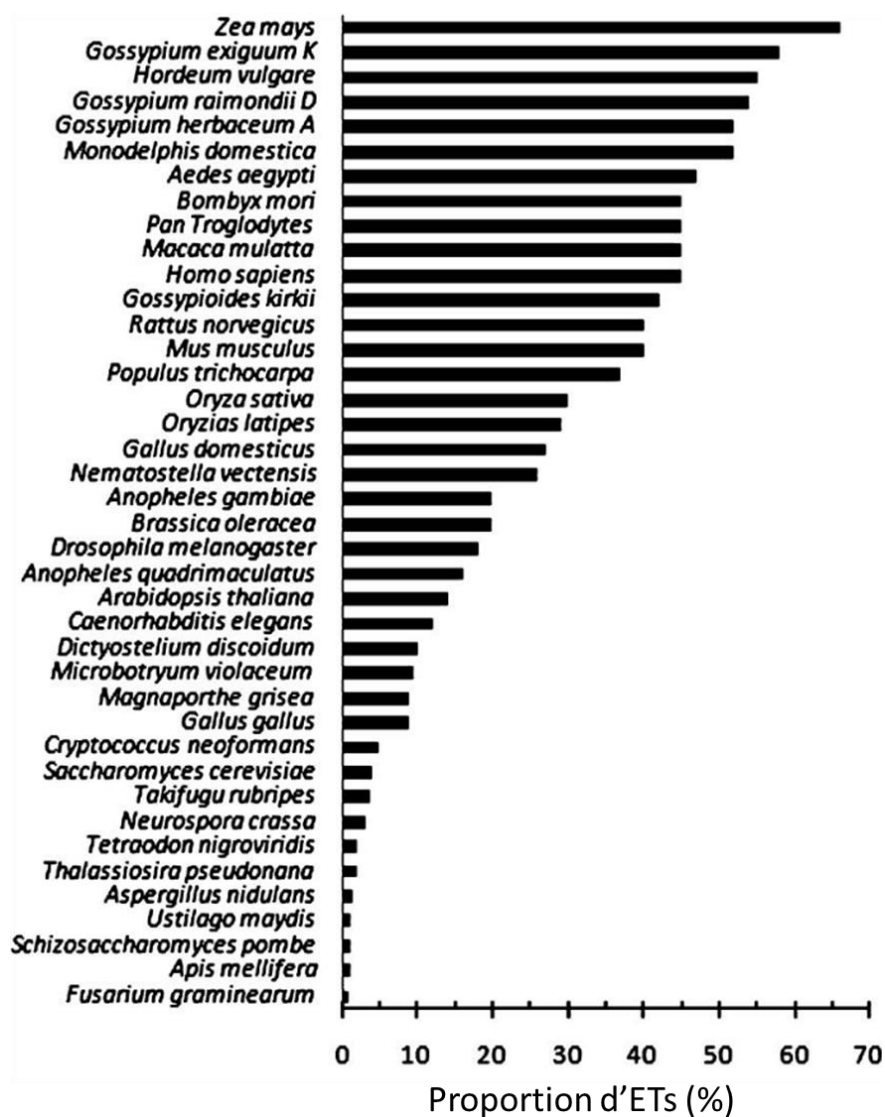


Figure 15 : Proportion d'ETs dans le génome de différents organismes.

Image provenant de Biemont (2010).

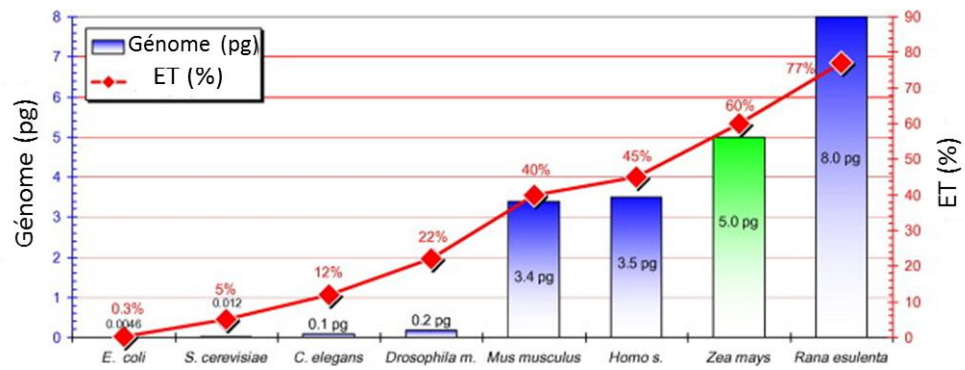


Figure 16 : Corrélation entre la taille des génomes (ADN, picogramme) et la proportion en ETs qu'ils possèdent.

Image adaptée de Alzohairy et al. (2013).

3.3. La classification des ETs:

Historique de la classification

En 1989, David Finnegan propose la première classification des ETs, en suggérant de les répartir en deux classes selon leur mécanisme de déplacement (Finnegan, 1989). La Classe I regroupant les ETs qui se déplacent par la rétrotranscription d'un ARN intermédiaire (ADN-ARN-ADN) et la Classe II regroupant les ETs capables de se déplacer directement d'une région génomique à une autre (ADN à ADN). Cette première classification fut adoptée par la communauté scientifique. Suite au nombre croissant de nouveaux éléments découverts, celle-ci fut affinée par Capy et al (1997) par l'ajout de sous-classes, de superfamilles et de familles (Capy et al., 1997). Ensuite, Hull (2001) proposera l'union des ETs de Classe I avec les virus, en se basant sur les similarités qu'ils partagent avec les rétro-éléments viraux et les relations phylogénétiques de leur rétro-transcriptase (Hull, 2001). Cependant, le besoin croissant d'une classification unique des ETs continuera à être soulevé (Capy, 2005). Plus tard, Wicker et al proposeront en 2007 une méthode de classification des ETs pour les eucaryotes. Ce premier système de classification hiérarchique a été conçu sur la base du mécanisme de transposition, des similarités de séquences et de leurs relations structurelles (Wicker et al., 2007). Cette méthode de classification est aujourd'hui la plus utilisée par la communauté scientifique travaillant sur les ETs. La classification de Wicker et al 2007 est en concurrence avec celle proposée par Kapitonov et Jurka (Kapitonov et Jurka, 2008) dont le laboratoire gère la base de données Repbase. Cette dernière utilise une nomenclature différente.

Ce manuscrit se base sur la méthode de classification de Wicker et al 2007, en raison de son utilisation prépondérante par la communauté scientifique.

Méthode générale de classification selon Wicker et al (2007)

Cette classification permet à des scientifiques non-experts de regrouper les ETs en Classes, Sous-classes, Ordres, Superfamilles et familles (Figure 17). Le premier niveau comprend deux Classes distinctes prenant en compte le mécanisme de transposition : pour la Classe I, la transposition inclut le passage par un ARN. Cette classe est donc identique à celle décrite par Finnegan (Finnegan, 1989). En revanche, la Classe II est subdivisée en deux Sous-classes, dont les éléments qui les composent sont distingués par le nombre de brins d'ADN coupés lors de la transposition (Wicker et al., 2007). A l'intérieur de ces deux Classes, les ETs sont ensuite différenciés dans des Ordres selon le type d'enzymes qu'ils codent (Figure 17). Pour chaque Ordre, les ETs sont rassemblés en Superfamilles, selon la structure des séquences, leurs domaines protéiques ou leur TSD (*Terminal Side Duplication*). Le TSD correspond à un motif que la transposase recopie au niveau du point d'insertion. Les ETs sont ensuite regroupés en familles par similarité nucléique. Deux ETs appartiennent à la même famille, s'ils partagent : au moins 80% d'identité sur au moins 80% de leur séquence nucléique, et ceci, pour des séquences ayant une longueur supérieure à 80 pb. Cette règle nommée *80-80-80* permet ainsi d'apporter une méthodologie simple et unifiée de la classification des ETs.

Distinction des éléments autonomes et non-autonomes

Un élément autonome code pour les gènes nécessaires à sa transposition. En revanche, il peut ne pas être actif. Les ETs autonomes actifs utilisent la machinerie cellulaire de l'hôte pour exprimer leurs gènes et synthétiser les protéines essentielles à leur transposition. Un ET non-autonome ne code pas les gènes permettant la synthèse des protéines participant à son déplacement. L'absence de ces gènes est soit originelle, soit ces gènes ont été perdus au cours du temps, du fait, entre autre, d'éventuelles mutations. Cependant, un élément non-autonome peut tout de même être déplacé en *trans*, via l'action de protéines codées par des ET autonomes. En conclusion, il est donc important de garder à l'esprit que les génomes sont la plupart du temps constitués d'un ensemble d'ETs structurellement différents, pouvant être potentiellement autonomes, non-autonomes, tronqués ou sous forme de reliques. Ces derniers sont les témoignages d'anciens ETs complets qui ont fortement divergés au cours du temps (Lerat et al., 2003).

Classification		Structure	TSD
Order	Superfamily		
Class I (retrotransposons)			
LTR	<i>Copia</i>	→ GAG AP INT RT RH →	4-6
	<i>Gypsy</i>	→ GAG AP RT RH INT →	4-6
	<i>Bel-Pao</i>	→ GAG AP RT RH INT →	4-6
	<i>Retrovirus</i>	→ GAG AP RT RH INT ENV →	4-6
	<i>ERV</i>	→ GAG AP RT RH INT ENV →	4-6
DIRS	<i>DIRS</i>	→ GAG AP RT RH YR ←	0
	<i>Ngaro</i>	→ GAG AP RT RH YR → → →	0
	<i>VIPER</i>	→ GAG AP RT RH YR → → →	0
PLE	<i>Penelope</i>	← RT EN →	Variable
LINE	<i>R2</i>	RT EN	Variable
	<i>RTE</i>	APE RT	Variable
	<i>Jockey</i>	ORF1 APE RT	Variable
	<i>L1</i>	ORF1 APE RT	Variable
	<i>I</i>	ORF1 APE RT RH	Variable
SINE	<i>tRNA</i>		Variable
	<i>7SL</i>		Variable
	<i>5S</i>		Variable
Class II (DNA transposons) - Subclass 1			
TIR	<i>Tc1-Mariner</i>	↔ Tase* ↔	TA
	<i>hAT</i>	↔ Tase* ↔	8
	<i>Mutator</i>	↔ Tase* ↔	9-11
	<i>Merlin</i>	↔ Tase* ↔	8-9
	<i>Transib</i>	↔ Tase* ↔	5
	<i>P</i>	↔ Tase ↔	8
	<i>PiggyBac</i>	↔ Tase ↔	TTAA
	<i>PIF-Harbinger</i>	↔ Tase* ORF2 ↔	3
	<i>CACTA</i>	↔ Tase ORF2 ↔	2-3
Crypton	<i>Crypton</i>	YR	0
Class II (DNA transposons) - Subclass 2			
Helitron	<i>Helitron</i>	RPA Y2 HEL	0
Maverick	<i>Maverick</i>	C-INT ATP CYP POL B	6

→ Terminaisons LTRs ↔ Terminaisons TIRs — Régions codantes — Régions non codantes
 — Motifs —/— Régions présentant des ORFs additionnels
 TSD: Duplication du site cible

Figure 17 : Système de classification des ETs proposé par Wicker et al. en 2007.

Image adaptée de Wicker et al. (2007)

Classe I : Les Rétrotransposons

La Classe I, ou rétro-transposons, regroupe l'ensemble des ETs pouvant transposer *via* l'intermédiaire d'un ARN. Cet ARN est ensuite utilisé comme brin matrice par la transcriptase inverse, qui est généralement codée par l'élément pour synthétiser un ADN complémentaire. Chaque cycle de transposition produit une ou plusieurs nouvelles copies de l'élément. Par ce mécanisme, les rétrotransposons peuvent envahir massivement le génome hôte et participent ainsi à l'augmentation de la taille de celui-ci. Ces éléments sont regroupés en cinq ordres distincts selon leur mécanisme, leur organisation et la phylogénie de leur transcriptase inverse. La Classe I est composée des LTR (*Long Terminal Repeats*), DIRS (*Dictyostelium-like element*), Penelope (ou PLE pour *Penelope-Like Elements*), LINE (*Long Interspersed Nuclear Element*) et les éléments non-autonomes SINE (*Short Interspersed Nuclear Element*). Parmi la Classe I, les Ordres LTR, LINE et SINE sont les plus étudiés du fait de leur ubiquité chez les eucaryotes. Ils sont présentés plus en détail ci-dessous.

Les LTR (*Long Terminal Repeats*)

Le nom LTR fait référence aux longues terminaisons LTRs identiques qui bordent la partie interne de la séquence nucléique de ces éléments (Figures 17, 18 et 19). Les rétrotransposons LTR sont divisés en cinq Superfamilles : les Copia, les Gypsy, les Bel-Pao, les Rétrovirus et les ERV (Rétrovirus endogènes) (Figure 17). Les rétrotransposons LTR codent pour des protéines à travers deux principaux cadres de lecture (ORF, *Open Reading Frame*) nommés GAG et POL (Figure 17). La protéine GAG peut être soit une partie du cadre de lecture ouvert de la polyprotéine, soit une partie distincte. La polyprotéine POL contient une protéinase aspartique (AP), une intégrase (IN), une rétro-transcriptase (RT) et une Ribonucléase H (RH) (Figures 17 et 18). La protéine GAG produit des particules virales qui se lient et enveloppent l'ARN du rétro-transposon, les RT, RNase H et IN. Les enzymes RT et RNase permettent la rétro-transcription de la matrice ARN, et l'IN permet l'insertion de la nouvelle copie du rétrotransposon LTR dans une région du génome (Figure 18) (Sabot et Schulman, 2006; Schulman, 2015). Les Superfamilles Rétrovirus et ERV sont également capables de synthétiser une protéine enveloppe (ENV). Les rétrovirus sont similaires à la Superfamille Gypsy. Il est possible que leur histoire évolutive soit liée, bien que cette relation ne soit pas claire (Lerat et Capy, 1999; Hayward, 2017). Chez les éléments LTR, les séquences régulatrices sont présentes dans chacune des terminaisons LTRs (Mager et Stoye, 2015). Au cours du temps, les éléments autonomes peuvent accumuler des mutations dans leur séquence interne, de sorte que certains ne pourront plus être capables de coder les protéines nécessaires à leur déplacement. Dans ce cas précis, si leurs séquences terminales restent intactes, ces éléments pourront être déplacés *in trans* dans le génome hôte, via les enzymes synthétisées par d'autres éléments autonomes. Ces derniers éléments sont nommés TRIM (*Terminal-repeat Retrotransposons In Miniature*, séquence < 4000 bp) ou LARD (*Large Retrotransposon Derivative*, séquence > 4000 pb) selon leur taille (Figure 19) (Witte et al., 2001; Kalendar et al., 2004; Sabot et al., 2006).

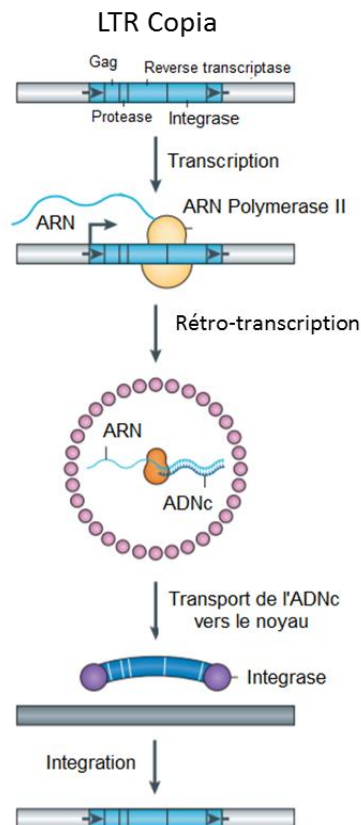


Figure 18 : Représentation schématique du mécanisme de transposition d'un rétrotransposon LTR Copia.

La transcription est réalisée à partir du promoteur situé dans une des terminaisons LTRs. Le transcrit est ensuite exporté en dehors du noyau et enveloppé dans des particules virales synthétisées par la protéine GAG. La rétro-transcription de l'ARN en ADNc est alors réalisée par la RT. Les particules se déplacent vers le noyau et l'ADNc est ensuite intégré dans le génome sous l'action de l'intégrase (IN). Image adaptée de Levin et Moran (2011).

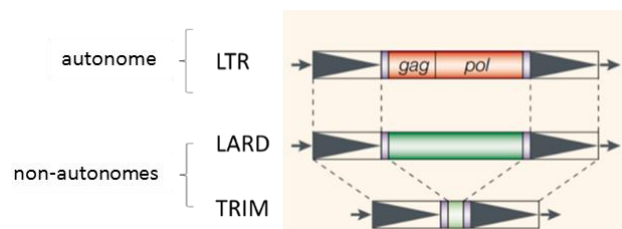


Figure 19 : Illustration de la relation entre les LTR, LARD et TRIM.

Un rétrotransposon LTR autonome pourra devenir un LARD ou un TRIM (non-autonomes) s'il subit des mutations à l'intérieur de sa séquence (insertions ou délétions), tout en conservant ses séquences terminales intactes. Image adaptée de Feschotte et al. (2002).

Les LINE (*Long Interspersed Nuclear Element*)

Contrairement aux LTR, les LINE sont des éléments ne possédant pas de terminaisons aux extrémités de leur séquence (Figure 17). Ils sont subdivisés en cinq Superfamilles : R2, RTE, Jockey, L1 et I. Les éléments de Superfamille L1 sont particulièrement étudiés car ils représentent près de 20% du génome humain (Lander et al., 2001). Ils peuvent participer à des mutations génétiques pouvant être héréditaires et être responsables de certains types de cancers (Chénais, 2015). Les LINE codent pour la synthèse d'une transcriptase inverse (RT) et d'une nucléase (APE) (Figures 17 et 20). D'autres Superfamilles synthétisent des protéines complémentaires comme une enveloppe (ENV) pour les R2 ou une Ribonucléase H (RH) pour les I. Enfin, les LINE possèdent également un promoteur spécifique de ARN Polymérase II dans leur région 5' UTR (5' *Untranslated Transcribed Region*) (Richardson SR, Doucet AJ, et al 2015).

Les SINE (*Short Interspersed Nuclear Elements*)

Les SINE sont des petites séquences non-autonomes qui ont pour origine la rétrotransposition accidentelle de transcrits d'ARN Polymérase III. Ces séquences possèdent ainsi une ARN polymérase III interne qui leur permet d'être exprimées et d'être mobilisées *in trans* par les protéines synthétisées par des LINE (Figure 20). Certains SINE ne peuvent se déplacer que par l'intervention de protéines générées par un LINE spécifique, tandis que d'autres peuvent se déplacer par l'intervention de LINE différents. Il existe trois Superfamilles de SINE, les ARNt, 7SL et 5S. La famille de SINE la plus étudiée est celle des éléments *Alu* dont les répétitions constituent plus de 11% du génome humain (Lander et al., 2001).

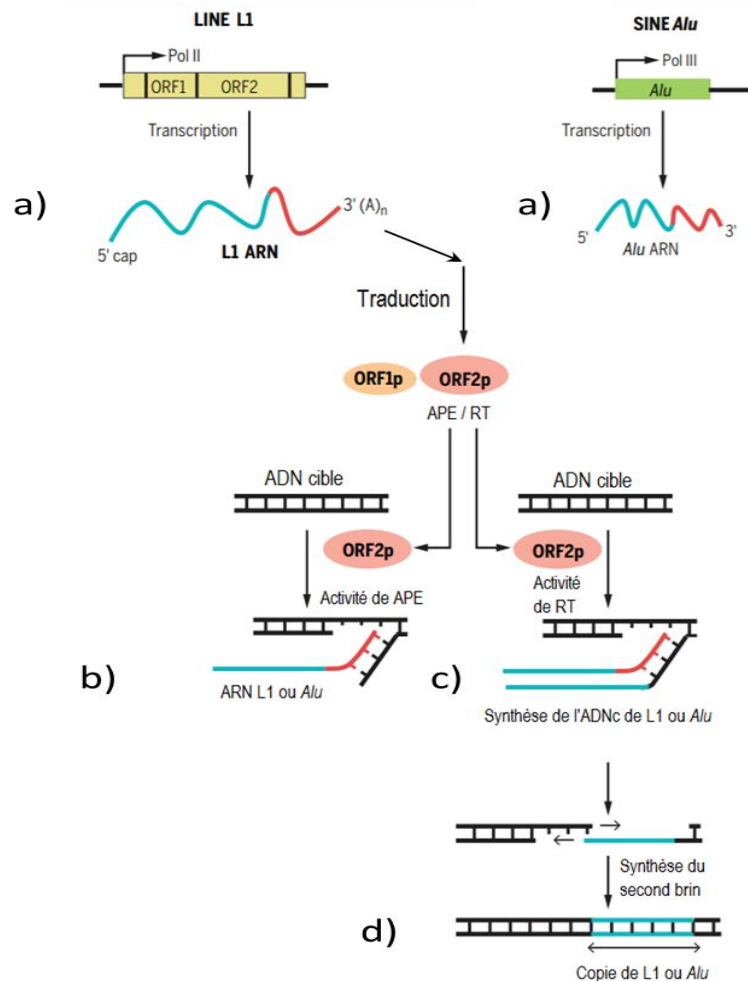


Figure 20 : Illustration des mécanismes de transposition des LINE et des SINE.

a) La transcription d'un LINE L1 ou d'un SINE *Alu* produit un ARN. b) Par l'activité de l'endonucléase, cet ARN s'amorce sur un site cible au niveau de sa queue poly-dA. c) L'ADNc est transcrit par rétrotranscription par la RT et l'ARN est séparé de l'ADNc. d) Enfin, le brin complémentaire d'ADN est synthétisé, permettant d'achever le processus d'intégration. Image adaptée de Elbarbary et al. (2016).

Classe II : Les transposons à ADN

Les transposons à ADN comportent deux sous-classes qui se distinguent par le nombre de brins d'ADN coupés pendant la transposition. La première est composée des Ordres des TIR (*Terminal Inverted Repeats*) et des Crypton, tandis que la seconde regroupe les Ordres des Helitron et des Maverick (ou Polinton). Les Helitron se déplacent par un mécanisme dit de « cercle roulant » (Yang et Bennetzen, 2009), tandis que les Maverick sont supposés se déplacer par une réplication extra-chromosomale de leur séquence (Kapitonov et Jurka, 2007). En raison de leur ubiquité, seul l'Ordre des TIR est détaillé ci-dessous.

Les TIR (*Terminal Inverted Repeats*)

L'Ordre des TIR regroupe les éléments dont les extrémités du transposon sont constituées de terminaisons inversées et répétées nommées TIRs (*Terminal Inversed Repeats*) (Figure 17). Ces éléments possèdent une région centrale comportant une ORF codant pour une transposase (TASE). Les terminaisons TIRs sont reconnues par les protéines TASE qui excisent l'élément de sa position initiale et l'insère dans une nouvelle portion génomique (Figure 21). Ils sont regroupés en neuf Superfamilles : Mariner, hAT, Mutator, Merlin, Transib, P, PiggyBac, PIF-Harbinger et CACTA. Les Superfamilles de TIR peuvent être en partie identifiées par leur TSD (*Target Site Duplication*) (Figure 21). Par exemple, les éléments *Mariner* visent le site d'insertion composé de deux nucléotides « TA ». Ce motif est dupliqué par la transposase lors de l'insertion, donnant lieu au TSD (Figure 21). Les TIR possèdent également un promoteur ARN Polymérase II. Ces ETs peuvent générer des copies en transposant pendant la réplication chromosomique, d'une position déjà répliquée à une autre que la fourche de réplication n'a pas encore atteinte (Figure 22) (Feschotte et Pritham, 2007). Ils peuvent également exploiter la réparation du « vide », laissé après l'excision d'un élément, pour créer une copie supplémentaire sur le site du donneur (Figure 22) (Feschotte et Pritham, 2007). Ils constituent des éléments non-autonomes lorsqu'ils possèdent une délétion interne importante et des extrémités intactes. Ceux présentant une taille d'une centaine de pb sont nommés MITE. Les TIR non-autonomes peuvent être déplacés en *trans* par l'intermédiaire de transposases synthétisées par d'autres éléments autonomes (Figure 23).

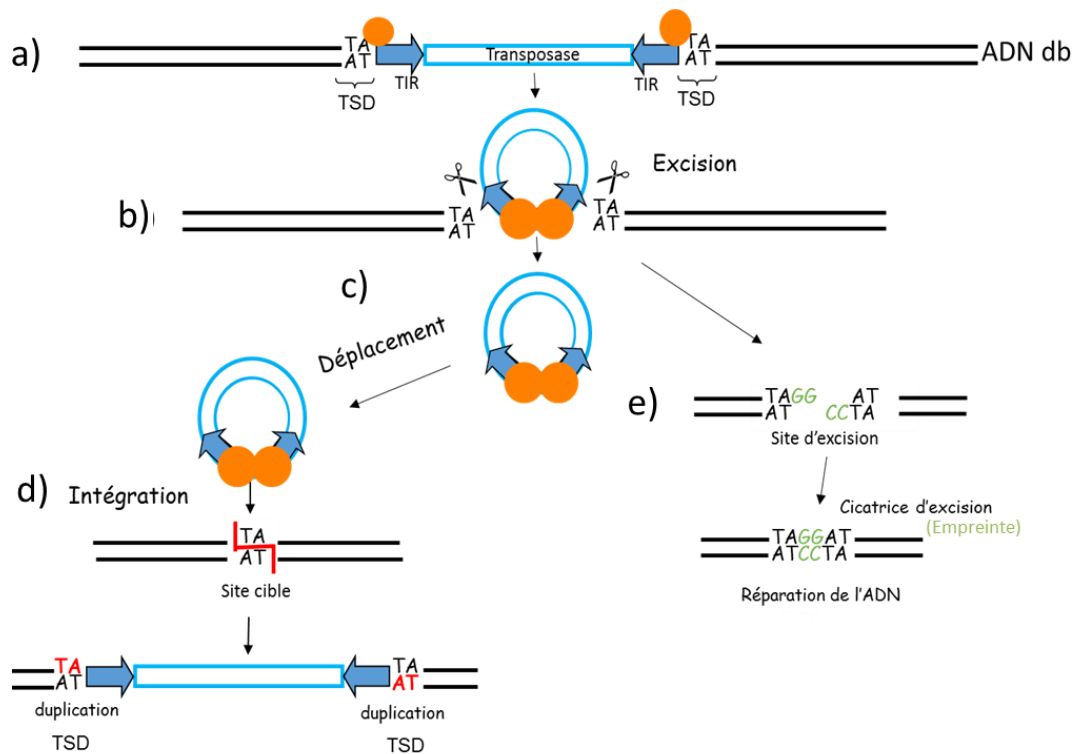


Figure 21 : Illustration du mécanisme de transposition d'un TIR Mariner.

a) L'expression d'un TIR Mariner permet la synthèse de transposases par la cellule. b) Les transposases reconnaissent les terminaisons TIRs du Mariner, qui est alors excisé. c) Le TIR Mariner est déplacé de sa position d'origine et est d) inséré dans une nouvelle région du génome. Pour cela, il cible un site cible composé de deux nucléotides « TA ». Son intégration génère la duplication de ce site d'insertion nommé TSD. e) Son départ peut laisser une cicatrice (ou empreinte) dans sa position d'origine (quelques nucléotides provenant des terminaisons TIRs).

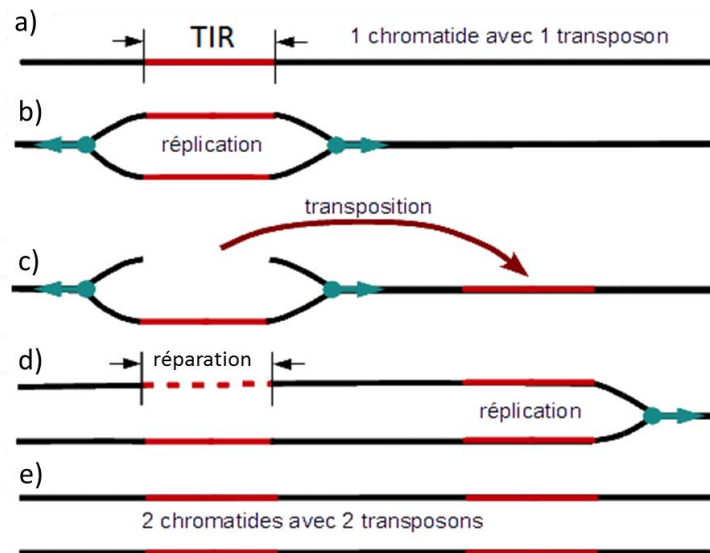


Figure 22 : Illustration de deux mécanismes permettant la duplication d'un élément TIR.

a) Un TIR est actif sur une chromatide (ADN double brin), b) un transposon peut se dupliquer si sa transposition est réalisée lors de la réplication de l'ADN. c) Le transposon se déplace de sa région d'origine ou l'ADN a été répliqué vers une nouvelle région encore non répliquée. d) Le transposon qui vient de transposer est alors répliqué à nouveau (premier mécanisme). Ensuite, la cassure qu'il a laissée peut être réparée (*gap repair*) en prenant comme matrice le transposon situé sur le deuxième brin d'ADN (second mécanisme). e) Chaque chromatides possède alors deux copies du TIR. Image provenant de <http://www.evolution-biologique.org/mecanismes/diversite-genetique/elements-mobiles.html>

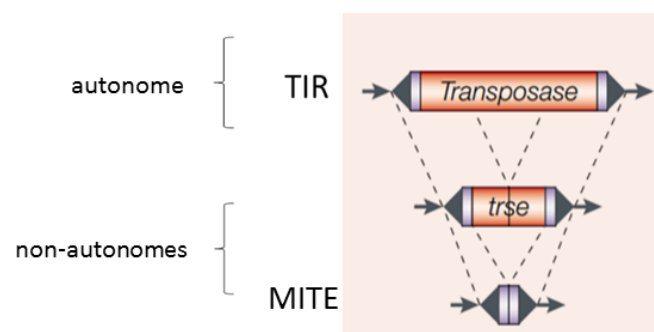


Figure 23 : Illustration de la relation entre des TIR autonomes et non-autonomes.

Un TIR autonome peut devenir un élément non-autonome si sa séquence interne subit des délétions et qu'il conserve ses séquences terminales TIRs intactes. Ceux présentant une taille d'une centaine pb sont nommés MITE. Image adaptée de Feschotte et al. (2002).

3.4. Le cycle de vie des ETs

La majorité des ETs présents dans les génomes sont généralement inactifs. Soit, parce qu'ils ont subi des mutations au cours du temps, soit parce qu'ils sont inactivés par des mécanismes de défense de l'hôte, telle que la méthylation de l'ADN (cf. Introduction § 3.5). La proportion d'ETs actifs chez l'homme a été estimée à moins de 0,05% (Mills et al., 2007). Comme évoqué précédemment, les ETs possèdent des séquences régulatrices similaires à des promoteurs. Ces promoteurs sont régulés par de nombreux facteurs temporels ou spatiaux dépendants du type cellulaire, ou en réponse à des événements extérieurs de stress abiotiques ou biotiques (Lisch, 2012).

Les mécanismes de transferts d'ETs

Pour persister au cours du temps, un ET actif doit être transféré d'un individu à un autre. Le transfert vertical est reconnu comme le mécanisme principal de transfert des ETs. Chez les organismes sexués, les nouveaux événements de transposition doivent avoir lieu dans les cellules germinales afin qu'ils puissent persister par hérédité. Les transferts horizontaux des ETs (HTT, *Horizontal Transfer of Transposable elements*) entre des génomes sont un moyen de transmission alternatif correspondant au transfert d'un ET entre des individus d'espèces différentes. Le premier cas de HTT identifié chez les eucaryotes a été celui de l'élément *P* entre différentes espèces de *Drosophiles* (Daniels et al., 1990). Aujourd'hui l'équivalent de 2 855 événements (HTT-database, 10 juin 2018) a été recensé chez des Métazoaires et Viridiplantae. Des rétrotransposons LTR similaires ont également été retrouvés dans les génomes des Diatomées *Phaeodactylum tricornutum* et *Thalassiosira pseudonana*, suggérant des événements de HTT entre ces deux espèces (Maumus et al., 2009). Les mécanismes permettant cette transmission sont encore peu connus (Schaack et al., 2010). Parmi les vecteurs potentiels, les virus sont de plus en plus suggérés (Schaack et al., 2010; Gilbert et Feschotte, 2018).

Les différentes phases de vie des ETs

Selon Kidwell et Lisch (2001), le cycle de vie d'un ET peut se diviser en trois phases : la phase d'invasion, la phase de maturité et la phase de sénescence (Kidwell et Lisch, 2001). La phase d'invasion correspond à l'insertion d'un ET actif dans un génome hôte, via un transfert vertical ou horizontal (Figure 24). Elle correspond à l'augmentation du nombre de copies de celui-ci dans le génome. Pour les éléments de Classe I, cette phase peut s'accompagner par une expansion massive des copies de l'éléments dans le génome (Chénais et al., 2012). Lorsque le nombre de copies actives devient inférieur au nombre de copies inactives, l'ET débute sa phase de maturité (Figure 24). L'apparition de mutations au cours du temps et les systèmes de régulation/défense (cf. Introduction § 3.5) limitent le nombre de copies actives d'un ETs (Figure 24). Enfin, quand toutes les copies de cet élément sont inactives, c'est la phase de sénescence (Figure 24). Pour échapper à ce cycle, un ET peut envahir de nouveaux organismes *via* des transferts horizontaux (Figure 24). Il peut ainsi amorcer un nouveau cycle de vie dans un nouveau génome hôte (Kidwell et Lisch, 2001).

3.5. Les mécanismes de régulation des ETs

Par leur mobilité, les ETs constituent une menace potentielle pour le génome hôte car ils sont susceptibles de générer des mutations possiblement délétères. Face à l'activité des ETs, les organismes ont mis en place divers mécanismes de défense pour contrôler l'expression et la transposition de ces éléments. Plusieurs mécanismes de défenses face au ETs sont attribués à des phénomènes épigénétiques, qui participent à la régulation de l'expression des génomes, sans modifications irréversibles, et transmissibles par des divisions cellulaires et héréditaires (Daxinger et Whitelaw, 2012). Différents mécanismes ont pu être démontrés chez diverses espèces, tels que les mécanismes de modification des histones, la méthylation de l'ADN ou la dégradation des transcrits d'ETs *via* des petits ARNs non codants (Slotkin et Martienssen, 2007). Enfin, il est également suggéré que des événements de larges délétions d'ADN pourraient constituer un mécanisme de défense du génome hôte pour contrer l'expansion des ETs actifs (Kapusta et al., 2017).

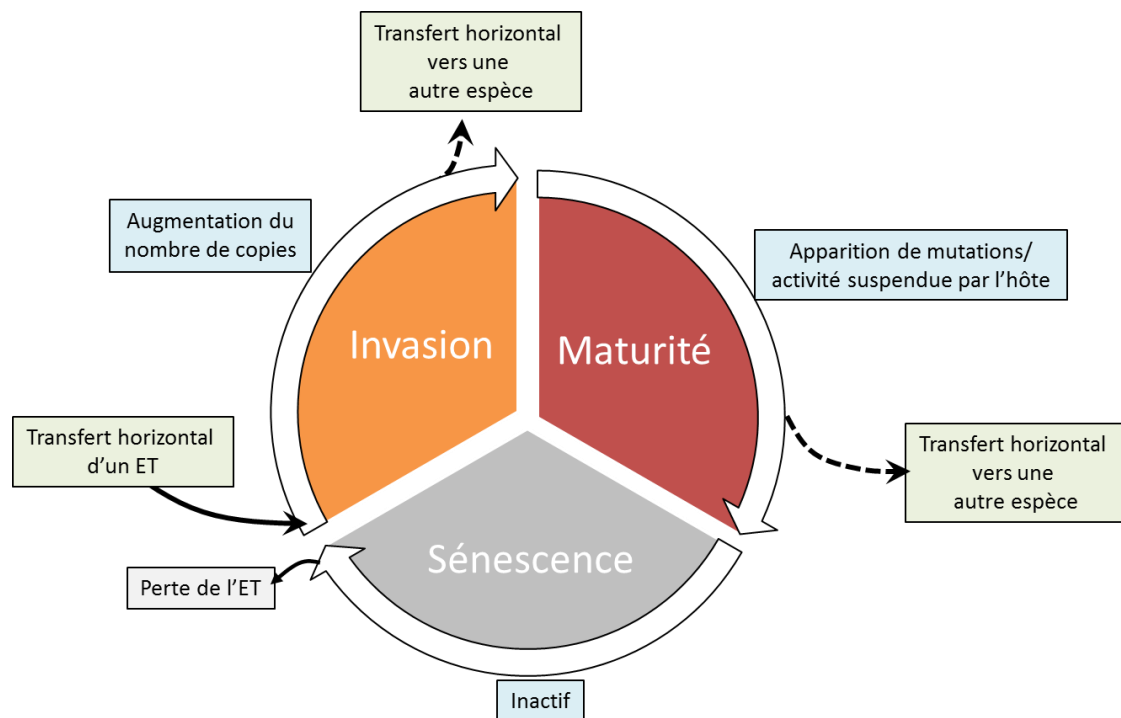


Figure 24 : Schéma représentatif du cycle de vie théorique des ETs.

Le cycle d'un ET est composé de trois phases : la phase d'invasion, la phase de maturité et la phase de sénescence. Les flèches en pointillé représentent des étapes « facultatives » du cycle. Image adaptée de Muñoz-López et García-Pérez (2010).

3.6. Impacts des ETs sur les génomes

Dans quelles régions des génomes les ETs sont-ils présents ?

Lorsqu'un ET est actif, la réalisation de nouvelles insertions est décisive pour la suite de son cycle de vie (Figure 24). Cependant, selon la région où l'élément s'insère, la mutation qu'il génère peut être délétère pour l'hôte. Dans ce cas de figure, cette nouvelle insertion ne sera pas sélectionnée par la population dans le milieu naturel. Des auteurs suggèrent que certains éléments transposent vers des zones génomiques propices à leur sélection (Levin et Moran, 2011). En effet, certains ETs sont retrouvés majoritairement au niveau de régions transcrites, riches en gènes. Il est possible que ces régions favorisent l'activité des ETs et leur transposition de novo. Par exemple, les TIR *Mutator* chez le maïs, les TIR *P* chez la *Drosophile* ou les rétrotransposons LTR *Tf1* chez *Schizosaccharomyces pombe* sont fortement présents dans la région 5' des gènes (Liao et al., 2000; Liu et al., 2009b; Guo et Levin, 2010). Chez *S. pombe*, l'étude de plus de 70 000 événements d'intégration de *Tf1* montre que 95% d'entre elles sont situées en amont de gènes (Guo et Levin, 2010). Au contraire, certains éléments sont retrouvés dans des régions pauvres en gènes. Il est à penser que cela minimise leurs possibles impacts délétères sur l'hôte et maximise ainsi leur sélection neutre. Une étude a également rapporté que les rétrotransposons LTR *Ty5* de la levure s'insèrent au sein de zones pauvres en gènes en condition normale, mais ciblent des régions riches en gènes en condition de stress (Levin et Moran 2011). D'une autre manière, certains LINE semblent s'insérer davantage dans les télomères chez la drosophile ou la microalgue Chlorophyte *Chlorella* sp. (Higashiyama et al., 1997). Ainsi, leurs nouvelles insertions participent au maintien de l'intégrité des chromosomes (George et al., 2010). Enfin, certains ETs ne sont pas retrouvés au sein de régions spécifiques, tels que les LINE *L1* et les SINE *Alu*, dont la forte présence est observée en différents points du génome humain (Lander et al., 2001; Levin et Moran, 2011).

Impacts des nouvelles insertions d'ETs sur les gènes

Malgré les différents mécanismes employés par les organismes pour limiter l'activité des ETs de nouvelles insertions de ces éléments peuvent avoir lieu, et possiblement impacter les gènes du génome hôte.

Si un ET s'insère dans un exon ou intron d'un gène, cela peut conduire à l'inhibition ou à un changement d'expression de celui-ci (Figure 25). Par exemple, la sélection d'un phénotype albinos du téléostéen *Oryzias latipes* a été attribuée à la présence d'un TIR hAT dans le premier exon du gène codant la tyrosinase. Cette enzyme participe à la biosynthèse de la mélanine. La présence de cet ET induit l'apparition d'un codon Stop prématuré provoquant ainsi une perte de fonctionnalité du gène en question (Koga et al., 1995). Chez l'homme, de nombreuses maladies sont causées par des déplacements d'ETs dans le génome (Chénais, 2015). Pour exemple, la neurofibromatose de type 1 (ou maladie de Recklinghausen) est une maladie génétique fréquente (1 personne atteinte sur 3 000 à 4 000) dont l'apparition peut être due à une mutation provoquée par un ET. En effet, le gène nommé *NF1* est un gène suppresseur de tumeur qui code pour une protéine importante, la neurofibromine. Une étude a décrit que l'insertion d'un élément SINE *Alu* dans un des introns du gène *NF1* induisait une altération de son épissage, participant ainsi à l'apparition de la maladie (Wallace et al., 1991).

L'insertion d'un ET au niveau d'une région régulatrices 5' d'un gène peut également impacter celui-ci, en modifiant son expression (Figure 25). Par exemple, chez la vigne, l'insertion d'un rétrotransposon LTR dans le promoteur du gène *VvmybA1*, qui participe à la synthèse des anthocyanines, inhibe son expression et conduit à un phénotype de type « chardonnay » (Figure 26) (Lisch, 2012). De nombreux exemples d'impacts d'insertions d'ETs sur les gènes sont recensés et illustrés Figure 25.

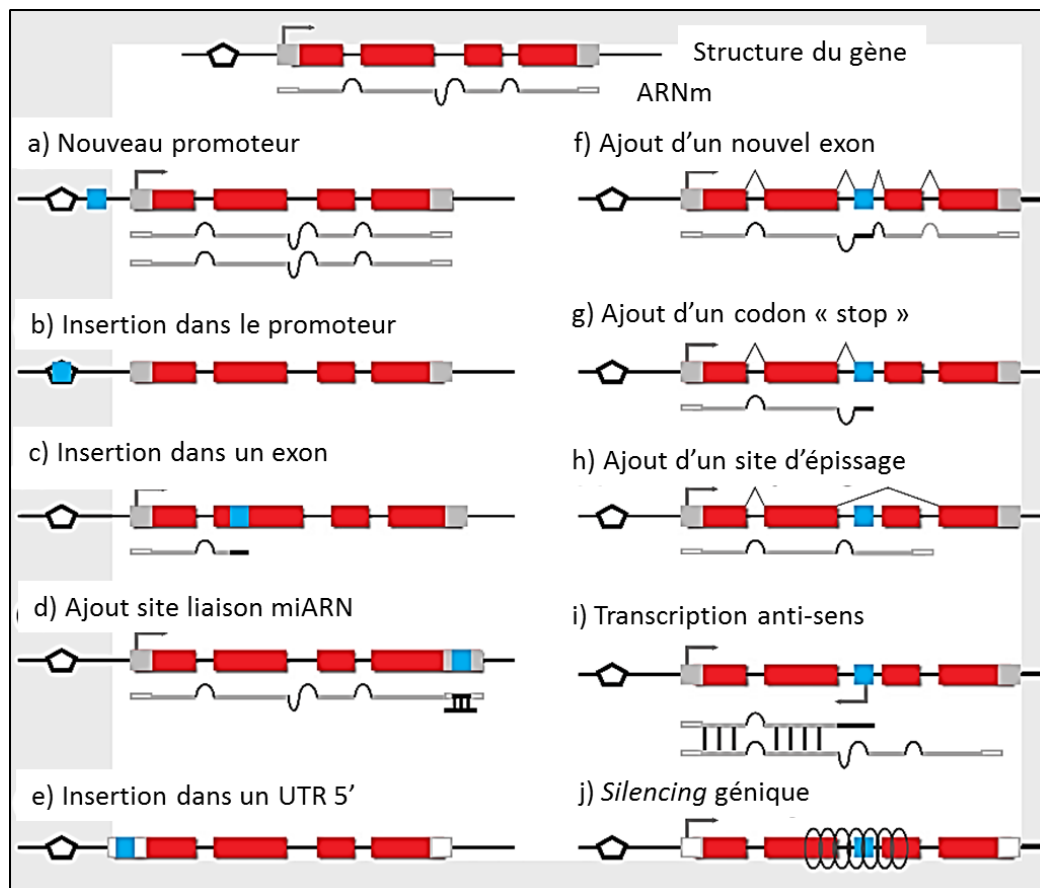


Figure 25 : Exemple de différents impacts pouvant être causés par l'insertion de novo d'un ET dans un gène ou à sa proximité.

a) Les ETs peuvent s'insérer au niveau de la région 5' d'un gène et apporter un nouveau promoteur pouvant conduire à la surexpression ou la répression du gène. Ils peuvent s'insérer dans b) un promoteur et perturber l'expression du gène, ou c) dans un exon et changer son cadre de lecture, ou introduire un codon stop. d) En s'insérant, un ET peut ajouter un site de liaison de micro ARN. d) Un ET peut aussi s'insérer au niveau e) d'un UTR 3' d'un gène et changer la séquence régulatrice de celui-ci. Il peut aussi s'insérer au niveau d'un intron et avoir divers impacts : f) être incorporé comme un nouvel exon, g) introduire un codon « stop » conduisant à un transcrit tronqué, h) introduire de nouveaux sites d'épissages et créer des épissages alternatifs, i) provoquer une transcription anti-sens qui peut interférer avec le transcrit sens, j) partager une inhibition épigénétique et inactiver un gène. Légende : les pentagones représentent des régions régulatrices, les carrés gris les *UTRs*, les rectangles rouges des exons et les carrés bleus des ETs. Image adaptée de Casacuberta et González (2013).

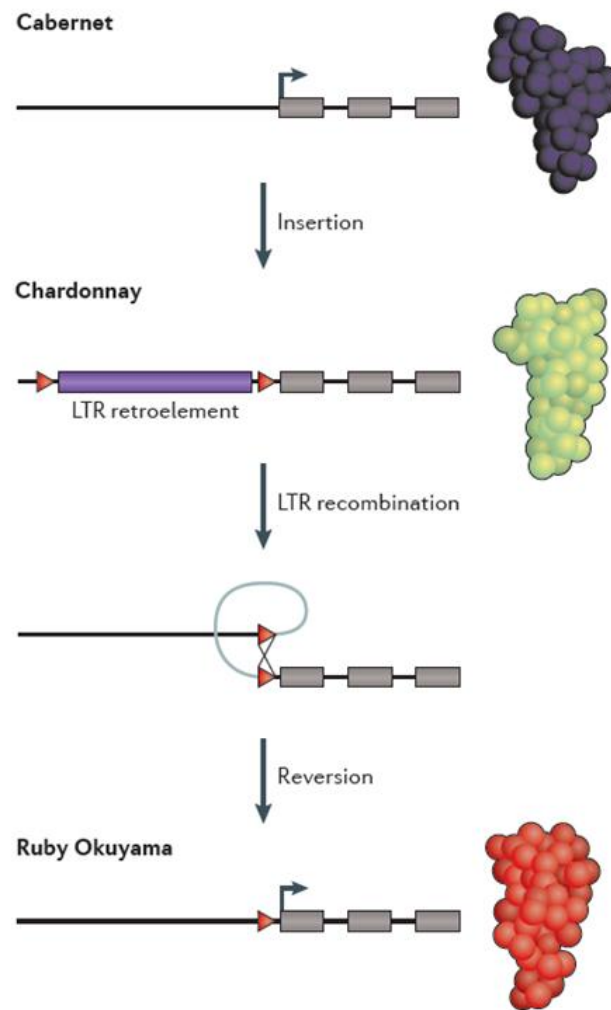


Figure 26 : Impacts d'un rétrotransposon LTR sur la couleur de différentes variétés de vigne.

Pour la variété « Cabernet », aucune insertion d'ET n'est présente à proximité du gène *Vvmyb1A* qui participe à la synthèse des anthocyanes. L'insertion du rétrotransposon LTR *Gret1* dans le promoteur de *Vvmyb1A* inhibe l'expression de ce gène et interrompt la production d'anthocyane, conduisant ainsi à la variété « Chardonnay ». La recombinaison homologue des terminaisons du rétrotransposon LTR *Gret1* donne lieu à présence d'une terminaison LTRs seule, qui diminue l'expression du gène *Vvmyb1A* et participe à l'obtention de la variété « Ruby Okuyama ». Image adaptée de Lisch (2012).

Réarrangements chromosomiques

En plus de générer de nouvelles insertions pouvant causer des maladies, les ETs participent fréquemment à différents types de réarrangements chromosomiques. De nombreux réarrangements attribués aux TIR sont bien décrits dans la littérature. Par exemple, l'élément *Ac* chez le maïs, dont les transpositions aberrantes peuvent entraîner des réarrangements telles que délétions, des inversions ou encore des translocations (Yu et al., 2011). Des réarrangements chromosomiques sont également attribués à des ETs de Classe I. Par exemple, chez la vigne, la recombinaison homologue des terminaisons LTRs d'un rétrotransposon LTR *Gret1* donne lieu à la répression du gène voisin nommé *VvmybA1* (Figure 26). Ce gène est connu pour coder un facteur de transcription qui participe à la synthèse des pigments anthocyanes, qui octroient la pigmentation rouge aux grains de raisin. La répression de *VvmybA1* donne naissance au phénotype de couleur rosé « Ruby Okuyama » (Kobayashi et al., 2004) (Figure 26). Autre phénomène important de remaniement, des recombinaisons inter-chromosomiques peuvent survenir entre des LINE ou des SINE, pouvant conduire à des délétions et/ou des réarrangements chromosomiques (Elbarbary et al., 2016).

Duplication de gènes

Les ETs peuvent également participer au déplacement ou à la duplication de gènes. Par le processus nommé *trans* duplication, un ET qui a incorporé un gène ou un fragment de gène peut réaliser son déplacement et/ou sa duplication vers une nouvelle position du génome hôte. Ce phénomène a été rapporté par exemple chez le maïs, l'arabette ou le riz (Talbert et Chandler, 1988; Yu et al., 2000; Turcotte et al., 2008). Chez ce dernier, les éléments Pack-MULE sont présents à environ 3 000 copies et contiennent des fragments dérivés de plus de 1 000 gènes (Jiang et al., 2004). Un autre processus de duplication de gènes, nommé la rétroposition, peut avoir lieu lorsque les protéines codées par un rétrotransposon LTR réalisent la rétrotranscription de l'ARNm d'un gène voisin. Une copie de ce gène, alors nommé rétro-gène, peut être intégrée dans une nouvelle portion du génome. Ce mécanisme est notamment responsable de la forme arrondie de certaines variétés de tomates (e.g. Roma) (Xiao et al., 2008).

Domestication des ETs

Un ET peut être « domestiqué » par le génome et devenir un gène à part entière lorsque celui-ci participe à l'augmentation du « *fitness* » de l'hôte, c'est-à-dire à l'augmentation de sa valeur sélective (Alzohairy et al., 2013). C'est notamment le cas d'un TIR Mutator qui a été domestiqué au cours de l'évolution du génome de *A. thaliana*. Ce TIR correspond aujourd'hui à un gène nommé *Mustang* qui est impliqué dans le maintien de l'homéostasie hormonale de la plante (Kajihara et al., 2012). L'inactivation de ce gène conduit à une forte diminution du succès reproducteur de l'hôte (Figure 27) (Cowan et al., 2005).

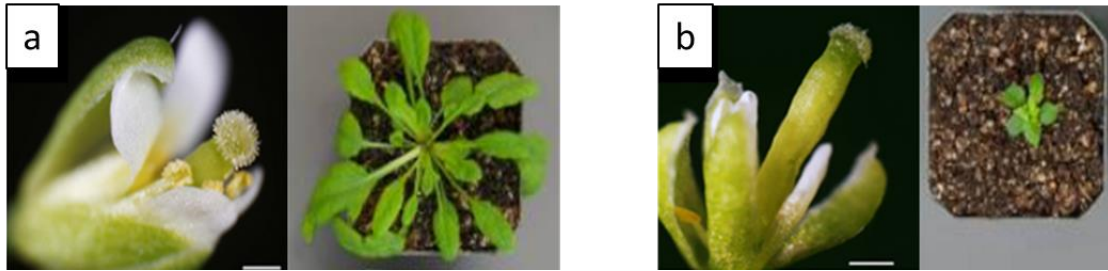


Figure 27 : Impact de l'inactivation de l'ET domestiqué *Mustang* sur la physiologie d'*Arabidopsis thaliana*.

a) Organisme sain avec un gène *Mustang* fonctionnel, b) Organisme mutant avec un gène *Mustang* éteint. Image adaptée de Cowan et al. (2005).

3.7. Les ETs : moteurs de l'évolution adaptative des espèces

Les ETs peuvent participer à l'évolution adaptative des espèces, par la génération de mutations bénéfiques chez un hôte soumis à une pression de sélection. Dans le milieu naturel, les organismes sont impactés par divers stress biotiques ou abiotiques. Ces stress peuvent devenir une pression de sélection lorsqu'ils impactent la survie des individus d'une espèce. L'une des réponses des organismes face à une pression de sélection est l'évolution adaptative (Visser, 2008; Casacuberta et González, 2013). L'évolution adaptative a lieu par la sélection naturelle d'individus présentant de meilleures capacités à survivre et à se reproduire. Ils présentent ainsi une aptitude supérieure à transmettre leurs gènes aux générations suivantes. Le variant génétique permettant cet avantage sélectif augmentera ainsi en fréquence dans la population. B. McClintock fut la première à suggérer que les changements génomiques causés par les ETs peuvent représenter une stratégie évolutive pour l'hôte lors de conditions défavorables (McClintock, 1984). Depuis lors, la capacité des ETs à répondre face à des changements environnementaux est reconnue et a été décrite chez de nombreux organismes (Casacuberta et González, 2013). Cependant, les mécanismes expliquant la susceptibilité des ETs à répondre face à des stress sont encore peu connus.

Dans certains cas, l'activation des ETs est considérée comme une conséquence directe du relâchement des mécanismes de contrôle épigénétique de ces éléments. Par exemple, chez la drosophile, le syndrome de dysgénésie hybride provoque un arrêt des mécanismes du contrôle épigénétique. Celui-ci a pour effet d'induire la mobilité d'ETs qui sont inactifs en condition normale (Petrov et al., 1995).

Dans d'autres cas, des études suggèrent que l'activité de certains ETs est liée à une condition de stress en particulier. Ceci constituerait une réponse spécifique des organismes face aux stress environnementaux. Pour supporter cette idée, des séquences régulatrices sensibles aux hormones ou aux chocs thermiques (HSP, *Heat-Shock Proteins*) ont été identifiées dans des séquences d'ETs, suggérant ainsi une réponse spécifique de ces éléments à des variations environnementales (Feder et Hofmann, 1999). Par exemple chez la microalgue *P. tricornutum*, le rétrotransposon LTR *Surcouf* a été rapporté comme surexprimé lors d'un stress thermique. Les auteurs ont souligné la présence de séquences régulatrices sensibles aux chocs thermiques dans l'une des terminaisons LTRs de cet

élément (Egue et al., 2015). Ainsi, dans certains cas, la surexpression des ETs pourrait correspondre à une stratégie d'adaptation à long terme face à des stress environnementaux. Cette caractéristique transmise à la descendance permettrait de favoriser l'adaptation évolutive des organismes sous une pression de sélection particulière (Capy et al., 2000; Casacuberta et González, 2013).

La création de mutations bénéfiques par les ETs, conduisant à l'adaptation d'individus d'une espèce face à un stress environnemental a été fréquemment rapportée chez les eucaryotes et les procaryotes (Casacuberta et González, 2013). Par exemple, chez le moustique *Culex pipiens*, une mutation associée à un ET a conduit à l'apparition et la sélection d'une souche résistante à une toxine microbienne, utilisée comme larvicide. En effet, un ET s'est inséré dans la région du gène *cpm1*, codant pour la protéine réceptrice de cette toxine. Cette variation génétique a conduit à la création d'un nouvel intron dans le gène *cpm1*. Les individus de cette souche produisent un récepteur non fonctionnel qui est incapable d'interagir avec la toxine, permettant à l'insecte de survivre (Darboux et al., 2007). Dans ce cas de figure, la mutation bénéfique pourra être transmise à la descendance par transfert vertical (Casacuberta et González, 2013).

Des auteurs se sont également interrogés sur le possible rôle de transferts horizontaux d'ETs dans l'adaptation des organismes face à des stress environnementaux (Casacuberta et González, 2013; Gilbert et Feschotte, 2018). Chez les eucaryotes, de nombreux transferts horizontaux de gènes ont permis l'adaptation d'organismes face à des pressions extérieures. Par exemple, le transfert de gènes codants pour des protéines antigèle chez certaines espèces de poissons (Graham et al., 2008). Au niveau des ETs, plusieurs transferts horizontaux ont été rapportés, pouvant parfois participer à des changements évolutifs et impacter le phénotype de l'hôte (Gilbert et Feschotte, 2018). Cependant, à notre connaissance, aucun exemple d'ET acquis par transfert horizontal et ayant activement participé à une adaptation évolutive n'a été découvert. Des études suggèrent cependant que de tels événements pourraient exister (Keeling et Palmer, 2008; Casacuberta et González, 2013).

Chez les procaryotes, quelques exemples suggèrent des transferts d'ETs entre espèces, pouvant ainsi participer à leur adaptation évolutive lors de stress extérieurs. Par exemple, des auteurs se sont intéressés au rôle des IS chez une souche de la bactérie *Cupriavidus metallidurans*, isolée dans un milieu riche en métaux lourds. Ils ont montré que certains de ces IS sont associés à des réarrangements génomiques ayant modifiés la capacité de

croissance autotrophe de cette souche, lui permettant de survivre dans cet environnement chargé en métaux lourds. Parmi les IS impliqués, les auteurs ont identifié des éléments identiques dans les génomes d'autres bactéries isolées dans des environnements similaires, suggérant des événements de transferts horizontaux d'IS entre ces espèces (Mijnendonckx et al., 2011). Ces transferts d'IS auraient ainsi participé à l'adaptation de cette souche de *Cupriavidus metallidurans*, face à un milieu riche en métaux lourds.

3.8. L'activité des ETs chez les microalgues

L'étude des ETs chez les microalgues a débuté en 1988 lorsque Day et al. décrivent une première famille de rétrotransposons (Classe I) chez la Chlorophyte *C. reinhardtii* (Day et al., 1988). Le premier transposon à ADN d'une microalgue sera également décrit chez cette espèce l'année suivante (Ferris, 1989). De nombreuses études se sont succédées sur la caractérisation, l'expression et parfois la mobilité de nouveaux ETs chez cette espèce (Day et Rochaix, 1991; Fan et al., 1995; Graham et al., 1995; Kumekawa et al., 1999; Wang et al., 1998; Lefebvre et Silflow, 1999; Perez-Alegre et al., 2005; Kojima, 2005; Kim, 2006; Merchant et al., 2007). Chez l'espèce multicellulaire *V. carterii*, des études similaires ont été entreprises. Ceci pour identifier des ET actifs et mobiles pouvant être utilisés pour cibler et cloner des gènes de cette espèce, à travers une méthode nommée « *transposons tagging* » (Figure 28) (Ferris et al., 1996; Kirk et al., 1999; Miller et Kirk, 1999; Nishii et al., 2003). Pour cette méthode, l'induction de la mobilité des TIR *Jordan* et *Itaden* avec des stress thermiques a été exploitée afin de créer des banques de mutants. Cette méthode a permis d'identifier plusieurs gènes participant au développement de la cellule (Figure 28) (Kirk et al., 1999; Miller et Kirk, 1999; Nishii et al., 2003). L'augmentation du taux de transcription du LINE *Zepp* a été aussi rapportée sous une forte température chez l'espèce *Chlorella vulgaris* (Higashiyama et al., 1995a, 1995b, 1997). Cette famille de LINE a été particulièrement explorée chez cette Chlorophyte en raison de son implication dans la formation et l'entretien des télomères chromosomiques de cette espèce (Higashiyama et al., 1997; Noutoshi et al., 1998; Yamamoto, 2003, 2003).

Plus récemment, chez *C. reinhardtii*, une étude a montré que l'insertion d'un MITE dans le génome de cette espèce, a conduit à l'évolution d'un clone devenu dépendant de la vitamine B₁₂, alors que cette espèce est indépendante à cette vitamine. Tandis que cette mutation était réversible pour certains clones évolués, la réversion du transposon a laissé une empreinte chez un clone, conduisant à sa dépendance à la vitamine B₁₂ (Figure 29) (Helliwell et al., 2015).

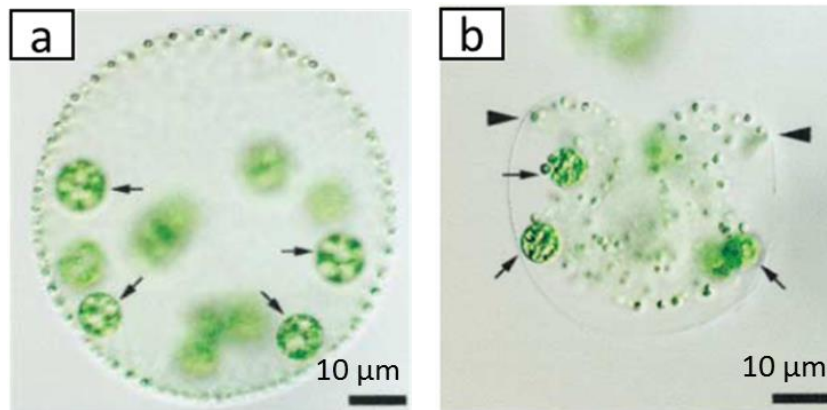


Figure 28 : Photographie par microscopie illustrant l'impact de l'insertion d'un ET dans le gène *InvA* de la microalgue multicellulaire *V. carteri*.

Le gène *InvA* participe au développement de la microalgue multicellulaire *V. carteri*. a) Organisme sain, b) Organisme mutant ayant subi une l'insertion d'un ET dans le gène *InvA*. Photographies provenant de Nishii et al. (2003).

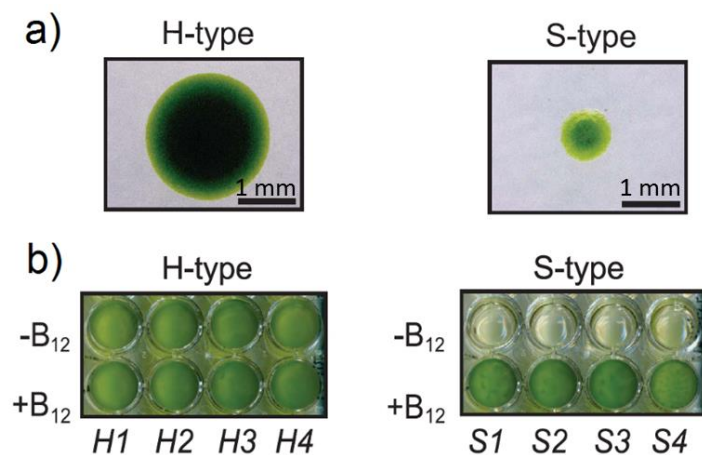


Figure 29 : Impact de l'insertion d'un ET dans le gène *MetE* dans le génome d'une souche clonale de *C. reinhardtii*.

a) Cellules de *C. reinhardtii* cultivées sur du milieu solide sans vitamine B₁₂. Le type H correspond à une cellule indépendante à la vitamine B₁₂ et la cellule de type S (plus petite), correspond à une cellule devenue dépendante à la vitamine B₁₂ après l'insertion d'un ET dans le gène codant pour la méthionine synthase (*MetE*). b) Croissance après 72h de quatre colonies de type H et type S avec ou sans vitamine B₁₂. Une très faible croissance est obtenue pour les souches type S (dépendantes à la vitamine B₁₂) dans un milieu sans vitamine B₁₂. Photographies provenant de Helliwell et al. (2015).

Ensuite, chez la microalgue *O. tauri*, l'implication d'un rétro-transposon a également été suggérée dans la résistance de certaines souches face à un virus (Derelle et al., 2018). Les auteurs ont identifié la surexpression d'une rétro-transcriptase chez des cellules infectées par rapport à des cellules saines. Ils suggèrent que cette surexpression pourrait être une réponse *O. tauri* face à l'attaque du virus. En impactant l'expression de gènes voisins et/ou en enclenchant sa transposition, l'ET pourrait potentiellement participer à la mise en place d'une résistance face à l'attaque virale (Derelle et al., 2018).

Chez les Diatomées, l'expression d'ETs a été étudiée à la suite du séquençage et de l'annotation des génomes des espèces *Thalassiosira pseudonana* et *Phaeodactylum tricornutum* (Armbrust, 2004; Bowler et al., 2008). L'expression de différents ETs de Classes I et Classes II a ainsi été mise en évidence chez ces espèces lors de différents stress environnementaux (Maumus et al., 2009; Norden-Krichmar et al., 2011). Chez *P. tricornutum*, les rétrotransposons LTR *Blackbeard* et *Surcouf* sont surexprimés respectivement lors d'une limitation en azote ou suite à une exposition au 2,4-décadiénal (Maumus et al., 2009). Cette étude a aussi indiqué que la déméthylation de l'ADN entraînait l'activation du LTR *Blackbeard* (Maumus et al., 2009). Une autre étude suggère que la surexpression du LTR *Blackbeard* pourrait aussi être corrélée à la densité cellulaire (Oliver et al., 2010).

Chez *P. tricornutum*, d'autres études se sont également intéressées à l'expression des rétrotransposons LTR *Blackbeard* et *Surcouf*, mais en condition de stress thermique. Ces études ont mis en évidence la surexpression du rétrotransposons LTR *Surcouf* sous différentes températures (Egue et al., 2015; Hermann, 2011). L'expression de TIR Mariner a aussi pu être démontrée sous un stress thermique chez les Diatomées *Amphora acutiuscula* et *Amphora coffeaeformis* (Hermann, 2011; Nguyen, 2014). Chez l'espèce *Thalassiosira oceanica*, la comparaison de données transcriptomiques a permis de souligner la surexpression d'une rétro-transcriptase en condition de limitation en Fer (Lommer et al., 2012). Récemment, une étude a rapporté la surexpression d'une polyprotéine associé à un plausible rétrotransposon lors de la mise en contact de la Diatomée *Skeletonema marinoi* avec des prédateurs (Amato et al., 2018).

Chez les Dinophytes, le séquençage du génome et le transcriptome de *Alexandrium catenella* et *Alexandrium ostenfeldii* ont permis de mettre en évidence les premières preuves de l'activité de rétrotransposons (Uribe et al., 2008; Jaekisch et al., 2011). La

surexpression de plusieurs familles de rétrotransposons LTR a récemment été rapporté pour l'espèce *Symbiodinium microadriaticum* soumise à une température élevée (Chen et al., 2017a).

Enfin, une étude s'est penchée sur l'identification de l'expression des rétrotranscriptase d'ETs de Classe I dans le milieu marin. Cette étude a été réalisée à partir de données méta-transcriptomiques, issues d'échantillons d'organismes planctoniques prélevées à différentes régions de la méditerranée. Elle a mise en évidence l'expression de rétrotranscriptases chez de nombreux taxons composés de microalgues, dont ceux des Straménopiles, Alvéolées, Cryptophytes et Haptophytes (Lescot et al., 2016). En revanche et à notre connaissance, aucune étude ne s'est particulièrement intéressée à l'analyse de l'expression des ETs chez les autres taxons de microalgues. En effet, la plupart des études se sont intéressées à la description ou à l'annotation de ces éléments comme par exemple chez les Cryptophytes ou les Haptophytes (Khan et al., 2007; Read et al., 2013).

Contexte de l'étude

1. Les microalgues : ressource d'avenir pour les biotechnologies

Les microalgues constituent de par leur incroyable diversité une formidable source de composés d'intérêt, tels que des lipides ou des enzymes, qui sont parfois uniques en leur genre. De par cela, ces microorganismes sont considérés comme une ressource d'avenir et sont de plus en plus exploités par l'homme dans différents domaines biotechnologiques (Milledge, 2011). Les principales applications utilisant des microalgues sont présentées ci-dessous.

1.1. Alimentation

Tandis que l'utilisation des Cyanobactéries comme ressource alimentaire par l'être humain remonte à des milliers d'années (Milledge, 2011), l'intérêt porté sur les microalgues est plus récent. Les études ont réellement commencé il y a 60 ans. Tout d'abord dans le cadre de l'alimentation animale et humaine, en raison de leur composition en lipides, protéines, vitamines et leurs propriétés anti-oxydantes (Cadoret et al., 2012). Historiquement, les microalgues ont été tout d'abord utilisées en aquaculture pour la nutrition des bivalves, des crevettes et des poissons (Muller-Feuga, 2000; Hemaiswarya et al., 2011). Ce n'est que plus récemment, que leur potentiel en alimentation humaine est étudié. En France, neuf espèce de microalgues sont autorisées à la commercialisation pour l'alimentation humaine (principalement sous forme de compléments alimentaires) : les Chlorophytes *Dunaliella salina*, *Chlorella vulgaris*, *Chlorella sorokiniana*, *Scenedesmus vacuolatus*, *Haematococcus pluvialis*, *Parachlorella kessleri*, *Auxenochlorella protothecoides*, *Auxenochlorella pyrenoidosa*, ainsi que la Diatomée *Odontella aurita* (ANSES, Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail, Saisine n° 2017- SA-0086). Bien qu'il existe un certain engouement de la part des consommateurs, le marché reste encore anecdotique en France, hormis la culture de la Cyanobactérie *Arthrospira* nommée plus communément « spiruline » dont les unités de production ne cessent de croître. Actuellement, les microalgues sont principalement distribuées sous forme de compléments alimentaires, en raison de leur richesse en acides aminés essentiels,

vitamines et acides gras polyinsaturés *omega*-3 (Hemaiswarya et al., 2011; García et al., 2017). Dans l'agroalimentaire, elles sont aussi utilisées comme colorant naturel (Bêta-carotènes E160aiv).

1.2. Remédiation

Les microalgues commencent à être utilisées dans de petites stations d'épuration pour remédier des pollutions et contaminants des eaux usées (Martínez et al., 2000; Renuka et al., 2015). Il est également envisagé de les utiliser pour traiter les eaux d'effluents provenant des élevages (Godos et al., 2009) ou celles contaminées par des métaux lourds. Enfin, elles sont étudiées pour la synthèse de bioplastiques, par exemple à partir de l'extraction de polymères de protéines (Zeller et al., 2013). Leur utilisation pourrait permettre de diminuer la quantité de plastique actuellement élaboré à base de pétrole (Rasul et al., 2017).

1.3. Energie

Dans les années 2000, les microalgues ont suscité un engouement particulier comme matière première pour la production de biocarburant de troisième génération (Milano et al., 2016). L'utilisation de biocarburant à partir de microalgues permettrait l'utilisation d'une énergie renouvelable, biodégradable, non toxique et participerait à la diminution des émissions de gaz à effet de serre (Brune et al., 2009; Rasul et al., 2017). L'exploitation des microalgues est également envisagée pour la production de bio-hydrogène. Soit en utilisant leur capacité de production d'hydrogène à travers la photosynthèse, soit par fermentation anaérobie en les utilisant comme substrat (Nagarajan et al., 2017). Cependant, ces engouements semblent à l'heure actuelle retombés car les systèmes et méthodes de production de microalgues à grande échelle sont encore loin d'être économiquement viables face au prix du pétrole (Islam et al., 2017).

1.4. Molécules actives

Les microalgues sont aussi examinées pour leur contenu en métabolites secondaires. En effet, elles représentent un réservoir potentiel de molécules bioactives encore inconnues. Des composés tels que leurs acides gras, pigments ou leurs polysaccharides sont étudiés pour leurs propriétés anti-inflammatoires, anti-oxydantes, antivirales ou antimicrobiennes (Mimouni et al., 2015; Rasul et al., 2017). Ces premiers résultats semblent prometteurs mais l'efficacité de ces molécules reste encore à être évaluée cliniquement. Ainsi, les microalgues suscitent un fort intérêt avec pour espoir d'identifier de nouvelles molécules d'intérêt en santé.

1.5. Des verrous à lever pour l'utilisation des microalgues

Physiques : Production de biomasse à grande échelle

A l'heure actuelle, la filière microalgue est pleine de promesses et de potentialités. Cependant, de nombreuses limites économiques, techniques et biologiques restent encore à être levées pour que leur utilisation devienne économiquement viable à grande échelle (Vigani et al., 2015). Par exemple, pour l'alimentation humaine ou animale, les principales limites techniques pointées par les spécialistes sont la production de biomasse à grande échelle et surtout, la séparation physique des microalgues de l'eau (Figure 30).

Biologiques : Obtention de souches domestiquées

De nombreuses études portant sur la conception de systèmes de cultures innovants pour obtenir une meilleure productivité ont été réalisées. Elles ont permis d'obtenir des gains substantiels en terme de productivité massique et/ou de productivité pour des métabolites d'intérêt (e.g. lipides, pigments, sucres). Cependant, les études réalisées utilisent principalement des souches sauvages (souches provenant de banques et/ou isolées dans l'environnement). L'obtention et l'utilisation de souches domestiquées présentant des caractères phénotypiques d'intérêt représentent ainsi un autre levier important pour améliorer la productivité (Figure 30) (Vigani et al., 2015). A l'image des progrès de l'agriculture engendrés par la domestication des plantes et des animaux, pour devenir économiquement rentable, l'exploitation des microalgues requiert la domestication de souches sauvages (Larkum et al., 2012; Vigani et al., 2015).

**Nombre d'experts
ayant indiqué ces facteurs (%)**

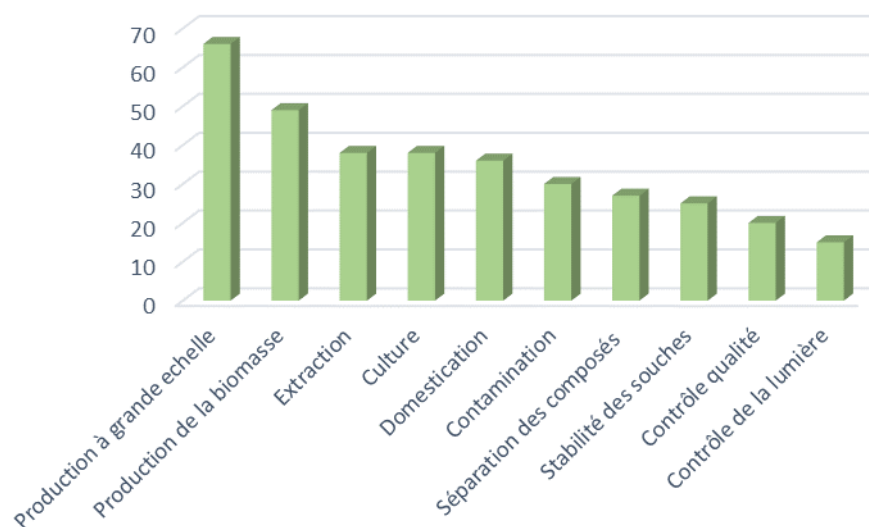


Figure 30 : Facteurs les plus influents sur le coût de production d'aliments à base de microalgues.

Image adaptée de Vigani et al. (2015).

2. La domestication des microalgues

La domestication est l'action ou le processus réalisé par l'homme sur un organisme afin de s'en approprier un bénéfice pour répondre à un besoin. La domestication est très ancienne, elle a commencé chez les plantes lors de la mise en place de l'agriculture il y a plus 10 000 ans (Meyer et Purugganan, 2013; Larson et al., 2014). Les premiers processus de domestication étaient très simples et avaient pour base la sélection massale. Cette approche consiste à sélectionner les meilleurs individus pour la prochaine génération. Au fil des siècles, les processus de domestication sont devenus de plus en plus élaborés « jusqu'à la révolution verte » qui débute dans les années 1940 (e.g. croisements contrôlés, généalogie, hybrides) (Meyer et Purugganan, 2013). A titre d'exemple, en 50 ans (1950-2000), le rendement a doublé pour la culture du colza, triplé pour les betteraves, quadruplé pour le maïs (Doré et Varoquaux, 2006).

Chez les microalgues, la domestication ou l'amélioration de ces espèces a débuté il y a une vingtaine d'années à peine. Actuellement, les souches de microalgues exploitées proviennent directement du milieu naturel et sont des souches dites « sauvages ».

L'amélioration des microalgues a été mise en œuvre pour différentes finalités, à partir d'une large variété d'espèces. Le choix de l'espèce de microalgues à améliorer dépendra de la molécule ou du caractère recherché. Plusieurs espèces appartenant à différents taxons ont fait l'objet de programmes d'amélioration. Les premières souches domestiquées présentent une amélioration de productivité allant de 30 à 400%. Pour une grande partie d'entre elles, les études se sont centrées sur la possibilité d'augmenter la teneur ou la productivité en lipides pour la production de biocarburant (Alonso et al., 1994; Grima et al., 1995; Meireles et al., 2003; Chaturvedi et Fujita, 2006; Lian et al., 2010; Iskandarov et al., 2010; Bougaran et al., 2012; Doan et Obbard, 2012; Vigeolas et al., 2012; Ota et al., 2013a; Yang et al., 2013; Zayadan et al., 2013; Jaeger et al., 2014; Terashima et al., 2015). D'autres études ont permis d'accroître la production de pigments, notamment de lutéine chez *Chlorella sorokiniana* (Cordero et al., 2011), d'astaxanthine chez *Haematococcus pluvialis* (Chen et al., 2003), de beta-carotène chez *Dunaliella bardawil* (Shaish et al., 1991). Enfin, quelques recherches se sont focalisées sur l'amélioration de souches présentant de meilleures capacités de culture, par exemple une optimisation de la niche thermique pour *T. lutea* (Bonfond et al., 2017).

L'obtention de souches domestiquées est réalisée à travers la mise en place de programmes d'amélioration. Ces programmes d'amélioration profitent des méthodologies et des outils développés chez les autres organismes. Cependant les connaissances biologiques ne permettent pas certaines approches. En effet, chez la plupart des espèces de microalgues, les croisements ne sont pas possibles car leurs cycles biologiques ne sont pas encore connus. Actuellement, quatre différentes stratégies d'amélioration ont été appliquées. Elles peuvent se distinguer de la manière suivante : l'amélioration par 1) démarche de pression de sélection ou « approche d'évolution dirigée », 2) crible des individus ou « approche de sélection », 3) augmentation de la variabilité génétique suivie d'un crible d'individus ou « approche de mutation-sélection » et 4) transformation ou « approche de génie génétique ».

Amélioration par « génie génétique »

L'approche par génie génétique est certainement la plus controversée pour des microorganismes marins cultivés majoritairement dans des bassins en extérieur. Cependant, cette stratégie a été utilisée pour améliorer des souches de microalgues (Xue et al., 2015; Yang et al., 2016).

Amélioration par « évolution dirigée »

Des études utilisent la capacité d'adaptation des microalgues pour faire apparaître des phénotypes adaptés face à une pression de sélection. La sélection phénotypique est réalisée à travers la croissance. Les cellules sont maintenues dans un état de croissance active, de sorte que les variants adaptatifs, avec une vitesse de division plus rapide, puissent se fixer dans la population (Krasovec et al., 2017). Par exemple, ce type d'expérience a permis l'obtention d'une souche de *C. reinhardtii* présentant un taux de croissance amélioré dans un milieu nutritif riche en acétate (Perrineau et al., 2014) ou une souche de *T. lutea* adaptée à une niche thermique plus large que la souche d'origine (Bonnefond et al., 2017).

Amélioration par « sélection »

S'appuyant sur les méthodes d'amélioration des plantes terrestres, qui consistent à sélectionner des espèces ou écotypes performants parmi une diversité d'individus, certains programmes d'amélioration des microalgues utilisent une démarche de « sélection » d'individus qui présentent un phénotype d'intérêt. Plusieurs protocoles différents de sélection phénotypique peuvent être employés pour les microalgues. Ainsi, des études utilisent des méthodes de criblage d'individus d'intérêt. Par exemple, par sélection d'individus sur milieu gélosé sur boîtes de Pétri en présence ou non d'un inhibiteur métabolique (Chen et al., 2003; Van Wageningen et al., 2014). La sélection d'individus qui présentent des traits d'intérêt peut également être réalisée manuellement au travers d'approches de dilution limite ou de micromanipulations (Grima et al., 1995). De même, l'utilisation de la cytométrie en flux a également été proposée par Mendoza (2008) pour la sélection d'une souche de *Dunaliella salina* hyper-productrice de caroténoïdes. Couplée à l'emploi d'un fluorochrome comme le Rouge du Nil et à un dispositif de tri, cette technique permet de sélectionner à très haute fréquence des cellules qui présentent un phénotype intéressant (Mendoza et al., 2008; Bougaran et al., 2012; Rumin et al., 2015). Une autre stratégie est l'utilisation de dispositifs microfluidiques permettant l'isolement de cellules individuelles et donc le criblage de phénotypes d'intérêt (Yousuff et al., 2017).

Amélioration par « mutation-sélection »

Dans le but d'augmenter la variabilité génétique dans une souche de microalgues, une étape de « mutation » peut être réalisée en amont de l'étape de « sélection ».

Lors de l'étape de mutation, sont distingués les mutagènes chimiques et physiques. Parmi les mutagènes physiques, les radiations ultraviolet de courte longueur d'onde (UV-c), sont les plus fréquemment employées (Shaish et al., 1991; Bougaran et al., 2012), bien que certaines études aient également fait appel à des radiations UV-a de plus faible énergie (Zayadan et al., 2014). D'autres techniques font appel au bombardement par ions lourds (Ma et al., 2013; Ota et al., 2013b) ou aux jets de plasma (Zhang et al., 2014). Des mutagènes chimiques sont aussi utilisés pour la mutagenèse des microalgues. Dans la littérature, les agents les plus fréquemment utilisés sont le méthanesulfonate d'éthyle (EMS) (Chaturvedi et Fujita, 2006) et le méthylnitronitrosoguanidine (MNNG) (Cordero et al.,

2011). Quelques études réalisent plusieurs cycles successifs de mutagenèse, faisant appel à plusieurs agents mutagènes différents (chimique ou physique) (Chen et al., 2003; Lian et al., 2010). La mise en œuvre de ces techniques de mutagenèse conduit à des résultats d'amélioration phénotypique variables. Selon les travaux, des gains de quelques pourcents seulement sont obtenus, alors que d'autres parviennent à des améliorations très significatives (400%) du caractère recherché (Chen et al., 2003; Cordero et al., 2011; Bougaran et al., 2012).

De manière générale, l'impact de ces différents programmes d'amélioration sur le génome des souches et la stabilité génétique et phénotypique des caractères acquis ont été peu étudiés (Perrineau et al., 2014; Carrier et al., 2018).

3. Présentation de l'espèce d'étude : *Tisochrysis lutea*

Le laboratoire de Physiologie et Biotechnologie des Algues utilise principalement comme modèle d'étude l'Haptophyte *Tisochrysis lutea* que ce soit au niveau écophysiologie ou moléculaire. La microalgue *T. lutea* (Figure 31) appartient au phylum des Haptophytes, à l'ordre des Isochrysidales et la famille des Isochrysidacées (Figure 32) (Bendif et al., 2013). La première souche de cette espèce a été découverte à partir d'un échantillon collecté en Polynésie française dans un bassin de crevettes par Haine en 1977. Cette souche est nommée CCAP927/14 et conservée à la *Culture Collection of Algae and Protozoa* (CCAP) situé en Ecosse. De nombreuses autres souches ont ensuite été identifiées dans différentes régions du globe éloignées géographiquement (Figure 33) (Efimova et al., 2016). En raison de sa facilité de culture et de son fort taux de croissance, cette espèce est historiquement utilisée en aquaculture afin de nourrir les larves et juvéniles de mollusques et de crustacés (Guedes et Malcata, 2012). Plus récemment, cette microalgue est étudiée pour son intérêt en alimentation animale et humaine car elle possède un profil lipidique riche en acides gras polyinsaturés à longue chaîne de type *omega*-3, comme le DHA (acide docosahexaénoïque) et l'EPA (acide eicosapentaénoïque) (Ryckebosch et al., 2014). Son profil pigmentaire est également intéressant car ses chloroplastes contiennent des pigments caroténoïdes tels que la fucoxanthine aux propriétés antioxydantes (Delbrut et al., 2018). Enfin, la teneur et la composition en lipides neutres de *T. lutea* font que cette espèce a été l'objet de plusieurs études pour son utilisation dans la production de biocarburant de troisième génération (Bougaran et al., 2012; Driver et al., 2014).

Dans la littérature, les caractéristiques biologiques de cette espèce ont été analysées au niveau éco-physiologique (Saoudi-Helis et al., 1994; Bougaran et al., 2012; Lacour et al., 2012; Marchetti et al., 2013, 2018) et biochimique (Saoudi-Helis et al., 1994; Custódio et al., 2014; Alkhamis et Qin, 2015; Delbrut et al., 2018). De même, de nombreuses études « omiques » commencent à s'accumuler à son sujet (Carrier et al., 2018; Charrier et al., 2015; Garnier et al., 2014; Hernández Javier et al., 2018; Thiriet-Rupert et al., 2016, 2018) et la première méthode de transformation génétique a récemment été développée (Endo et al., 2018). *T. lutea* a également fait l'objet de différents programmes de domestication qui ont abouti à l'obtention d'un large panel de souches domestiquées (Bougaran et al., 2012; Carrier et al., 2018).

Récemment, le premier assemblage du génome de *T. lutea* a été publié (Carrier et al., 2018). Pour cela, les génomes d'une souche sauvage et de deux souches domestiquées ont été séquencés. Ces souches domestiquées ont été obtenues par un programme d'amélioration de type « mutation-sélection ». La comparaison des génomes de ces souches a abouti à la prédiction de polymorphismes nucléotidiques (SNPs) et d'événements d'insertions et de délétions de séquences nucléotidiques (indels). Parmi ces variations génétiques spécifiques, les auteurs ont identifié la contribution de larges indels présentant des similarités à des séquences d'ETs. A la suite de ces observations, les auteurs ont suggéré la présence probable d'ETs actifs chez *T. lutea*.

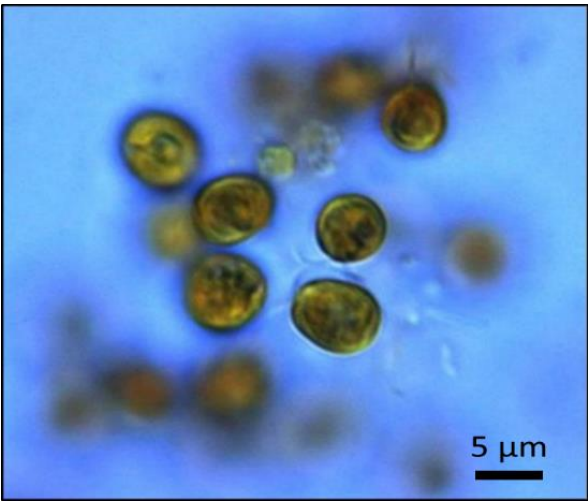


Figure 31 : Photographie au microscope de plusieurs cellules de *T. lutea*.

Photographie de R. Kaas (PBA, Ifremer).

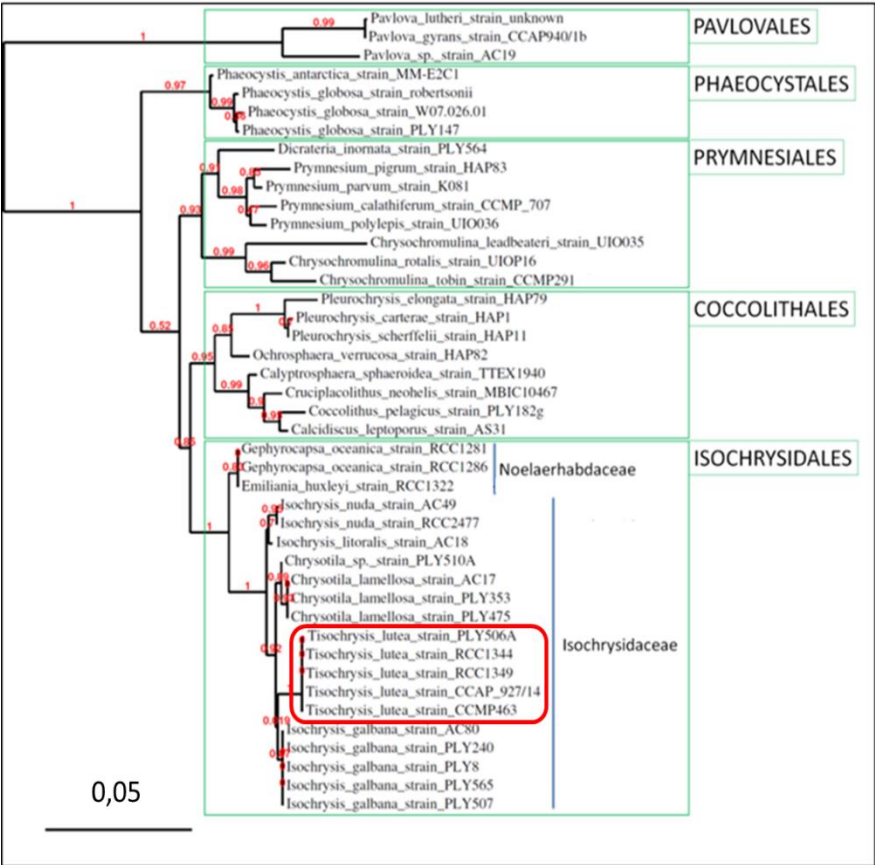


Figure 32 : Position phylogénique de l'espèce *T. lutea* chez les Haptophytes.

Construction phylogénétique basée sur l'étude de l'ARNr 18S. Image provenant de Carrier et al. (2018).

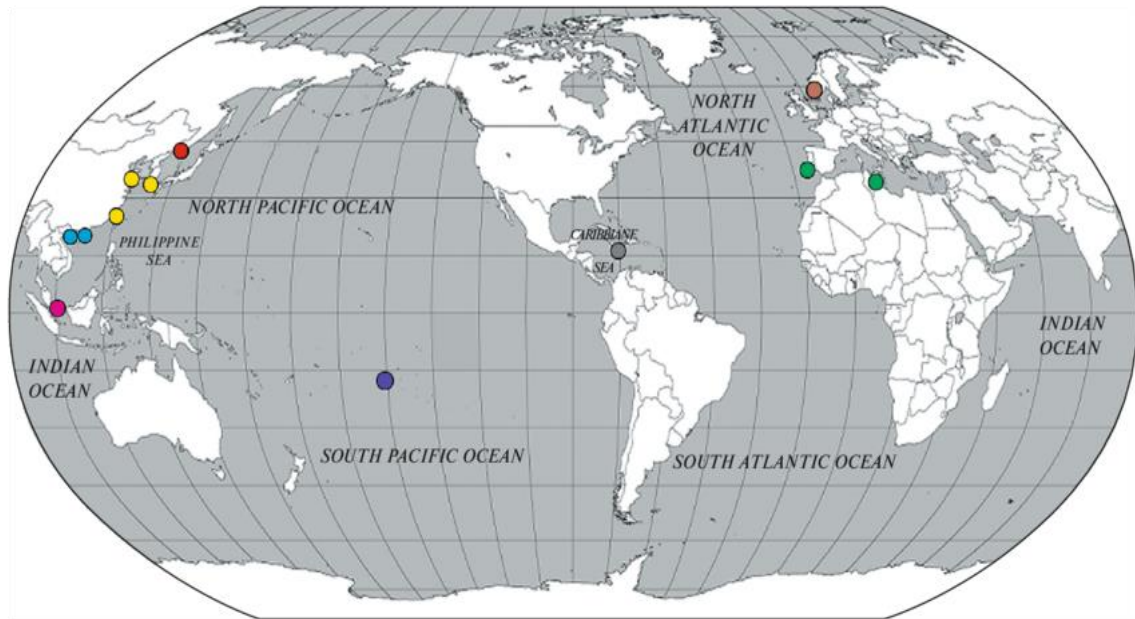


Figure 33 : Distribution géographique des différentes souches isolées de *T. lutea*.

Les points de différentes couleurs représentent les régions où les souches de *T. lutea* ont été isolées : Rouge : Mer du Japon, violet : région de Tahiti, jaune : Mer Jaune et Mer de Chine orientale, bleu : Mer de Chine méridionale, rose : Mer de Chine méridionale des côtes du Vietnam, vert : Région méditerranéenne, marron : Mer du Nord, gris : Mer des Caraïbes. Image provenant de Efimova et al. (2015).

Objectifs de la thèse

Dans ce contexte d'obtention de souches domestiquées de microalgues, la stabilité et l'origine moléculaire des caractères améliorés n'ont que très peu été abordés. Une première étude sur l'impact des programmes d'amélioration sur le génome de *T. lutea* a mis en évidence une part non négligeable de larges indels (Carrier et al. 2018). Leurs origines pourraient être associées à une possible activité des ETs. Les ETs sont connus pour être des acteurs importants de l'évolution des génomes et jouent un rôle majeur sur la variation des traits phénotypiques d'intérêt (Lisch, 2012).

L'objectif de ce travail est d'initier les premières études sur la dynamique du génome de l'Haptophyte *T. lutea* à travers l'étude des ETs. Plus précisément, l'objectif consiste à 1) annoter exhaustivement les ETs chez *T. lutea* et étudier leur activité, 2) analyser leur dynamique au sein de deux souches domestiquées de cette espèce.

Ainsi, les objectifs de cette thèse peuvent se traduire par trois questions principales :

- (1) Quelle est la diversité des ETs chez *T. lutea* ?
- (2) Quelles sont les familles potentiellement actives chez cette espèce ?
- (3) Quelle est l'importance de la dynamique des ETs chez deux souches domestiquées de *T. lutea* ?

Cette étude apporte les premières connaissances concernant la diversité, l'activité et la dynamique des ETs chez une Haptophyte et contribue à l'acquisition de nouvelles informations sur l'importance des ETs dans la dynamique du génome des microalgues.

Ce travail de thèse se décline sous la forme de deux chapitres :

Chapitre 1 : Annotation et étude de l'activité des ETs de *T. lutea*

Chapitre 2 : Etude de la dynamique des ETs chez deux souches domestiquées de *T. lutea*

Chapitre 1

Annotation et étude de l'activité des ETs de *T. lutea*

L'objectif du premier chapitre de cette thèse consiste à annoter les ETs présents dans le génome de *T. lutea* et plus particulièrement d'identifier des ETs potentiellement autonomes et actifs. Pour faciliter et améliorer cette annotation, un nouvel assemblage du génome de *T. lutea* a été réalisé. Afin d'obtenir une annotation des ETs la plus exhaustive possible, un nouveau pipeline « PiRATE » a été développé. La détection, la classification et l'annotation des ETs de *T. lutea* ont ainsi pu être achevées. Parmi les ETs qui ont été identifiés comme potentiellement autonomes, certains ont été validés comme potentiellement actifs en s'appuyant sur l'analyse de données transcriptomiques et protéomiques disponibles. Tous ces résultats ont permis la rédaction et la publication d'un article dans la revue *BMC Genomics*, présenté dans le paragraphe 4 de ce chapitre. Enfin, des résultats complémentaires à cet article, ont également été décrits.

1. Un nouvel assemblage du génome de *T. lutea*

1.1. Contexte de l'étude

Au début de cette thèse, une première version de l'assemblage du génome de *T. lutea* était disponible (Carrier et al., 2018). Celle-ci avait été obtenue à partir de données produites par la technologie de seconde génération Illumina HiSeq (*paired-end*, 100 pb). Comme vu précédemment, cette technologie permet d'amplifier et de séquencer de courtes séquences d'ADN nommées *short reads*. Ces courtes séquences permettent de réaliser des assemblages de génomes mais ceux-ci sont très souvent fragmentés en plusieurs milliers de *contigs* (Alkan et al., 2011). Cette fragmentation résulte en partie d'un assemblage imparfait de ces *short reads*, notamment en ce qui concerne les régions répétées des génomes (Treangen et Salzberg, 2011). Ces *short reads* ne permettent généralement pas de couvrir entièrement les longues régions répétées des génomes tels que les ETs, conduisant à la création de *contigs* chimériques (Figure 34). La première version du génome de *T. lutea* est composée de 7 659 *contigs* (N50=10 571 pb), pour une taille estimée à 54 Mpb. Celle-ci est particulièrement fragmentée, et de ce fait, ne permet pas l'annotation correcte des différentes copies d'ETs présentes dans le génome de *T. lutea*. Par conséquent, une nouvelle version du génome de *T. lutea* a été initiée au cours du présent travail.

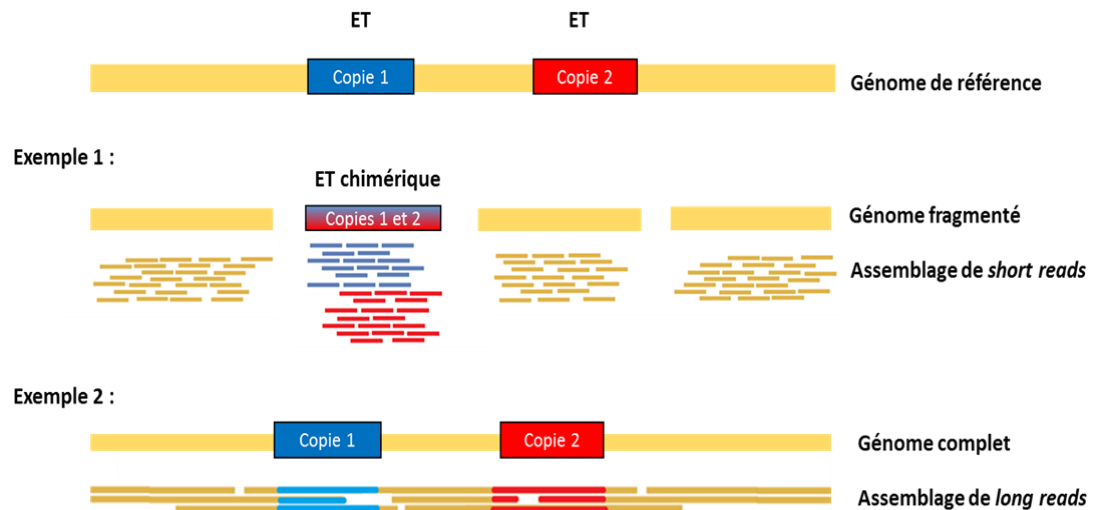


Figure 34 : Illustration schématique de l'assemblage d'une région d'un génome présentant deux copies d'un ET à l'aide de données *short reads* ou *long reads*.

L'exemple 1 illustre la création d'un *contig* chimérique à partir de l'assemblage de *short reads* issues de deux copies d'un ET. L'exemple 2 illustre l'utilisation de *long reads* permettant de dissocier et assembler correctement les deux copies d'un ET.

1.2. Matériel génétique

La souche CCAP927/14 de l'espèce *T. lutea* a été utilisée comme matériel biologique pour l'extraction de l'ADN total. Cette souche est communément utilisée par la communauté scientifique et a ainsi été choisie pour le premier séquençage du génome (Marchetti et al., 2012; Bougaran et al., 2012; Liu et al., 2013; Carrier et al., 2014; Bonnefond et al., 2017; Hernández Javier et al., 2018). Cette souche a été axénisée à l'aide d'antibiotiques. Comme la plupart des souches de microalgue, cette souche n'est pas mono-clonale. Elle se caractérise par une variabilité génétique intra-populationnelle qui est induite par la présence conjointe d'individus dont le nombre reste à déterminer précisément (plus de 200 haplotypes détectés, Saint-Vanne et al., 2017 mémoire de Master 2). Cette variabilité génétique intra-populationnelle complexifie la démarche envisagée pour l'assemblage d'un génome de référence. Cependant, le temps imparti à la thèse ne nous a pas permis d'envisager le séquençage du génome d'une souche clonale. Un isolement d'une cellule unique à partir de la souche CCAP927/14 de *T. lutea* est possible mais l'obtention de la biomasse suffisante pour un séquençage aurait demandé trois mois de culture.

1.3. Stratégie de séquençage

Afin d'obtenir un assemblage moins fragmenté du génome de *T. lutea*, un nouveau séquençage du génome a été entrepris à l'aide de la technologie de 3^{ème} génération Pacifique Biosciences. Cette technologie permet de séquencer directement des molécules d'ADN de grandes tailles, qui peuvent ainsi couvrir de longues régions répétées du génome. A partir de ces *long reads*, les outils d'assemblage peuvent plus facilement reconstruire les régions répétées des génomes et permettent ainsi d'obtenir un assemblage de génome peu fragmenté (McCoy et al., 2014) (Figure 34). A partir de sept séries de séquençages réalisées par un Pacific Biosciences RSII, 735 115 *long reads* ont été produites présentant une taille moyenne de 5 000 pb. Une profondeur moyenne de séquençage de 58,5X a été obtenue pour l'ensemble du génome (Figure 35).

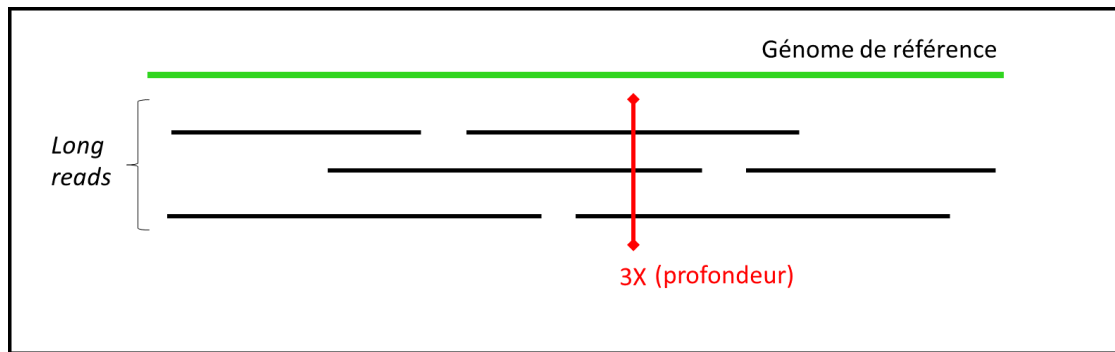


Figure 35 : Schéma illustrant la notion de profondeur de séquençage d'une région d'un génome.

1.4. Assemblage du génome de *T. lutea*

La nouvelle version de l'assemblage du génome de *T. lutea* a été réalisée en partenariat avec l'IRHS (Institut de Recherche en Horticulture et Semences, INRA) d'Angers qui a récemment réalisé l'assemblage du génome du pommier (Daccord et al., 2017). A partir des *long reads* obtenues par la technologie Pacifique Bioscience, un premier assemblage du génome a été réalisé en utilisant le programme Canu (Koren et al., 2017). L'outil Canu est divisé en trois étapes distinctes : la correction, la réduction et l'assemblage des *long reads* (Koren et al., 2017). L'étape de correction sélectionne les meilleurs chevauchements des *long reads* et génère automatiquement des séquences corrigées. L'étape de réduction permet d'enlever les régions de séquences de mauvaise qualité ainsi que les adaptateurs issus de l'étape de préparation des bibliothèques. Enfin, la phase d'assemblage engendre des séquences consensus qui sont ensuite assemblées sous forme de *contigs*. L'outil Canu a permis d'obtenir une première version de l'assemblage du génome de *T. lutea*. Celui-ci est constitué de 260 *contigs*.

En raison du taux élevé d'erreur par base des *long reads* séquencées par la technologie Pacifique Bioscience (10%), deux étapes supplémentaires de correction ont été appliquées. Une première étape de correction exploitant la redondance de séquençages de chaque base (58,5X de profondeur dans notre cas) a été réalisée à partir des *long reads*. Pour cela, les *long reads* ont été préalablement alignées avec le programme bwa (Li et Durbin, 2009) sur l'assemblage du génome obtenu par Canu. Cet alignement a ensuite été soumis à l'outil Quiver (Chin et al., 2013). Celui-ci a permis d'améliorer l'assemblage (Tableau 2). Une seconde étape de correction a été réalisée à partir des *shorts reads* précédemment produites à l'aide de la technologie Illumina. Cette méthodologie combinant les deux technologies de séquençage est en effet recommandée dans la communauté scientifique pour significativement améliorer la qualité de l'assemblage d'un génome (Phillippy, 2017; Tyson et al., 2018). L'outil Pilon (Walker et al., 2014) a été appliqué sur l'assemblage du génome obtenu à la suite de Quiver et a permis d'apporter des corrections (Tableau 2).

Pour finir, une étape de vérification manuelle a été réalisée afin d'identifier d'éventuels *contigs* similaires (identité de 98%). Ces *contigs* similaires n'ont pu être fusionnés durant l'assemblage à cause d'un polymorphisme allélique trop important entre les différents clones séquencés de la souche. Par cette approche, 67 *contigs* identifiés ont été retirés.

Tableau 2 : Aperçu des corrections réalisées par les outils Quiver et Pilon.

Ces outils ont permis de corriger des SNPs et indels.

	Correction par l'outil Quiver	Correction par l'outil Pilon
Nombre SNPs	25 663	696
Nombre insertions	133 867	29 888
Nombre délétions	1 766	1 604

Au final, nous disposons d'un nouvel assemblage du génome de *T. lutea*. Il est composé de 193 *contigs* et présente une taille de 82 Mpb. En comparaison avec la première version du génome établi par Carrier et al. (2018), celui-ci possède un gain de taille de 28 Mpb (54 Mpb vs 82 Mpb) et a l'avantage d'être beaucoup moins fragmenté (7 659 *contigs* vs 193 *contigs*). L'obtention d'un assemblage du génome ayant une taille plus importante, ainsi que la diminution de sa fragmentation, suggèrent qu'une part importante des régions répétées a pu être correctement assemblée. Ce nouvel assemblage du génome de *T. lutea* est donc mieux adapté pour permettre l'annotation la plus exhaustive possible des ETs. L'annotation automatique des gènes et la prédiction de leur fonction ont été réalisées à l'aide des outils MAKER2 et BLAST2GO (Conesa et al., 2005; Holt et Yandell, 2011). Avec ces outils, 15 972 gènes ont été identifiés, mais une fonction putative n'a pu être attribuée qu'à 26% d'entre eux.

En complément, une estimation de la taille du génome de *T. lutea* a été réalisée avec la méthode des k-mer. A l'aide de celle-ci, une taille de 85 Mpb a été obtenue, indiquant que l'assemblage du génome est proche de la taille réelle, mais qu'il semble encore incomplet. De même, afin de confirmer qu'aucun génome bactérien n'a été séquencé lors de ce processus, des données transcriptomiques de *T. lutea* ont été alignées sur l'assemblage du génome afin de vérifier que chaque *contig* était bien recouvert et correspondait bien à cette espèce.

2. Elaboration d'un pipeline d'annotation des ETs de *T. lutea*

2.1. Introduction

L'identification et l'annotation des ETs est une tâche difficile en raison de la diversité de ces éléments, la structure complexe de leurs séquences, leur présence en plusieurs copies dans les génomes et de la faible conservation de leur séquences nucléiques entre taxons. Une stratégie fortement utilisée pour obtenir une annotation exhaustive des ETs d'un génome se décline en trois étapes successives : premièrement la détection de séquences potentielles d'ETs, deuxièmement leur classification qui s'accompagne par la création d'une librairie de séquences d'ETs identifiés et troisièmement, l'annotation du génome à partir des séquences contenus dans la librairie (Flutre et al., 2012).

Les outils bio-informatiques qui mettent en œuvre cette stratégie et qui sont communément utilisés dans la littérature sont les pipelines d'annotation REPET et RepeatModeler (Smit et Hubley, 2010; Flutre et al., 2012). Ces outils utilisent des méthodes de détection *de novo*, principalement basées sur la détection de séquences répétées. Ils ont été calibrés à partir d'assemblages de génomes complets, bien décrits et issus d'organismes dits « modèles ». L'annotation à l'aide de ces outils chez les organismes dits « non-modèles » est fortement assujettie à deux facteurs : 1) la qualité de l'assemblage du génome et 2) la faible conservation des séquences nucléiques des ETs entre taxons.

En premier lieu, la qualité de l'assemblage influe directement sur l'identification exhaustive des ETs. En effet, ces outils s'appuient sur la détection des ETs par la recherche de séquences présentes en plusieurs copies dans le génome. Ainsi, l'utilisation d'un assemblage de génome fragmenté et de moindre qualité diminue l'efficacité de l'approche de ces outils (Bergman et Quesneville, 2007; Reinart, 2016 mémoire de thèse). La seconde limitation est produite par le manque d'informations concernant les ETs dans les taxons d'organismes peu étudiés. En effet, l'étape de classification est basée sur l'interrogation de banques de données de séquences d'ETs de référence. Ainsi, le nombre d'ETs répertoriés et la proximité phylogénétique de l'organisme étudié par rapport à ceux présents dans les

banques, impactent directement l'efficacité de la classification (Bergman et Quesneville, 2007; Hoen et al., 2015).

Afin de surmonter au mieux ces deux limitations lors de l'annotation des ETs chez *T. lutea*, un nouveau pipeline bio-informatique nommé PiRATE (*Pipeline to Retrieve and Annotate Transposable Elements*) a été développé et un génome de qualité a été produit (cf. Chapitre 1 § 1.4) (Bertheliet et al., 2018). L'objectif de ce nouveau pipeline est d'annoter les ETs présents dans les génomes d'espèces non-modèles et en particulier les ETs pouvant être autonomes. L'outil PiRATE est composé de trois étapes successives : la détection, la classification et l'annotation des ETs (Figure 36). L'étape de détection a été optimisée en combinant plusieurs outils présentant des approches différentes et celle de classification a été améliorée en enrichissant manuellement le répertoire des banques de séquences d'ETs référents.

2.2. Détection de séquences d'ETs

L'objectif de cette étape est la détection de séquence pleine longueur des copies des diverses familles d'ETs présentes dans l'assemblage d'un génome. La détection de séquences d'ETs pleine longueur facilite leur reconnaissance lors de la classification, et leur utilisation comme séquences de références conduit à une annotation plus précise du génome. Afin d'optimiser l'étape de détection, la combinaison de plusieurs logiciels présentant des stratégies différentes est conseillée (Hoen et al., 2015). Une recherche bibliographique a permis de recenser une quarantaine de programmes de détection (Tableau 3). Tous ces outils ont été regroupés en quatre approches, selon le type de méthode qu'ils mettent en œuvre : 1) recherche de séquences similaires à celles d'ETs référents ; 2) recherche de structures et de motifs de séquences d'ETs ; 3) recherche des séquences répétées présentes dans l'assemblage du génome ; 4) construction des séquences répétées du génome à partir de données de séquençage brutes.

Douze logiciels complémentaires couvrant ces quatre approches de détection ont été sélectionnés pour intégrer l'étape de détection de PiRATE. La méthodologie de ces différents outils est détaillée ci-dessous.

Tableau 3 : Liste des outils disponibles visant la détection de séquences d'ETs.

Les outils ont été regroupés en quatre approches. Ceux sélectionnés pour construire l'étape de détection de PiRATE ont été soulignés.

Nom	Site internet	Auteurs
Recherche par similarité		
Censor	girinst.org/censor/download.php	Jurka et al., 1996
<u>RepeatMasker</u>	repeatmasker.org/	Smit et Hubley, 1996
TEannot	urgi.versailles.inra.fr/Tools/REPET	Quesneville et al., 2005
Windowmasker	ftp.ncbi.nlm.nih.gov/pub/agarwala/windowmasker	Morgulis et al., 2006
TransposonPSI	transposonpsi.sourceforge.net/	Haas, 2007
TARGeT	target.iplantcollaborative.org	Han et al., 2009
TESeeker	repository.library.nd.edu/view/27/teseecker	Kennedy et al., 2011
Recherche par structure		
TRANSCO	alggen.lsi.upc.es/recerca/search/transpo/transpo.html	Santiago et al., 2002
LTR_STRUC	mcdonaldlab.biology.gatech.edu/finalLTR.htm	McCarthy et McDonald, 2003
LTR_par	eecs.wsu.edu/~anant/software.htm	Kalyanaraman et Aluru, 2006
LTR_FINDER	tlife.fudan.edu.cn/ltr_finder/	Xu et Wang, 2007
find_ltr	darwin.informatics.indiana.edu/cgi-bin/evolution/ltr.pl	Rho et al., 2007
<u>LTRharvest</u>	zbh.uni-hamburg.de/LTRharvest/index.php	Ellinghaus et al., 2008
MUST	healthinformatics.org/supp/	Chen et al., 2009
<u>HelSearch</u>	sourceforge.net/projects/helsearch/files	Lixing et Bennetzen, 2009
MGEScan-LTR	darwin.informatics.indiana.edu/cgi-bin/evolution/daphnia_ltr.pl	Tang, 2009
<u>MGEScannonLTR</u>	darwin.informatics.indiana.edu/cgi-bin/evolution/nonltr/nonltr.pl	Tang, 2009
<u>MITE-Hunter</u>	target.iplantcollaborative.org/mite_hunter.html	Han et Wessler, 2010
<u>Sine-Finder</u>	plantcell.org/content/early/2011/09/07/tpc.111.088682	Wenke et al., 2011
MITE-Digger	ncbi.nlm.nih.gov/pmc/articles/PMC3680318	Yang, 2013
HelitronScanner	sourceforge.net/p/helitronscanner	Xiong et al., 2014
detectMite	sourceforge.net/projects/detectmite	Congtig et al., 2016
SINE_scan	github.com/maohlzj/SINE_Scan	Mao et Wang, 2016
MUSTv2	healthinformatics.org/supp/resources.php	Ge et al., 2017
MITE Tracker	github.com/INTABiotechMJ/MITE-Tracker	Crescente et al, 2018

Nom	Site internet	Auteurs
Recherche les séquences répétées dans un assemblage de génome		
REPuter	bibiserv2.cebitec.uni-bielefeld.de/reputer	Kurtz et Schleiermacher, 1999
RepeatFinder	cbcb.umd.edu/software/RepeatFinder	Volfovsky et al., 2001
RECON	selab.janelia.org/recon.html	Bao et Eddy, 2002
GROUPER	urgi.versailles.inra.fr/Tools/REPET	Quesneville et al., 2003
PILER	drive5.com/piler	Edgar et Myers, 2005
<u>RepeatScout</u>	repeatscout.bioprotects.org	Price et al., 2005
Repseek	wwwabi.snv.jussieu.fr/public/RepSeek	Achaz et al., 2006
PClouds	evolutionarygenomics.com/ProgramsData/PClouds	Gu et al., 2008
Tallymer	zbh.uni-hamburg.de/?id=211	Kurtz et al., 2009
<u>TEdenovo</u>	urgi.versailles.inra.fr/Tools/REPET	Flutre et al., 2011
RepeatModeler	repeatmasker.org/RepeatModeler.html	Smit et Hubley, 2014
Red	toolsmith.ens.utulsa.edu	Girgis, 2015
phRAIDER	github.com/karroje/phRAIDER	Schaeffer et al., 2016
Construit les séquences répétées du génome à partir de données brutes		
ReAS	ftp.genomics.org.cn/pub/ReAS/software	Li et al., 2005
<u>RepeatExplorer</u>	galaxy.umbr.cas.cz:8080	Novák et al., 2013
Tedna	urgi.versailles.inra.fr/Tools/Tedna	Zytnicki, 2014
<u>RepARK</u>	https://github.com/PhKoch/RepARK	Koch et al., 2014
<u>dnaPipeTE</u>	lbbe.univ-lyon1.fr/-dnaPipeTE	Goubert et al., 2015
Transposome	github.com/sestato/Transposome	Staton et Burke, 2015
REPdenovo	github.com/Reedwarbler/REPdenovo	Chu et al., 2016
RepLong	github.com/ruiguo-bio/replong	Guo et al., 2018

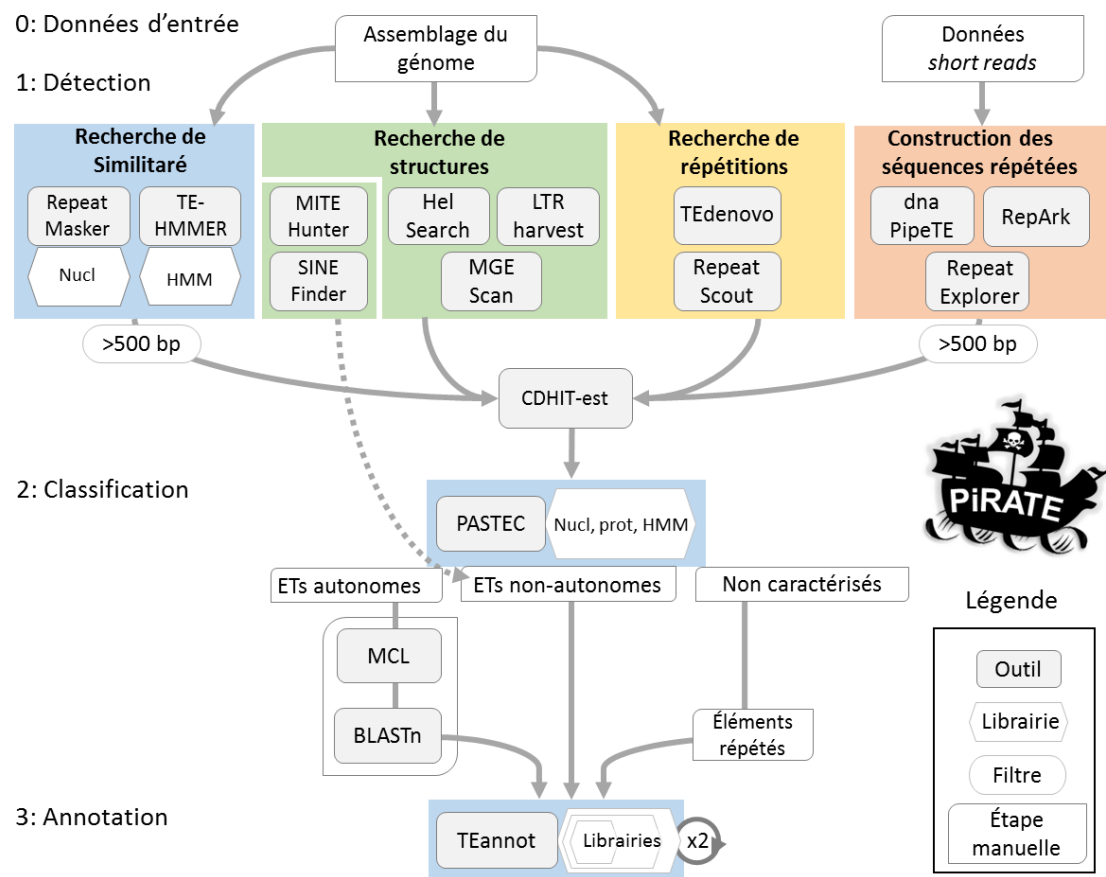


Figure 36 : Représentation schématique des différentes étapes du pipeline PiRATE.

Image adaptée de Bertheliet et al. (2018).

Approche 1 : recherche de séquences similaires à celles d'ETs référents

La première approche permet la détection de séquences d'ETs, en comparant les séquences du génome de l'espèce d'étude à des banques de données composées de séquences d'ETs référents ou de profils *Hidden Markov Model* (profils HMMs) qui ont été construits à l'aide de séquences d'ETs référents (Figure 37). Les outils de cette approche permettent la détection d'ETs présents en une seule ou plusieurs copies dans l'assemblage du génome d'étude. Cependant, l'efficacité des outils utilisant cette approche sera dépendante de la composition de leurs banques de données. L'identification des séquences des ETs est ainsi peu précise lors de l'analyse de génomes d'organismes non modèles (Hoen et al., 2015). Dans ce cas de figure, les outils n'identifient pas des séquences d'ETs pleines longueurs, mais des fragments de séquences (Hoen et al., 2015). De même, cette approche n'est pas pertinente pour la détection d'éléments non-autonomes composés de séquences non-codantes comme les MITE et SINE (Bergman et Quesneville, 2007). Elle ne permet pas non plus la détection de nouvelles Superfamilles ou Ordres d'ETs.

Pour cette première approche, le logiciel RepeatMasker (Smit et al., 1996) a été choisi pour intégrer PiRATE. Pour le compléter, un nouveau programme nommé TE-HMMER a été conçu dans le cadre de cette thèse.

Le programme RepeatMasker permet de détecter des régions nucléiques du génome assemblé, similaires aux séquences d'ETs présentes dans la banque de données *ad hoc*. En complément, TE-HMMER permet la détection d'ETs dans un génome assemblé à partir d'une banque de profils HMMs (Figure 37). Pour cela, il convertit tout d'abord l'assemblage du génome en acide aminées, et ceci dans les six cadres de lectures possibles. Puis il utilise le programme HMMER (Eddy et al., 1995) pour rechercher des séquences protéiques similaires aux profils HMMs de sa banque de données. Enfin, l'outil tBLASTn est utilisé pour retrouver les séquences nucléiques correspondantes dans le génome de l'organisme concerné.

Pour augmenter l'efficacité de ces deux outils lors de l'analyse du génome de *T. lutea*, leurs banques de données respectives ont été enrichies par des séquences d'ETs détectées dans des génomes de microalgues (cf. Chapitre 1 § 2.4).

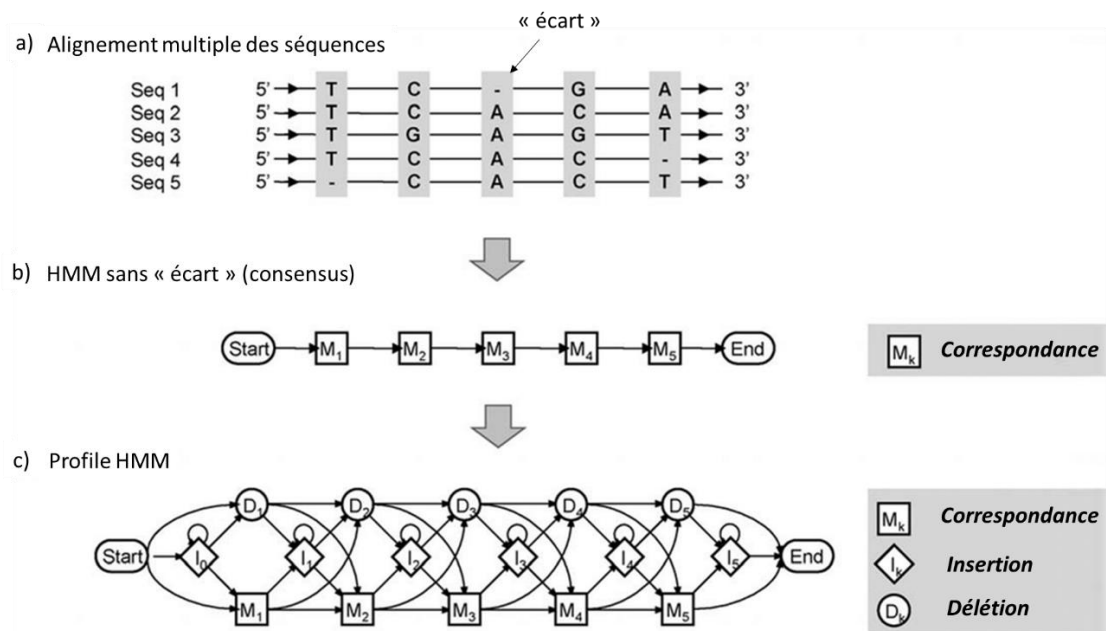


Figure 37 : Méthode de construction de profils HMMs (*Hidden Markov Models*).

a) Un alignement multiple de séquences est réalisé b) Le HMM sans « écarts » représente la séquence consensus de l'alignement. c) Le profil HMM final permet l'incorporation d'informations tels que des évènements d'insertions et de délétions. Image adaptée de Yoon (2009).

Approche 2 : recherche de structures et de motifs de séquences d'ETs

Les outils de cette approche utilisent des méthodologies basées sur la recherche de structures, motifs ou domaines protéiques spécifiques d'ETs. Ces outils permettent d'améliorer la détection des séquences du type d'ETs dont ils sont spécialisés. Les outils peuvent généralement détecter des familles d'ETs présents en peu de copies dans l'assemblage du génome. Néanmoins, les outils de cette approche peuvent présenter des limites lors de l'étude d'organismes non-modèles. En effet, leurs critères de recherche s'appuient généralement sur la structure ou des motifs d'ETs qui ont été caractérisés chez des organismes modèles appartenant à un taxon en particulier. De même, certains Ordres d'ETs présentent des structures de séquences plus hétérogènes et seront ainsi plus difficiles à détecter (e.g. Helitron) (Bergman et Quesneville, 2007). Dans cette approche, les programmes LTRharvest, MGEScan non-LTR, SINE-Finder, MITE-Hunter et Helsearch ont été sélectionnés. Ils permettent d'améliorer la détection de séquences complètes de LTR, LINE, SINE, MITE et Helitron dans l'assemblage du génome de l'organisme d'étude. Tout comme les outils de l'approche précédente, ils ne sont pas conçus pour aboutir à la découverte de nouveaux Ordres ou de nouvelles Superfamilles d'ETs.

Le programme LTRharvest (Ellinghaus et al., 2008) détecte des séquences d'ETs de l'Ordre des LTR. Il identifie des paires de terminaisons LTRs pouvant appartenir à un même élément LTR (Figure 38). Il a été conçu pour identifier des éléments portant deux terminaisons LTRs et ne détecte pas les ETs ayant perdu une de leurs terminaisons. Il peut identifier des séquences autonomes ou non-autonomes (TRIM et LARD) des différentes Superfamilles de rétrotransposons LTR.

Ensuite, MGEScan-nonLTR (Rho et Tang, 2009) est utilisé pour la détection d'éléments LINE. Cet outil recherche des domaines protéiques connus chez les LINE avec HMMER ainsi que les régions de liaison de ces domaines. Il permet la détection de LINE complets, mais également de LINE tronqués ayant perdu l'un de leurs domaines protéiques.

L'outil SINE-Finder (Wenke et al., 2011) permet la détection de séquences SINE. Ces ETs sont difficiles à identifier en raison de leur petite taille et de leur structure complexe. Pour cela, SINE-Finder recherche différents motifs et structures de séquences qui ont été caractérisés chez des SINE provenant de génomes de plantes (Figure 39).

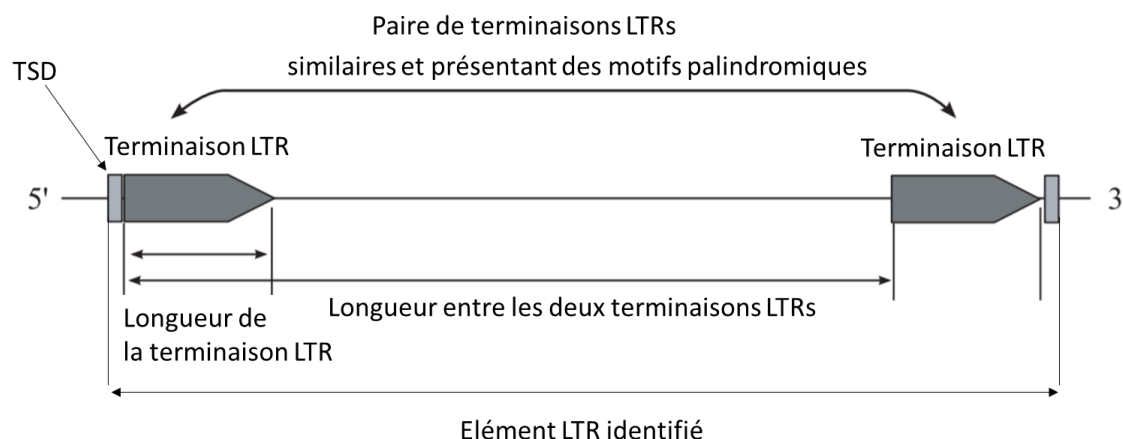


Figure 38 : Représentation schématique des paramètres recherchés par LTRharvest.

L'outil cherche à identifier des paires de séquences terminales LTRs pouvant appartenir à un rétrotransposon LTR. Divers filtres sont appliqués pour sélectionner une possible paire de terminaisons LTRs : une taille (en pb) minimum et maximum, une distance (en pb) minimum et maximum entre les deux LTRs, un pourcentage d'identité minimum entre les deux terminaisons, l'identification de TSD et de motifs palindromiques. L'ET identifié correspond à la séquence comprise allant d'une terminaison LTR à l'autre. Image adaptée de Steinbiss et al. (2009).

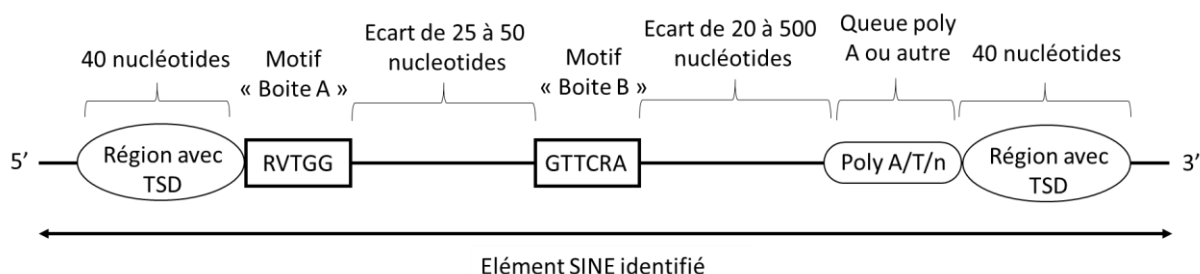


Figure 39 : Représentation des motifs et structures recherchés par SINE-Finder.

Une séquence sera détectée comme appartenant à un SINE si elle présente les caractéristiques suivantes : un motif TSD situé au niveau d'une région 5' de 40 nucléotides suivi par un motif « RVTGG » caractéristique de la polymérase l'ARN polymérase III, un écart de 25 à 50 nucléotides, un motif « GTTCRA » caractéristique de l'ARN polymérase III, un espace de 20 à 500 nucléotides, puis une extrémité poly-dA de six adénines et enfin une région de 40 nucléotides présentant un motif TSD en 3' de l'élément.

Avec une stratégie similaire, le programme HelSearch (Yang et Bennetzen, 2009) recherche des séquences d'Helitrons présents en plusieurs copies (minimum deux) dans le génome de l'organisme d'étude. Ces ETs sont particulièrement difficiles à détecter en raison de leurs structures nucléiques complexes et hétérogènes (Hoen et al., 2015). De façon similaire à l'outil SINE-Finder, HelSearch recherche différents motifs et structures nucléiques spécifiques d'Helitron identifiés chez des plantes (Figure 40).

Enfin, le programme MITE-Hunter (Han et Wessler, 2010) est utilisé pour la détection de MITE. Les MITE sont des éléments non-autonomes, difficiles à identifier en raison de leur petite taille et de leurs séquences non codantes (Feschotte et Mouches, 2000). MITE-Hunter recherche des séquences répétées qui présentent des structures similaires à des terminaisons TIRs et des TSDs. Les séquences identifiées comme des MITE candidats sont ensuite utilisées pour retrouver avec BLASTn des copies de ceux-ci dans le génome étudié. L'ensemble des copies sont alors regroupées en fonction de leur degré de similarité. Pour chaque groupe, un alignement multiple des séquences est réalisé, permettant la construction d'une séquence consensus qui sert de référence.

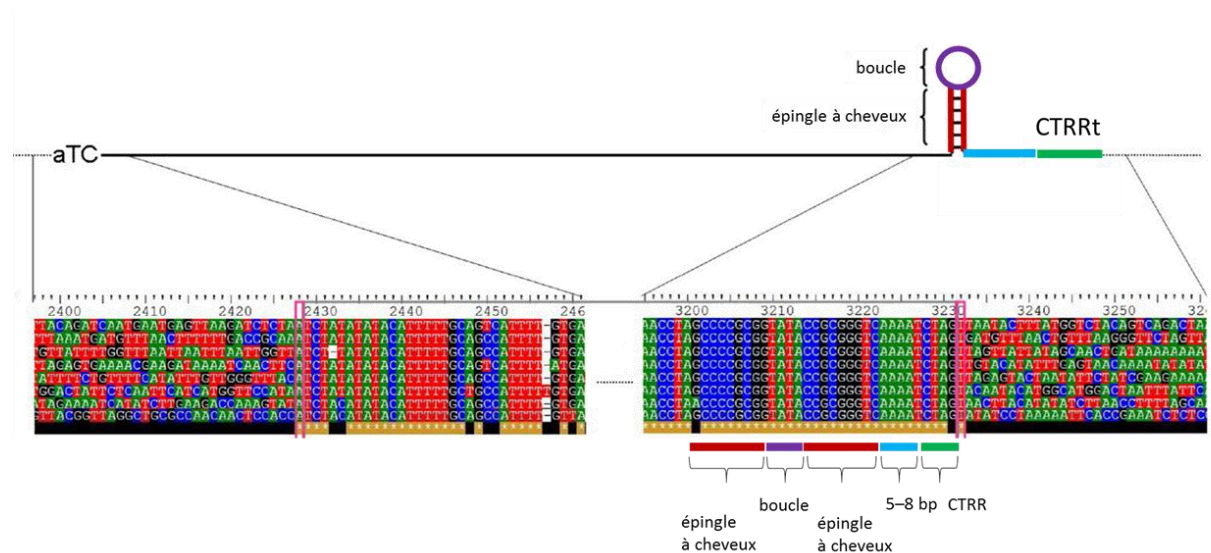


Figure 40 : Représentation des motifs et structures recherchés par HelSearch.

Cet outil examine différentes structures connues de terminaisons 3' d'Helitron identifiés chez des plantes. Dans cette figure, des séquences nucléiques Helitrons sont alignées afin de mettre en avant les motifs spécifiques recherchés par cet outil. HelSearch vise l'identification de séquences présentant les structures suivantes : un motif « CTRRt », puis un minimum de 6 paires de nucléotides en « épiingle à cheveux » situées en amont. Il cherche ensuite la présence d'une boucle de 2 à 4 bp et enfin un espace de 5 à 8 nucléotides entre la séquence en « épiingle à cheveux » et le « CTRRt ». Les séquences candidates sont regroupées en fonction de leur structure « d'épiingle à cheveux ». Image adaptée de Yang et Bennetzen (2009).

Approche 3 : recherche des séquences répétées présentes dans l'assemblage du génome

Cette approche utilise une démarche *de novo* qui consiste à rechercher l'ensemble des séquences répétées présentes dans un assemblage de génome. La détection de tous types d'ETs, sans connaissance préalable de leur structure ou leur séquence, représente un atout particulier pour l'étude des génomes d'espèces non-modèles comme *T. lutea*. En revanche, les outils utilisés dans cette approche sont moins efficaces lors de l'utilisation d'un assemblage de génome fragmenté et peuvent ne pas détecter les familles d'ETs présentes en peu de copies (Bergman et Quesneville, 2007; Reinart, 2016 mémoire de thèse). Enfin, ces outils ne sont pas spécifiques à la détection d'ETs mais permettent la détection de tous types de séquences répétées. Une étape de classification est donc nécessaire pour identifier les séquences d'ETs détectées. Pour cette approche, les programmes RepeatScout et TEdenovo ont été sélectionnés.

Le programme RepeatScout recherche des régions nucléiques répétées dans l'assemblage d'un génome. En premier lieu, cet outil utilise une stratégie d'identification de séquences répétées dites de « k-mer », c'est à dire qu'il recherche les motifs répétés de taille k nucléotides et compte leur abondance dans l'assemblage du génome (par exemple pour toutes les séquences répétées de taille k=10). Après identification de ces séquences, il cherche à les « étendre » et à obtenir de plus longues séquences consensus. Une illustration de la méthodologie de RepeatScout est proposée en Figure 41.

Le pipeline TEdenovo (Flutre et al., 2012) permet aussi l'identification des ETs à travers la recherche des séquences nucléiques répétées. Pour cela, TEdenovo découpe l'assemblage du génome en fragments chevauchants. Puis, il compare ces fragments entre eux avec l'outil BLASTER (Quesneville et al., 2003) et cherche à identifier des paires de séquences similaires correspondant à des séquences répétées du génome. Les séquences répétées sont ensuite regroupées par l'utilisation de trois outils différents : GROUPE (Quesneville et al., 2003), RECON (Bao et Eddy, 2002) et PILER (Edgar et Myers, 2005). Chaque groupe obtenu est ainsi constitué de plusieurs séquences similaires. Un alignement multiple des séquences est réalisé pour chaque groupe, permettant ensuite la construction d'une séquence consensus. Ces séquences consensus sont ainsi représentatives de séquences répétées du génome et donc potentiellement de familles d'ETs.

a) k-mer (k=10)

10-mer	abondance
CAACGTCTGC	5
TTGGCTTCAA	3
GGCCTTACGC	10
...	...
TACCGTTACTC	15



b)

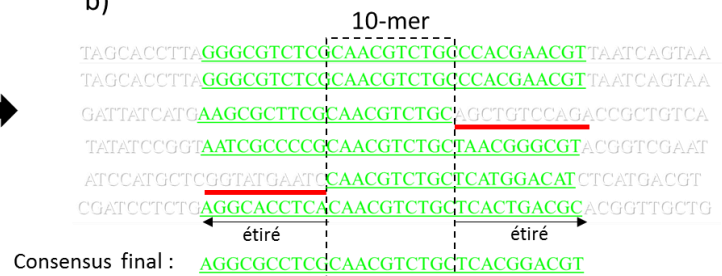


Figure 41 : Illustration de la méthodologie de RepeatScout.

RepeatScout utilise une stratégie d'identification de séquences répétées dites de « k-mer », c'est-à-dire qu'il recherche les motifs répétés de taille k nucléotides et compte leur abondance dans l'assemblage du génome. a) Dans cet exemple RepeatScout recherche et liste l'abondance des k-mer de taille 10 nucléotides (10-mer) dans l'assemblage du génome de l'organisme d'étude. b) les régions nucléiques contenant les 10-mer sont alignées et les 10-mer d'intérêt sont « étirés » nucléotide par nucléotide. Si une séquence de l'alignement est trop divergente (soulignée en rouge), elle n'est plus utilisée dans la suite de « l'étirement » du consensus. La séquence consensus cesse d'être « étirée » lorsque la majorité des séquences ne s'aligne plus. Image adaptée de la présentation de A. Price et al. (2005) bix.ucsd.edu/repeatscout/repeatscout-ismb.ppt

Approche 4 : construction des séquences répétées du génome à partir de données de séquençage brutes

Cette approche est basée sur la détection d'ETs chez une espèce, en utilisant des données de séquençage non assemblées afin de construire les séquences répétées présentes dans son génome. Les outils de cette approche sont donc très pertinents pour la détection des ETs chez les organismes non-modèles car ils permettent de s'affranchir de la qualité de l'assemblage. En revanche, bien que de tels outils réalisent la détection *de novo* de tous types d'ETs, ils ne permettent généralement pas la construction pleine longueur de leurs séquences (Chu et al., 2016). De plus, ces outils ne sont pas spécifiques à la construction des ETs, mais de tout type de séquences répétées. Comme pour l'approche précédente, une étape de classification est donc nécessaire pour identifier les séquences d'ETs détectées. Les outils RepeatExplorer (Novak et al., 2013), RepArk (Koch et al., 2014) et dnaPipeTE (Goubert et al., 2015) ont été sélectionnés pour cette approche.

L'outil RepeatExplorer permet la détection d'ETs à travers la construction de séquences répétées, à partir de données *short reads*. Dans un premier temps, cet outil réalise une comparaison par paires de celles-ci afin d'identifier celles qui sont similaires et chevauchantes. Cette information permet la construction d'un graphique composé de *short reads* chevauchantes, qui sont théoriquement issues de régions répétées du génome (Figure 42). Ce graphique est ensuite partitionné en groupes de *short reads* qui possèdent un plus fort degré de similarité entre elles, plutôt qu'avec le reste des *short reads* du graphique (Novak et al., 2013). Les *short reads* de chaque groupe sont alors assemblées pour construire une séquence répétée du génome qui peut correspondre à un ET.

L'outil dnaPipeTE reconstruit aussi les séquences répétées du génome d'étude à l'aide de données *short reads*. Il effectue un échantillonnage aléatoire des séquences *short reads* disponibles. L'assemblage des *short reads* est ensuite réalisé par au moins deux itérations de l'outil d'assemblage Trinity (Haas et al., 2013). Pour chaque itération, les *short reads* précédemment assemblées sont ajoutées à l'échantillonnage suivant afin d'améliorer l'assemblage des séquences répétées. Au final, les séquences reconstruites peuvent correspondre à des ETs.

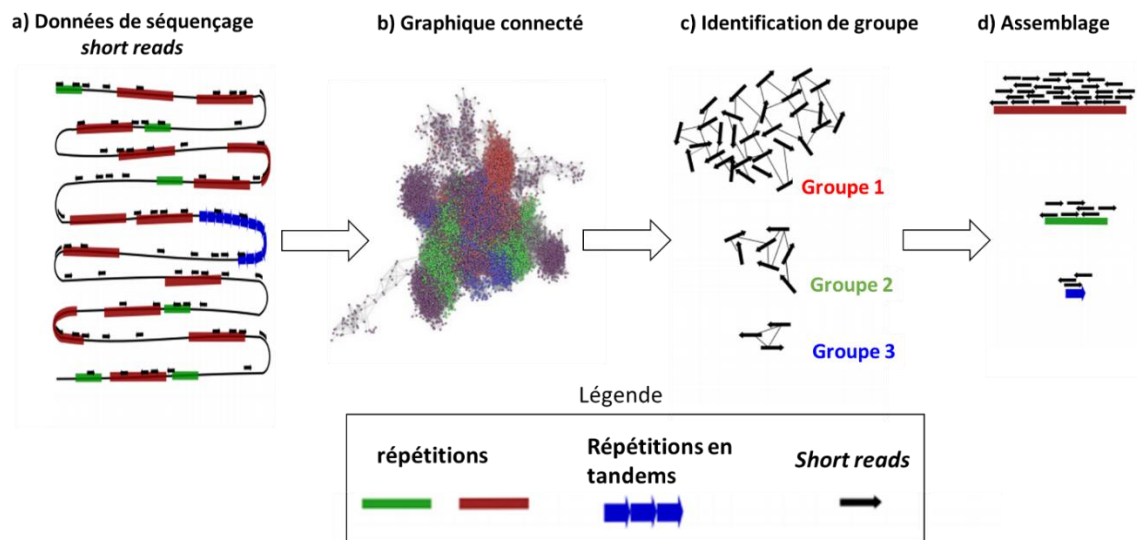


Figure 42 : Illustration de la méthodologie de RepeatExplorer.

a) RepeatExplorer utilise des données de séquençage *short reads*. b) Une comparaison par paires des *short reads* permet ensuite la construction d'un graphique de *short reads* chevauchantes. Lors de la construction du graphique, les *shorts reads* qui correspondent à des séquences non-répétées du génome donneront des nœuds isolés. En revanche, les séquences répétées formeront des groupes de nœuds connectés. c) Ce graphique est ensuite partitionné en groupes de *short reads*, en fonction de leur degré de similarité (Novák et al., 2010). Pour chaque groupe, les *short reads* sont assemblées pour reconstruire une séquence répétée. Image adaptée de repeatexplorer.org/wp-content/uploads/Macas_repeat_identification_2017.pdf

Enfin, le programme RepARK vise également la construction des séquences répétées d'un génome à partir de données *short reads*. Pour cela, RepARK vise à identifier les séquences répétées dites de « k-mer » dans les *short reads* (Figure 43). Les k-mer les plus abondantes sont ensuite isolées et assemblées afin de construire les séquences répétées (Zerbino, 2010). Les séquences construites peuvent ainsi correspondre à des ETs.

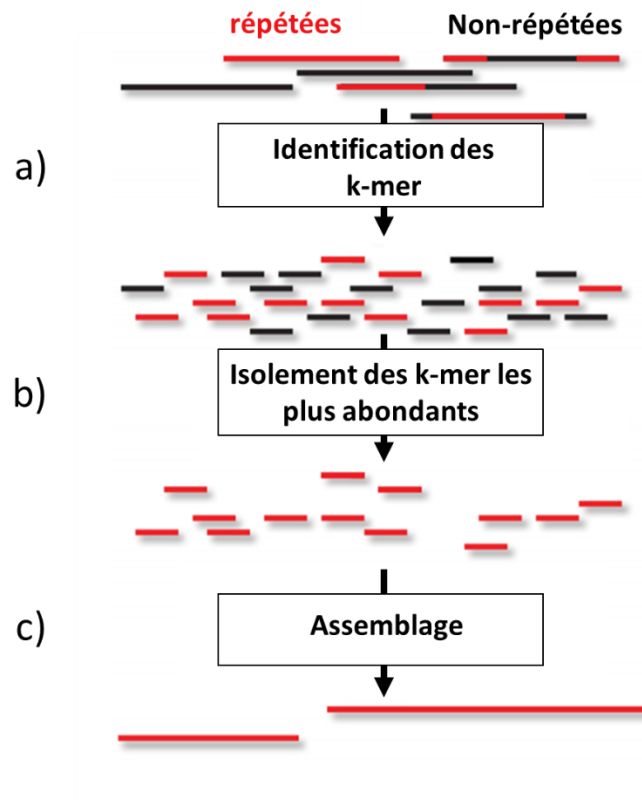


Figure 43 : Représentation schématique de la méthodologie de RepARK.

a) Les séquences répétées en k-mer sont recherchées dans les *short reads*. b) Les k-mer les plus abondants sont ensuite isolés. c) Finalement RepARK cherche à reconstruire les séquences répétées du génome à l'aide de l'outil d'assemblage Velvet (Zerbino, 2010). Image adaptée de Koch et al. (2014).

Ces différents outils ont été utilisés avec des données génomiques de *T. lutea* (génomome assemblé ou *short reads*). Au final, ils ont permis de détecter 514 337 séquences pouvant correspondre à des ETs (Tableau 4). Différents filtres ont été appliqués afin d'en diminuer le nombre.

Tableau 4 : Nombre de séquences potentielles d'ETs détectées par les outils, sans et avec filtres.

Une « * » indique que le filtre considéré n'a pas été appliqué pour ces données.

Outils :	Nombre séquences détectées		
	Sans filtre	Filtre « >500 »	Filtre « CDHIT-est »
RepeatMasker	19 207	365	259
TE-HMMER	9727	9540	288
LTRharvest	1503	1503*	1454
MGEScan-nonLTR	11	11*	8
SINE-Finder	10	10*	10*
MITE-Hunter	25	25*	25*
Helsearch	1407	1407*	919
RepeatScout	4126	4126*	2540
TEdenovo	5936	5936*	5483
RepeatExplorer	416 448	12 416	12 209
RepArk	41 220	1965	1138
dnaPipeTEs	14 717	2206	2143
Total :	514 337	39 510	26 476

2.3. Réduction du nombre de séquences détectées

Les séquences identifiées dans les données génomiques de *T. lutea* sont majoritairement de taille inférieure à 500 pb ce qui est trop faible pour constituer des ETs autonomes complets. Ainsi, un filtre de taille (supérieur à 500 pb) a été appliqué pour les séquences détectées par les outils des approches 1 et 4 (Tableau 4). Ensuite, les séquences restantes pour ces deux approches et celles détectées par les outils des approches 2 et 3 ont été regroupées. Toutes ces séquences ont ensuite été soumises à l'outil CD-HIT-est (Li et al., 2001) afin de retirer les séquences redondantes de petite taille et de ne garder que des séquences uniques de plus grande taille (Figure 44). Au final, 26 476 séquences ont été soumises à une étape de classification automatique. Seule exception, les séquences détectées par les logiciels MITE-Hunter et SINE-Finder n'ont pas été soumises à ces filtres car les ETs que ces outils détectent peuvent être de taille inférieure à 500 pb et ont donc été directement conservés pour l'étape de classification (35 séquences).

2.4. Classification des séquences détectées

L'étape de classification permet de reconnaître et de classer les séquences potentielles d'ETs qui ont été détectées lors de la première étape de PiRATE. Pour rappel, la classification des ETs fait encore l'objet de vifs débats parmi la communauté scientifique. Les deux principales classifications en concurrence sont celles proposées par Wicker et collaborateurs en 2007 (Wicker et al., 2007) et celles des créateurs de la base de données Repbase (Kapitonov et Jurka, 2008). Dans cette thèse, la classification de Wicker et al. (2007) a été choisie. Elle classe les ETs en Classes, Sous-classes, Ordres, Superfamilles et familles. Le programme PASTEC (*Pseudo Agent System for Transposables Elements Classification*, Hoede et al. 2014) permet de classer automatiquement les séquences d'ETs, en prenant en compte la classification de Wicker et al. (2007).

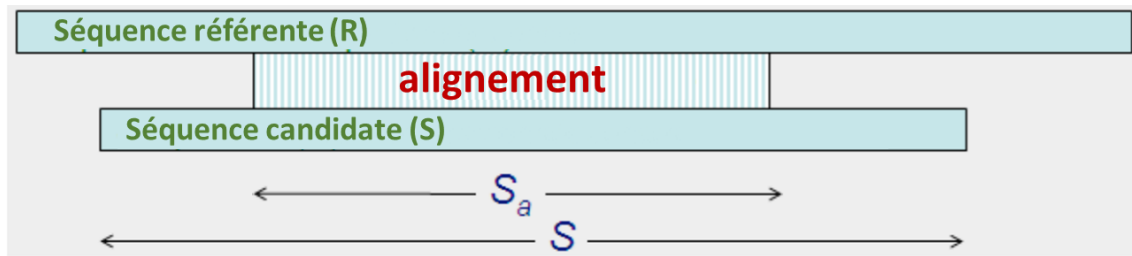


Figure 44 : Représentation schématique des paramètres de CD-HIT-est.

Dans PiRATE, le paramètre S_a/S est égal à 1 et le pourcentage d'identité est égal à 100%.

Image adaptée de bioinformatics.org/cd-hit/cd-hit-user-guide.pdf.

Présentation de l'outil PASTEC

Le programme PASTEC a pour objectifs 1) de prédire automatiquement si les séquences détectées et filtrées lors de la première étape de PiRATE correspondent à des ETs ; 2) de tenter de les assigner à une Classe et un Ordre (Hoede et al., 2014). Pour cela, l'outil recherche des caractéristiques structurelles : la longueur de la séquence, la présence de terminaisons LTRs ou TIRs, la présence de répétitions de séquences simples, l'existence d'une queue poly-dA et d'un éventuel cadre de lecture (Hoede et al., 2014). PASTEC utilise les programmes BLASTx et tBLASTx pour détecter des similitudes avec les séquences nucléotidiques ou de protéines d'ETs référents et répertoriées dans leurs banques de données respectives. Pour cela, des banques de données standards sont disponibles sur Repbase pour PASTEC. L'outil HMMER est utilisé pour détecter des domaines protéiques à partir d'une banque de données de profils HMMs. Une banque de données standard est distribuée par les concepteurs de cet outil (urgi.versailles.inra.fr/Tools/PASTECClassfier).

Concernant le pipeline d'analyse PiRATE, les banques de données utilisées par PASTEC ont été améliorées pour l'analyse des génomes de microalgues (cf. Chapitre 1 § 2.4). PASTEC analyse chaque séquence en testant toutes les classifications possibles des ETs et chaque possibilité de classification se voit attribuer un score en fonction des évidences identifiées (Hoede et al., 2014). Ainsi, il détermine si la séquence étudiée correspond à une séquence d'ETs, avant de lui attribuer une Classe et un Ordre (Hoede et al., 2014).

Amélioration de l'étape de classification

Afin de recenser les séquences des différentes familles d'ETs présentes dans les génomes des organismes eucaryotes et procaryotes, des banques de données spécifiques à ces éléments ont été créées et sont disponibles. En effet, certaines banques de données sont spécialisées pour un type d'élément particulier chez un organisme précis. Pour exemple, la base de données « Human Endogenous Retrovirus Database » (<https://herv.img.cas.cz/>) est spécialisée dans le recensement de rétrovirus identifiés dans le génome humain. D'autres bases de données sont davantage spécialisées dans le recensement de tous les types d'ETs. Par exemple « IS-finder », est spécialisé dans le recensement des différents groupes d'IS présents chez les procaryotes (isfinder.biotoul.fr/). La base de données Repbase (girinst.org/repbase/) est spécialisée dans l'inventaire des séquences répétées chez les

eucaryotes et en particulier les ETs. Cette dernière représente la base de données la plus complète et la plus utilisée pour des analyses de bio-informatique chez les génomes d'eucaryotes. Comme indiqué précédemment, les banques de données (nucléiques et protéiques) que Repbase distribue sont notamment utilisées par l'outil PASTEC et celles-ci ont ainsi été utilisées par PiRATE.

Etant donné que Repbase n'utilise pas la même classification d'ETs que le programme PASTEC, une version spécialement dédiée à cet outil et utilisant la classification de Wicker est disponible. Ces banques de données souffrent néanmoins de l'absence de données relatives à certains taxons, et notamment de données propres aux microalgues. Par exemple, chez les Haptophytes, seules 17 séquences d'ETs étaient recensées lors de cette thèse. Afin de diminuer la difficulté de classification liée à ce manque d'information, une recherche de séquences d'ETs de microalgues, non-référencées dans Repbase, a été effectuée. Les séquences correspondantes ont été majoritairement retrouvées dans la base de données NCBI. D'autres séquences provenant de banques de données ISfinder (isfinder.biotoul.fr/), SINEbases (sines.eimb.ru/) et Plant-MITE-databases (pmite.hzau.edu.cn/mite_more/) ont également été récupérées. L'ensemble des séquences récupérées a ensuite été ajouté à la banque de données de PiRATE. Ensuite, pour améliorer la reconnaissance des domaines protéiques d'ETs avec HMMER (utilisé par PASTEC), de nouveaux profils HMMs ont été créés. Pour la création de ces profils, l'utilisation de séquences d'ETs de microalgues ou d'organismes phylogénétiquement proches aux Haptophytes a été favorisée.

2.5. Création des librairies de séquences d'ETs référents

PASTEC a permis d'apporter une classification automatique des séquences détectées. L'objectif suivant a été de créer trois librairies de séquences d'ETs, qui seront utilisées pour l'annotation du génome de *T. lutea*. Ces trois librairies ont été construites selon l'idée de « poupées russes » (Figure 45). Les trois librairies sont les suivantes : une librairie contenant des séquences d'ETs potentiellement autonomes, une librairie contenant des séquences d'ETs potentiellement autonomes et non-autonomes et une librairie contenant l'ensemble des éléments répétés détectés (Figure 45). Les critères utilisés pour sélectionner les séquences référentes qui constituent ces librairies sont décrits ci-après.

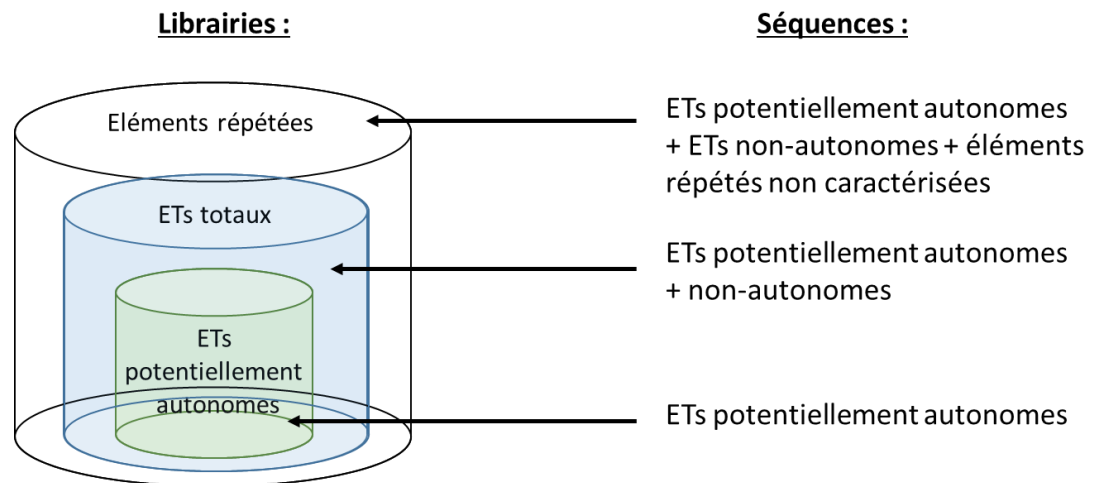


Figure 45 : Représentation schématique de la méthode de construction des trois librairies en « poupées russes ».

Les trois librairies sont les suivantes : une « librairie d’ETs potentiellement autonomes », une « librairie d’ETs totaux » et une « librairie des éléments répétés ».

Les séquences qui ont été classées par PASTEC comme appartenant à un Ordre d'ETs potentiellement autonomes (e.g. LTR, TIR) ont été manuellement sélectionnées et classées en Superfamilles, puis regroupées en familles. Afin de classer manuellement les séquences en Superfamilles, les résultats des requêtes des outils BLASTx, tBLASTx et HMMER, utilisé par PASTEC pour prédire la classification ont été analysés manuellement. Les séquences classées en Superfamilles ont ensuite été regroupées de façon automatique en familles en utilisant l'outil MCL. Cette approche a été préconisée par Hoede et al. (2014). Les familles constituées ont ensuite été manuellement affinées à l'aide de l'outil BLASTn (avec un minimum de 80% de similarité). L'utilisation de l'outil BLASTn permet de comparer entre elles les différentes séquences d'une même Superfamille pour s'assurer que la formation des familles est correcte et respecte la règle 80-80-80 de Wicker et al. (2007). Pour chaque famille, les séquences les plus complètes ont été sélectionnées comme séquences référentes pour former une première librairie nommée « librairie d'ETs potentiellement autonomes » (Figure 45).

Les séquences classées comme MITE, LARD, TRIM et SINE n'ont pas été vérifiées manuellement et ont été ajoutées dans la « librairie des ETs totaux » (Figure 45). Concernant les séquences classées comme LARD, seules les séquences provenant des outils TEdenovo et RepeatScout ont été sélectionnées. En effet, ces outils permettent de ne garder que des séquences répétées dans le génome, diminuant ainsi la présence de faux positifs. Enfin, les séquences classées comme noCAT par les outils TEdenovo et RepeatScout ont été ajoutées dans la librairie nommée « librairie des éléments répétés » (Figure 45). Ces séquences noCAT sont des séquences présentes en plusieurs copies, pouvant correspondre à des ETs mais dont PASTEC n'a pu assigner ni Classe, ni Ordre. Elles peuvent ainsi potentiellement être des ETs non identifiés ou correspondre à d'autres types de séquences répétées.

2.6. Annotation des ETs de *T. lutea*

L'annotation des ETs permet de cartographier la position des différentes copies de ces éléments dans le génome hôte.

L'étape d'annotation a été réalisée avec l'outil TEannot qui est intégré dans le package REPET (Quesneville et al., 2005). Le pipeline TEannot est divisé en six étapes décrites ci-après : 1) fragmentation de l'ensemble des *contigs* du génome en plusieurs séquences chevauchantes ; 2) alignement des séquences d'ETs présentes dans la librairie sur le génome fragmenté, via l'utilisation indépendante des logiciels BLASTER, RepeatMasker et CENSOR ; 3) les annotations obtenues par les trois logiciels précédents sont filtrées et combinées. Dans cette étape, les annotations chevauchantes sont éliminées et des connections entre séquences annotées sont réalisées avec la procédure « de rassemblement des séquences éloignées » (*long-join procedure*) ; 4) annotation des séquences simples répétées du génome à l'aide des outils TRF, Mreps et RepeatMasker (Smit et al., 1996.; Benson, 1999; Kolpakov, 2003); 5) comparaison de l'annotation des ETs avec celle des séquences simples répétées afin d'enlever ces dernières de l'annotation des ETs ; 6) élimination des doublons d'annotation des ETs et réaboutement des séquences annotées via le processus de « rassemblement des séquences éloignées ». Ce dernier processus, permet de connecter deux régions de la même copie d'un ET qui ont été séparées par une large insertion, affinant ainsi l'annotation (Flutre et al., 2012).

Pour réaliser l'annotation des ETs de *T. lutea*, les trois librairies construites précédemment (Figure 45), ont été utilisées séparément, afin d'obtenir trois annotations indépendantes du génome : une annotation des ETs potentiellement autonomes, une annotation de l'ensemble des ETs et une annotation de l'ensemble des séquences répétées.

La réalisation de ces trois annotations indépendantes permet de limiter les problèmes de compétition de séquences référentes qui peuvent s'aligner sur une même région du génome. En effet, lorsque deux séquences référentes peuvent annoter une même région, TEannot élimine les doublons d'annotation. Ceci est adéquat lorsque deux séquences référentes de la même famille d'ETs sont en « compétition », mais peut faire perdre des informations importantes lorsque deux séquences de niveau de classifications différentes sont situées dans la même région (e.g. séquence TIR vs séquence noCat).

TEannot est un outil efficace qui permet de combiner plusieurs méthodes d'annotation. En revanche, comme pour la plupart des outils d'annotation, le pourcentage de faux positifs est difficile à évaluer. TEannot détecte de nombreuses séquences très divergentes vis-à-vis des séquences référentes, aussi bien au niveau de leur similarité (min 60%) que de leur taille (min 200 pb). Afin de prévenir la présence de faux positifs, des filtres relatifs aux pourcentages d'identité et à la taille des séquences ont été appliqués lors de l'annotation des ETs de *T. lutea* (cf. Chapitre 1 § 4).

Un tableau récapitulatif des outils sélectionnés pour PiRATE est présenté en Annexe 2.

L'utilisation de PiRATE nous a permis d'estimer que 38,6 % du génome de *T. lutea* est constitué d'éléments répétés. Parmi ces éléments répétés, 20,8 % correspondent à des ETs et 3,8 % sont assignés à des ETs potentiellement autonomes. Ces résultats sont présentés plus en détail dans la publication relative à cette étude (cf. Chapitre 1 § 4).

3. Activité des ETs autonomes chez *T. lutea*

A l'image des gènes, les ETs autonomes peuvent être transcrits en ARNm via la machinerie cellulaire, témoignant ainsi de leur activité propre. Pour étudier l'activité des ETs autonomes de *T. lutea*, le laboratoire disposait de données transcriptomiques *short reads*. Celles-ci étaient produites par la technologie Illumina (*short reads*) sur des échantillons cellulaires obtenus en condition de limitation en azote (Carrier et al., 2014; Thiriet-Rupert et al., 2018). Ces *short reads* ont été alignées avec l'outil bwa sur le nouvel assemblage du

génomique de *T. lutea* pour identifier les régions transcrites du génome dans ces conditions de culture.

Parmi les 572 séquences d'ETs potentiellement autonomes et annotés (Chapitre 1 § 4) des transcrits ont été identifiés pour 17 familles d'ETs : une famille LTR Copia, quatre familles TIR Mariner, quatre familles TIR Harbinger et huit familles TIR hAT. Les séquences référentes propres à ces 17 familles ont ensuite fait l'objet d'une requête par BLASTx contre un protéome de *T. lutea* obtenu précédemment en condition de limitation azotée (Garnier et al., 2014; Thiriet-Rupert et al., 2016). Cette recherche a permis de mettre en évidence la présence d'une protéine de type transposase, respectivement pour une famille de TIR hAT et pour une famille de Mariner. Cette famille TIR hAT a été nommée *Ace* et est composée de 3 copies potentiellement autonomes dans le génome qui partagent une similarité de 99% d'identité. Concernant la famille TIR Mariner, celle-ci a été nommée *Luffy* et est formée par 20 copies potentiellement autonomes dans le génome. Ces copies présentent une similarité moyenne de 99% d'identité. De manière intéressante, ces deux familles font partie des trois transcrits les plus abondants en conditions de limitation azotée. L'expression et la synthèse des protéines de ces deux familles d'ETs indiquent ainsi une forte probabilité de mobilité de ces éléments autonomes et/ou pour des séquences MITE associées. Par conséquent, les éléments *Ace* et *Luffy* constituent des candidats très intéressants, pouvant participer à l'évolution du génome de *T. lutea*.

4. A transposable element annotation pipeline and expression analysis reveal potentially active elements in the microalga *T. lutea*

RESEARCH ARTICLE

Open Access



A transposable element annotation pipeline and expression analysis reveal potentially active elements in the microalga *Tisochrysis lutea*

Jérémy Bertheliet^{1,2*} , Nathalie Casse², Nicolas Daccord^{3,5}, Véronique Jamilloux⁴, Bruno Saint-Jean¹ and Grégory Carrier¹

Abstract

Background: Transposable elements (TEs) are mobile DNA sequences known as drivers of genome evolution. Their impacts have been widely studied in animals, plants and insects, but little is known about them in microalgae. In a previous study, we compared the genetic polymorphisms between strains of the haptophyte microalga *Tisochrysis lutea* and suggested the involvement of active autonomous TEs in their genome evolution.

Results: To identify potentially autonomous TEs, we designed a pipeline named PiRATE (Pipeline to Retrieve and Annotate Transposable Elements, download: <https://doi.org/10.17882/51795>), and conducted an accurate TE annotation on a new genome assembly of *T. lutea*. PiRATE is composed of detection, classification and annotation steps. Its detection step combines multiple, existing analysis packages representing all major approaches for TE detection and its classification step was optimized for microalgal genomes. The efficiency of the detection and classification steps was evaluated with data on the model species *Arabidopsis thaliana*. PiRATE detected 81% of the TE families of *A. thaliana* and correctly classified 75% of them. We applied PiRATE to *T. lutea* genomic data and established that its genome contains 15.89% Class I and 4.95% Class II TEs. In these, 3.79 and 17.05% correspond to potentially autonomous and non-autonomous TEs, respectively. Annotation data was combined with transcriptomic and proteomic data to identify potentially active autonomous TEs. We identified 17 expressed TE families and, among these, a TIR/Mariner and a TIR/hAT family were able to synthesize their transposase. Both these TE families were among the three highest expressed genes in a previous transcriptomic study and are composed of highly similar copies throughout the genome of *T. lutea*. This sum of evidence reveals that both these TE families could be capable of transposing or triggering the transposition of potential related MITE elements.

Conclusion: This manuscript provides an example of a de novo transposable element annotation of a non-model organism characterized by a fragmented genome assembly and belonging to a poorly studied phylum at genomic level. Integration of multi-omics data enabled the discovery of potential mobile TEs and opens the way for new discoveries on the role of these repeated elements in genomic evolution of microalgae.

Keywords: Transposable elements, Genome assembly, Pipeline, Tool, Annotation, Algae, Haptophyte, *Tisochrysis lutea*

* Correspondence: bertheliet.j@laposte.net

¹IFREMER, Physiology and Biotechnology of Algae Laboratory, rue de l'Île d'Yeu, 44311 Nantes, France

²Mer Molécules Santé, EA 2160 IUML - FR 3473 CNRS, Le Mans University, Le Mans, France

Full list of author information is available at the end of the article



© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

Background

Transposable Elements (TEs) are defined as DNA sequences able to move and spread within eukaryotic and prokaryotic genomes. These repeated elements constitute a variable fraction of eukaryotic genomes, ranging from 3% in the yeast *Saccharomyces cerevisiae*, 45% in human, to 80% in maize [1–3]. TEs were discovered by Barbara McClintock in the late 1940s, refuting the idea that genomes are stable but are, on the contrary, dynamic entities [4]. TEs are highly diverse and an unified classification system for eukaryotic TEs has been proposed, establishing two TE classes according to their transposition mechanisms, structures and similarities [5]. Class I (Retrotransposons) groups elements moving by a copy-paste mechanism through an RNA that is reversed transcribed. Class I is composed of several TE orders, named LTR, DIRS, PLE, LINE and SINE. Class II (DNA transposons) is composed of TEs using different cut-paste mechanisms to transpose. These elements are grouped into the orders TIR, Crypton, Helitron and Maverick. Although intact retrotransposons and DNA transposons are autonomous elements that can move by themselves, SINE elements are non-autonomous TEs and rely on LINE for their mobility, even though their origin is distinct. Other non-autonomous elements can also be distinguished. LTR elements can degenerate into non-coding structures known as LARD (> 4 kbp) or TRIM (< 4 kbp), and TIR elements can also degenerate into non-coding structures known as MITE. LARD, TRIM and MITE elements have intact termini and can thus move by exploiting the molecular machinery of related autonomous TEs [6]. Genomes also contain highly diverged TE fossils, accumulated over time and having no mobility capacity. Due to their mobility, TEs generate mutations in their host genome through new insertions/deletions and participate in genome evolution by impacting the DNA sequence, genome size [7, 8] and chromosome structure [9]. TE activity is known to be triggered during stressful events and, while the majority of transpositions are neutral or harmful to the organisms, transposition events are recognized to promote beneficial mutations [10]. New TE insertions can impact gene function and gene regulation [11]. They can also create new genes and participate in the rise of new phenotypes. The role of TEs has been widely studied in animals [12], land plants [13] and insects [14, 15], but work on their impact on microalgal genomes is only just beginning [16–19]. Microalgae form a diverse polyphyletic group composed of eukaryotic, unicellular and multicellular, photosynthetic organisms [20]. They live in all aquatic habitats whether these have fresh, brackish or salt water and have colonized different extreme habitats, ranging from hot springs, high altitude streams, ice sheets and desert sand crusts, highlighting their

evolutionary ability to adapt to broad range of ecosystems [21–25]. Currently around 150,000 species of algae have been described (<http://www.algaebase.org>), but the number of non-described species is likely to number hundreds of thousands or millions of species [26]. They are divided among different eukaryotic phyla, in Archaeplastidia (green and red lineage), Rhizaria, Alveolates, Stramenopiles (brown lineage), Cryptophytes, Haptophytes and Excavates [27]. Despite their high number and diversity, few genome-wide TE annotations have been performed for microalgae. For the green lineage, this task was realized for ten Chlorophyte species [28–37]. For the red lineage, TE annotation was only done for the Rhodophyte *Cyanidioschyzon* sp. [38]. TEs were annotated in three diatom genomes (brown lineage) [18, 39–41] and also in five dinoflagellate species [42–46]. In Haptophytes, TE annotation has only been performed for one species [47]. These studies reveal that the TE content of microalgae genomes is diverse and includes both retrotransposons and DNA transposons.

Concerning TE activity in microalgae, a few studies have reported evidence of expression or transposition events. Expression of two LTR/Copia families was identified under nitrate starvation or exposure to diatom-derived reactive aldehydes in the diatom species *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* [18]. Moreover, expression of LTR/Copia or TIR/Mariner elements was also reported under thermal stress in *P. tricornutum*, *Amphora acutiuscula*, *Amphora coffeaeformis* and *Symbiodinium microadriaticum* [16, 48–50]. Evidence of transposition events was only identified for a MITE element in a clone of *Chlamydomonas reinhardtii* in the presence of vitamin B₁₂, resulting in a new phenotype [17].

Concerning TE activity in Haptophytes, we previously compared genetic polymorphisms between genomes of several strains of *Tisochrysis lutea* [51]. We identified new insertions/deletions and suggested the implication of autonomous TEs in the genome evolution of this species. In this context, the goal of the present study was to inventory TEs in the *T. lutea* genome and to identify potentially autonomous TEs. This marine microalga is commonly used as a feed in aquaculture [52] and is particularly studied for biotechnological applications such as food and biofuel production [53, 54]. In addition, several domesticated strains of *T. lutea* have been obtained with different processes [55] and a large amount of omics data has been collected [51, 56–60].

In this study, we present a detailed TE annotation of the *T. lutea* genome. To achieve this, we designed a new pipeline named PiRATE (Pipeline to Retrieve and Annotate Transposable Elements). The efficiency of the detection and classification steps of PiRATE was evaluated with data of the model species *Arabidopsis thaliana*.

Moreover, to be as exhaustive as possible about the repeated content of *T. lutea*, a new genome assembly was performed by combining Pacific Bioscience and Illumina data. Finally, available transcriptomic and proteomic data were used to reveal potential active TE families.

Results

PiRATE: Pipeline to Retrieve and Annotate Transposable Elements of non-model organisms

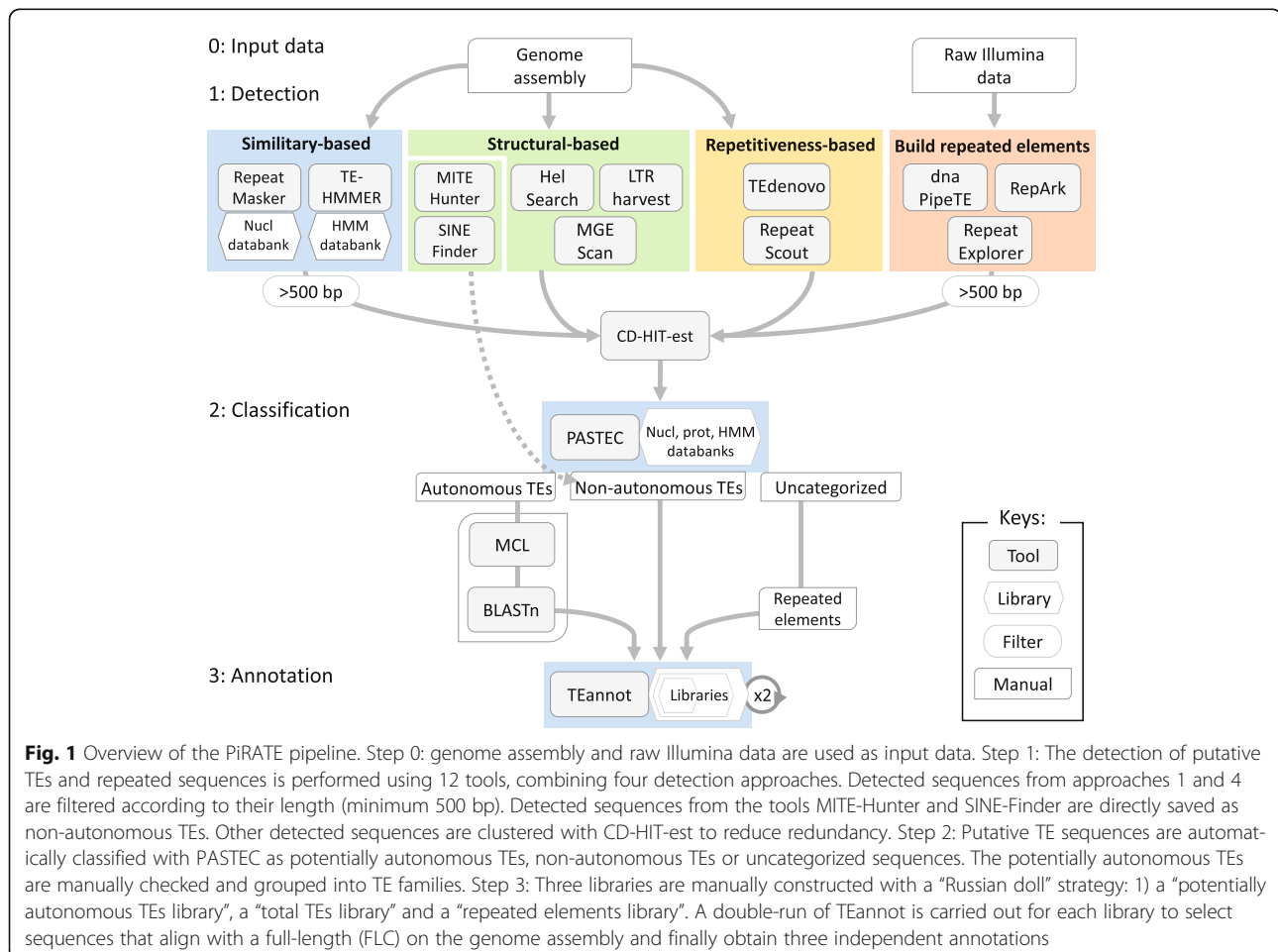
The goal of the present study was to inventory the TE content of the *T. lutea* genome and study the activity of potentially autonomous TEs. Annotation of TEs is a challenging task because of their diversity, their repetitive nature and the complexity of their structures (i.e. GC-rich regions, homopolymers and repeated motifs). Numerous tools have been designed to identify TEs (Additional file 1: Table S1), which can be grouped into four approaches according to their TE detection method: (1) similarity-based detection such as RepeatMasker [61], (2) structure-based detection such as MITE-Hunter [62], (3) repetitiveness-based detection such as

RepeatScout [63], and (4) tools building repeated elements from unassembled data such as dnaPipeTE [64].

Currently, the tool used most frequently to perform a TE annotation is RepeatMasker, which provides a rough estimation of the TE content in a genome assembly [61, 65]. However, this tool compares the genomic sequences with a databank of known TEs to realize the annotation and is therefore not suitable for realizing a de novo TE annotation [65–67]. To perform a de novo TE annotation, pipelines employing repetitiveness-based methods of detection, such as RepeatModeler and REPET, are commonly recommended [66–69]. Here we built PiRATE (Fig. 1) to conduct a de novo TE annotation in the genome of non-model species *T. lutea*. PiRATE is composed of detection, classification and annotation steps.

Detection of TEs

To date, genome assembly of non-model organisms has usually not been performed at the level of complete chromosomes but is instead highly fragmented. This fragmentation is recognized to be partly the result of a bad assembly of the TE copies due to their high



repetitive content, which increases the difficulty of their detection [70]. The optimization of the detection step of PiRATE was therefore a priority. We made an overview of tools related to TE detection (Additional file 1: Table S1) and 12 tools were selected according to the specificity and efficiency of their algorithms. These tools represent the four major TE detection approaches (presented above), so as to be as exhaustive as possible. Combining tools is recognised to improve TE detection efficiency [66, 67, 71]. We then applied a clustering method to decrease the redundancy of the detected sequences, by selecting the larger detected sequences of each cluster. The goal of this step was to promote the detection of full-length TE sequences. The detection of complete TE sequences bearing recognizable conserved domains or specific structures and motifs makes the classification step easier. Moreover, a complete TE sequence indicates a potentially autonomous element.

Classification of TEs

The classification step of PiRATE is performed by PASTEC [72], which partly uses databanks of known TEs to establish an automated classification of the detected sequences. To improve the classification step of PiRATE, its default databanks were upgraded, by adding 1240 TE sequences from other public databanks, non-inventoried algal TEs and by building 78 new profile HMMs (Hidden Markov Model). Adding non-inventoried data is important for improving the TE classification of species belonging to poorly studied phyla, which often have few described TEs in the databanks. This is common for numerous microalgal phyla (i.e. Haptophyta, Euglenophyta and Dynophyta). In our case, only 17 TE families belonging to the Haptophyte phylum are present in the most frequently used and complete TE databank Repbase [73, 74]. We also estimated that only 2609 TE families are described for microalgal taxa in Repbase. Compared with other taxa, this number is very low, for examples 29,503 TE families are described for Metazoa and 12,620 for Viridiplantae (Repbase, 10/29/2017). The putative TE sequences are classified following the Wicker et al. classification [5] and can be grouped as 1) potentially autonomous TEs, 2) non-autonomous TEs or 3) uncategorized sequences. Because we were interested in potentially autonomous TEs, these sequences were manually checked and grouped into families.

Annotation of TEs

For the annotation step, we built three libraries in order to then apply a method that we named “Russian doll”, due to its nesting strategy (Additional file 1: Figure S1). We built a “potentially autonomous TEs library” containing checked potentially autonomous TEs, a “total TEs library” also containing the non-autonomous TEs,

and a “repeated elements library” also containing the uncategorized repeated sequences. These nested libraries made it possible to perform several independent annotations in order to avoid a competition effect among sequences aligning on the same genomic regions.

Evaluation of PiRATE with *A. thaliana* genomic data

Evaluation of the detection step

The detection and classification steps of PiRATE were evaluated to highlight their strengths and weaknesses. This evaluation made it possible to define suitable rules for the manual check step. As a control, we used 359 consensus sequences of the described TE families of *A. thaliana*, available in Repbase. Genomic data of the model plant *A. thaliana* provided a suitable control because of its high quality genome assembly and high TE diversity. Class I and Class II *A. thaliana* TE families are well described for both autonomous and non-autonomous TEs. Detected sequences covering less than 40% of the full-length of a consensus sequence were considered too short to be efficiently classified and were not taken into account. The proportion of TE families detected with a complete length (coverage score $\geq 70\%$) or detected with at least a partial length (coverage score $\geq 40\%$) is given in Fig. 2a. PiRATE detected $\sim 81\%$ (292/359) of the TE families described in *A. thaliana* genome (Fig. 2a). PiRATE was especially effective for detecting sequences belonging to LTR (96%), LINE (79%), non-autonomous TIR (81%) and non-autonomous Helitron (94%) (Fig. 2a). It had a good efficiency for detecting TIR (62%) and Helitron (60%). However, it had difficulty detecting SINE elements (27%) (Fig. 2a). In addition, we compared the detection step of PiRATE to TEdenovo [68], LTRharvest [75], RepeatScout [63], RepeatMasker [61], dnaPipeTE [64], RepeatExplorer [76] and RepARK [77] (Fig. 2b). Overall, the detection step of PiRATE detected the highest percentage of TE families of *A. thaliana*. Compared to TEdenovo, which displayed the second highest percentage of detected TE families, PiRATE detected 21 additional TE families (+ 6%) (Fig. 2b and Additional file 1: Figure S2). PiRATE was particularly more effective for detecting LINE (+ 32%) and TIR (+ 10%) (Additional file 1: Figure S2).

Evaluation of the classification step

To evaluate the classification step of PiRATE, we used the 292 sequences detected by PiRATE during the evaluation of the detection step, which represent the largest detected sequences of the 292 TE families of *A. thaliana*. These 292 sequences were classified with PASTEC using the PiRATE databanks (excluding data from *Arabidopsis* species). To estimate the classification efficiency, we counted the number of detected TEs with correct classification at the order level and the number of sequences

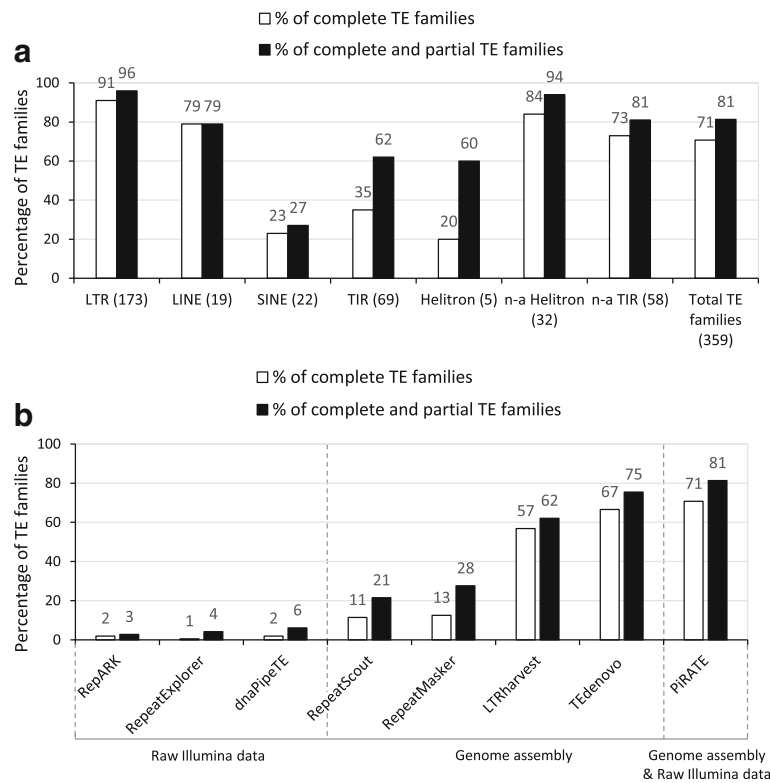


Fig. 2 Evaluation of the detection step of PiRATE with genomic data of *Arabidopsis thaliana*. **a**) Percentage of TE families detected in *A. thaliana*. For each TE order (x-axis) is indicated the percentage of TE families detected with a complete length (coverage score $\geq 70\%$, white bars) or detected with a partial and a complete length (coverage score $\geq 40\%$, black bars). The x-axis indicates the number of TE families for each order; “n-a” means non-autonomous. **b**) Comparison of the percentage of TE families of *A. thaliana* detected by PiRATE (Step 1), RepARK, RepeatExplorer, dnaPipeTE, RepeatScout, RepeatMasker, LTRharvest and TEdenovo. For each tool is indicated the percentages of TE families of *A. thaliana* detected with a complete length (coverage score $\geq 70\%$, white bars) or detected with a partial and a complete length (coverage score $\geq 40\%$, black bars). The x-axis indicates the tools and nature of the input data

that had an incorrect classification or that were uncategorized. We observed that 75% (218/292) of the detected TEs were correctly classified, 7% (21/292) were incorrectly classified and 18% (53/292) were uncategorized. The classification step of PiRATE was therefore efficient at correctly classifying autonomous TEs belonging to LTR (98%), LINE (87%), TIR (91%) and Helitron (100%) but had difficulty correctly classifying SINE (50%), non-autonomous TIR (27%) and non-autonomous Helitron (7%) (Additional file 1: Figure S3). Taking into account all of the above results, PiRATE is efficient enough to detect and correctly classify the majority of the autonomous TE families.

A new genome assembly of *T. lutea* to improve the TE annotation

We recently published a draft genome assembly of *T. lutea* obtained with Illumina short-read technology [51]. To obtain an improved genome assembly, the genome of *T. lutea* was re-sequenced with Pacific Bioscience long-read technology. A new genome assembly was

performed from the long reads and was improved with the Illumina short-read data, used to build the draft genome assembly [51]. The new genome assembly of *T. lutea* is composed of 193 contigs and is 82 Mb in size. A gain of around 30 Mb was obtained (+ 34%), compared with the previous 54 Mb genome assembly, which was composed of 7659 contigs [51]. The size of the coding regions increased slightly between these genome versions. While the new genome assembly encodes for 15,972 genes, corresponding to a coding region length of 32 Mb, the gene proportion of the previous draft genome version was 25 Mb, suggesting that the new assembled regions are mostly repeated elements. This new larger version of the genome seems to incorporate more assembled TEs.

Effect of genome quality on TE detection approaches

To estimate the contribution of each TE detection approach of PiRATE depending on the level of fragmentation of the genome assembly, the detection step (Fig. 1) of PiRATE was applied with raw Illumina data of *T.*

lutea and, either the draft genome version of *T. lutea* (7659 contigs) [51] or the new genome assembly of *T. lutea* (193 contigs). In both cases, the detected sequences were compared to the referent sequences of the TE families of *T. lutea* (described below). For each TE detection approach in PiRATE, we counted the number of *T. lutea* TE families detected, with the largest length (i.e. the most complete sequences, having the highest percentage of coverage compared to the reference TE sequences) and divide this number by the total of detected TE families. This provided an estimation ratio of the contribution of each TE detection approach depending on the input data (Fig. 3). With both types of dataset, the similarity-based approach had the weakest percentage and contributed to detecting only 2 or 3% of the *T. lutea* TE families. Using the draft genome assembly and the raw Illumina data, the structural-based approach contributed to detecting 1% of the TEs families of *T. lutea*, but 20% of the TE families of *T. lutea* with the new genome assembly and the raw Illumina data (Fig. 3). The repetitiveness-based approach contributed to detecting 7% of the TE families of *T. lutea* with the draft genome assembly and the raw Illumina data. However, it was the most efficient approach with the new genome and contributed to detecting 63% of the *T. lutea* TE families (Fig. 3). When a draft genome assembly is used as input, the fourth detection approach, using raw Illumina data to build repeated elements, was the most useful approach and contributed to detecting 67% of the TE families (Fig. 3).

Annotation of the repeated elements content of the *T. lutea* genome

We applied PiRATE to the new genome assembly of *T. lutea* and raw Illumina data. After the classification step,

we manually curated the sequences as potentially autonomous TEs, non-autonomous TEs or uncategorized repeated elements. Because we were interested in characterizing their activity, the potentially autonomous TEs were manually checked and grouped into families (see [Methods](#)). We identified six potentially autonomous families of LTR/Copia and four families of LTR/Gypsy (Table 1). We found 14 potentially autonomous families of LINE elements, similarly close to Tx1 elements, belonging to the L1 superfamily [78, 79]. We identified seven potentially autonomous families of TIR/Harbinger, six families of TIR/PiggyBac and eight families of TIR/Mariner. A high number of potentially autonomous hAT elements were detected. Due to their divergence, they were grouped into 129 putative families.

Three annotations were conducted with three nested libraries (Additional file 1: Figure S1). From the “potentially autonomous TEs library” composed of 240 referent sequences, we estimated that the proportion of the potentially autonomous TEs represent 3.79% of the *T. lutea* genome (Table 1). The annotation of the TE content was performed with the “total TEs library” containing 459 supplementary sequences corresponding to 14 sequences of potential SINE elements, 188 sequences of potential MITE, 240 sequences of potential TRIM and 17 sequences of potential LARD (Table 1). From this annotation, we estimated that the genome of *T. lutea* contains 20.84% of potentially autonomous and non-autonomous TEs (Table 1 and Additional file 1: Figure S4). Class I and Class II TEs represent 15.89 and 4.95%, respectively (Table 1). We found a large quantity of Gypsy (4.65%), LINE (3.87%) and hAT (2.12%) copies, suggesting ancient burst events for these elements (Table 1). We established that the proportion of non-autonomous TEs is 17.05% (Table 1). Then, we performed the annotation

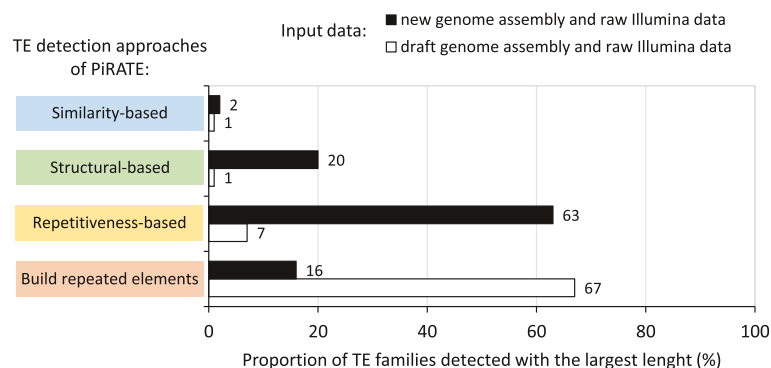


Fig. 3 Comparison of the contribution of the four TE detection approaches of PiRATE on the detection of the TE families of *Tisochrysis lutea*, depending on the input data. For each TE detection approach, we calculated the number of TE families detected with the largest length and divide this number by the total of detected TE families of *T. lutea*. The input dataset was either the draft genome assembly of *T. lutea* and raw Illumina data of *T. lutea* (white bars) or the new genome assembly of *T. lutea* and raw Illumina data of *T. lutea* (black bars). The similarity-based detection, structural-based detection and the repetitiveness-based detection use a genome assembly as input data. The last approach builds repeated elements from raw Illumina data

Table 1 Diversity and proportion of transposable element orders and classes in the genome assembly of *Tisochrysis lutea*. The abbreviations “a” and “n-a” indicate autonomous and non-autonomous transposable elements respectively

		Orders/ Superfamilies	Number of families (f) or detected sequences (s)	Number of potentially autonomous TEs	Proportion of the potentially autonomous TEs (%)	Proportion of total genome (%)
Class I	a	LTR/Copia	6 f	45	0.37	1.09
		LTR/Gypsy	4 f	242	2.56	4.65
		LINE/L1	14 f	59	0.25	3.87
	n-a	SINE	14 s			0.04
		LTR/LARD	17 s			0.76
		LTR/TRIM	240 s			5.48
Total Class I					15.89	
Class II	a	TIR/hAT	129 f	145	0.41	2.12
		TIR/Mariner	8 f	41	0.11	0.19
		TIR/Harbinger	7 f	26	0.05	0.34
		TIR/PiggyBac	7 f	14	0.04	0.26
	n-a	MITE	188 s			2.04
	Total Class II					4.95
Total TEs				572	3.79	20.84

of every repeated element by using the “repeated elements library” containing an additional 2680 uncategorized repeated sequences. From this annotation, we estimated that 17.79% of the *T. lutea* genome is represented by uncategorized repeated elements (Additional file 1: Figure S4). To estimate the proportion of the simple tandem repeats, we used the tool RepeatMasker and found that they made up 5.97% of the genome assembly of *T. lutea* (Additional file 1: Figure S4). By adding together the proportions of all the annotated repeats, we estimated that the total proportion of repeated elements in the *T. lutea* genome was 44.6%. Knowing that the coding gene proportion is of 38.49%, we estimated that 16.91% of the genome is non-characterized (Additional file 1: Figure S4).

Discovery of potentially active autonomous TEs in the *T. lutea* genome

In this study we chose to focus on the identification of potentially autonomous TEs to reveal potentially active elements. From the annotation obtained with the “potentially autonomous TEs library”, we performed the cartography of the 572 annotated TEs that are potentially autonomous (Fig. 4).

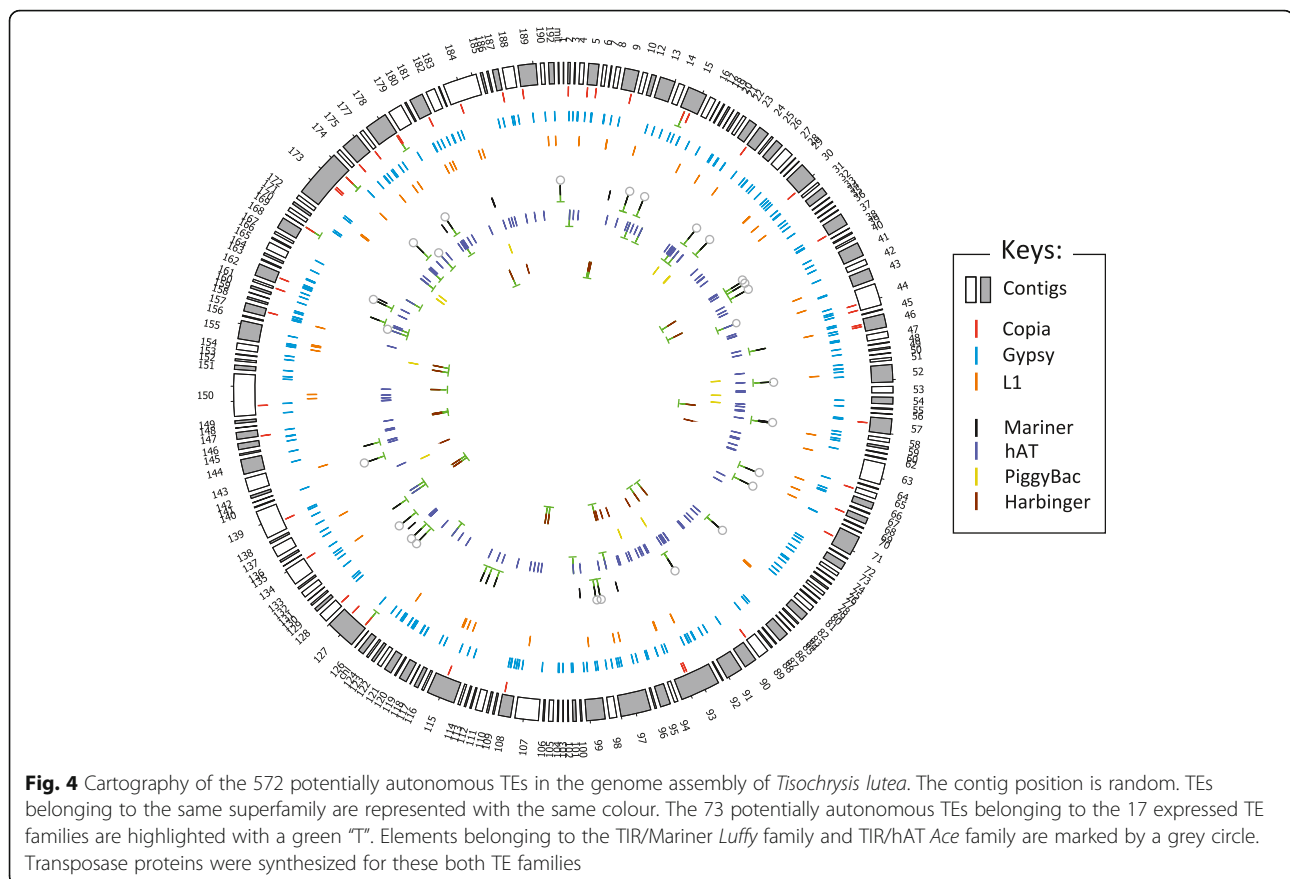
To identify potentially active TEs and have an estimation of the genome dynamic of *T. lutea*, transcriptomic data were mapped on the new genome assembly and crossed with the annotation of the 572 potentially autonomous TEs. Expression was identified for 17 TE families: one LTR/Copia, four TIR/Mariner, four TIR/Harbinger and eight TIR/hAT. These families represent 73 potentially autonomous TEs and their genomic position is illustrated in Fig. 4 and is indicated in

Additional file 2. Putative ancient transpositions were studied by looking for similarities between copies belonging to these 17 expressed TE families (Additional file 2). We identified that the Mariner-3 family is composed of 24 highly similar copies, which share a mean pairwise identity of 99.7%. Among them, 20 copies seem to be complete (Additional file 3). This high number of similar copies suggests that this family was/is active. The hAT-2 family is composed of three highly similar copies that share a mean pairwise identity of 99.8%. Moreover, eight similar copies were identified for the Harbinger-6 family and five similar copies for the Copia-3 family. Other details can be found in Additional file 2. TE copies belonging to these 17 expressed TE families were submitted to BLASTx on proteomic data of *T. lutea*, that we previously obtained under nitrogen limitation [58]. We identified that transposase proteins were synthesized for the Mariner-3 family and the hAT-2 family. The transposases of the Mariner-3 and hAT-2 families match with six and 36 peptides, respectively. The alignments with the matching peptides can be found for both families in Additional file 4. From transcriptomic data of a previous study, we highlight that these families were among the three higher expressed genes [58].

Discussion

PiRATE: Pipeline to Retrieve and Annotate Transposable Elements of non-model organisms

The goal of the present study was to inventory the TE content in the genome of *T. lutea* genome and study the activity of potentially autonomous TEs. We built PiRATE to counter the lack of knowledge about TEs in Haptophytes and the difficulty of identifying TEs in a



fragmented genome assembly [70]. The detection step of PiRATE has been optimized to promote the detection of full-length TE sequences and its classification step has been improved for algal genomes. The detection step of PiRATE combines multiple, existing analysis packages representing all major approaches for TE detection. The detection step of PiRATE was evaluated with genomic data of *A. thaliana* and compared to TEdenovo [68], LTRharvest [75], RepeatScout [63], RepeatMasker [61], dnaPipeTE [64], RepeatExplorer [76] and RepARK [77] (Fig. 2b). Overall, the detection step of PiRATE detected the highest percentage of TE families (81%) with a partial and complete length compare to the other tools used alone (Fig. 2b). This confirms that the combining of multiple tools, using several approaches improves the detection of different TE families, with complete sequences, as previously indicated [66, 67, 71]. In this comparison, TEdenovo was efficient and displayed the second highest percentage of detected TE families (75%) (Fig. 2b). LTRharvest also showed a good capacity to detect TE families of *A. thaliana* (62%) (Fig. 2b). This is due to the high content of LTR elements in the *A. thaliana* genome and because this tool detected TE families belonging to other TE orders. In this comparison, the least effective tools were RepARK (3%), RepeatExplorer

(4%) and dnaPipeTE (6%), which used raw illumina data as input (Fig. 2b). This is not surprising considering the challenge of building repeated elements from raw Illumina data, compared to the other tools using the complete genome assembly of *A. thaliana*.

A new genome assembly of *T. lutea* to improve the TE annotation

We recently published a draft genome assembly of *T. lutea* obtained with Illumina short-read technology [51]. While this technology has a very low sequencing error rate, its use alone often leads to fragmented assemblies, especially in TE-rich genomes, due to the incapacity of short-reads to entirely span repetitive elements [80]. To obtain an improved genome assembly, the genome of *T. lutea* was re-sequenced with Pacific Bioscience long-read technology and the assembly was corrected with short-read Illumina data. Indeed, the use of long-reads leads to a more complete and accurate assembly of long repeated elements such as TEs [81–83]. However, to date, this technology has a high sequencing error rate and its combination with short-read Illumina data has become a common way of partially overcoming this problem [84–86]. Compare to the previous draft genome

assembly, this new genome assembly is larger, less fragmented and seems to incorporate more assembled TEs.

Effect of genome quality on TE detection approaches

To estimate the contribution of the four TE detection approaches of PiRATE depending on the level of fragmentation of the genome assembly, the detection step (Fig. 1) of PiRATE was applied with raw Illumina data of *T. lutea* and, either the draft genome version of *T. lutea* (7659 contigs) [51] or the new genome assembly of *T. lutea* (193 contigs). The four TE detection approaches showed different contribution according to the level of fragmentation of the genome assembly (Fig. 3). By gathering these four detection approaches, PiRATE improves the TE detection of organisms having a genome assembly which is highly fragmented.

Annotation of the repeated elements content of the *T. lutea* genome

With PiRATE, we established that the total proportion of repeated elements in the *T. lutea* genome is represented by 20.84% of TEs, 17.79% of uncategorized repeated elements and 5.97% of simple tandem repeats (Additional file 1: Figure S4). The high percentage of uncategorized repeated elements could indicate the presence of unknown TEs. A high number of uncategorized sequences (30.9%) was also reported in the *Emiliania huxleyi* genome [40]. Here, we choose to focus on the identification of potentially autonomous TEs to reveal potentially active elements. The proportion of the potentially autonomous TEs represents 3.79% of the *T. lutea* genome, corresponding to 572 annotated TEs (Fig. 4). Interestingly, we found a potentially autonomous TIR/Mariner in the predicted mitochondrial genome and a potentially autonomous LTR/Copia and TIR/hAT in the predicted chloroplast genome.

Identification of potentially active TEs in *T. lutea*

Few studies have investigated TE activity in microalgal genomes and their role is poorly known. Regarding Class I TEs, some studies reported expression of LTR elements in dinoflagellate and diatom species under thermal stress or nitrogen limitation [16, 48–50]. Concerning Class II elements, a previous study reported a case of phenotypic evolution for the microalga *Chlamydomonas reinhardtii* caused by the transposition of a MITE in the presence of vitamin B₁₂ [17]. In the present study, we identified 17 expressed TE families and, among these, a TIR/Mariner *Luffy* and a TIR/hAT *Ace* family were able to synthesize their transposase under nitrogen starvation [58]. We highlight the presence of highly similar copies (Additional file 3) suggesting that these elements are able to transpose or could be able to trigger the transposition of potential derived MITE elements. Although

we cannot draw conclusions about their mobility, the investigation of the TE expression is a good indicator of the potential activity of TEs. Nitrogen limitation has been previously described as a stress condition in the diatom *Phaeodactylum tricornutum*, triggering overexpression of the LTR/Copia family named *Blackbeard* [18]. Although we cannot draw conclusions about de novo insertions, the evidence presented here indicates that these both TEs families are suitable candidates for mobility and could participate in the genome evolution of *T. lutea*.

Conclusion

Genome-wide TE annotation has rarely been performed in microalgae compared with animals, insects and land plants. This study opens the way to new searches about the role of TEs in the genome evolution of *Tisochrysis lutea* and their contribution to the microalgal adaptation process. In the present study, we built PiRATE to counter the lack of knowledge about TEs in Haptophytes and the difficulty of identifying TEs in a fragmented genome assembly. With PiRATE, we conducted a genome-wide detection and annotation of the repeated elements in a new genome assembly of *Tisochrysis lutea* and established that it is composed of 3.8 and 15.95% of potentially autonomous and non-autonomous TEs, respectively. The annotation of the potentially autonomous TEs was crossed with transcriptomic and proteomic data and evidence of expression was identified for 17 TE families. Among these, we discovered that transposase proteins were synthesized for both a Mariner (*Luffy*) and a hAT (*Ace*) family. Both these families have several highly similar copies throughout the genome and were among the three highest expressed genes in a previous transcriptomic study. All of this suggests that both these families could be able to transpose themselves or trigger the transposition of potential derived MITE elements.

Methods

Microalga strain and culture conditions

The *T. lutea* strain was provided by the Culture Collection of Algae and Protozoa (CCAP 927/14). This strain was isolated by Haines in the late 70s and stored in the algae bank. The strain was grown in two 1-L flasks, bubbled with 0.22 mm filtered-air. The culture was maintained at a constant temperature of 21 °C, under a constant irradiance of 50 $\mu\text{mol m}^{-2} \text{s}^{-1}$.

DNA extraction, sequencing, genome assembly and gene annotation

Total DNA was extracted from the *T. lutea* WT-strain using a phenol/chloroform protocol. DNA quality and concentration were assessed with gel electrophoresis and

Qubit Fluorometric Quantitation (ThermoFisher, Massachusetts, USA), respectively. *T. lutea* genome sequencing was performed with a PacBio RSII sequencer (Pacific Bioscience, California, USA) at the Plateforme GeT PlaGe (Toulouse, France); seven SMRT cells were performed. Filtered subreads were assembled using Canu1.3 [82]. The assembly was polished with Quiver (<https://github.com/PacificBiosciences/GenomicConsensus>) and its accuracy was improved with Pilon [87] using previous Illumina Hiseq mate-pair reads of *T. lutea* ([51]; SRA: SRR3156597). The annotation of the coding-gene region was performed with the pipeline MAKER2 [88–91].

TE annotation in the *T. lutea* genome using PiRATE

Step 1: TE detection

The new genome assembly of *T. lutea* and previous raw Illumina data ([51]; SRA: SRR3156597) were used as input. Putative TE sequences were detected using four approaches (Fig. 1). The first approach was represented by two tools using similarity-based detection: RepeatMasker (setting: -s, -no_low, -lib; with the PiRATE nucleotide databank; [61]) and TE-HMMER (with a homemade profile HMMs databank). TE-HMMER is a homemade tool using HMMER (default setting, [92]) and tBLASTn (setting: -evalue 10E-300, [93]). The second approach consisted of five tools using structural-based detection: LTRharvest (default setting, [75]), Helsearch (default setting, [94]), MGEScan-nonLTR (default setting, [95]), MITE-Hunter (default setting, [62]) and SINE-Finder (default setting, [96]). The third approach combines tools using repetitiveness-based detection: TEde novo (steps 1 to 4, default setting, [68]) and Repeat Scout (default setting, [63]). These tools cluster repeated sequences from a genome assembly to build consensus sequences. The last approach was composed of tools performing the assembly of repeated sequences from raw Illumina data (fasta or fastq). We used RepARK (default setting, [77]), dnaPipeTE (setting: %coverage: 0.6, [64]) and RepeatExplorer (setting: -paired, [76]). The sequences detected by the first and the last approaches that were below 500 bp in length were removed with a perl script. The sequences detected with SINE-Finder and MITE-Hunter were directly saved for the second step. Other detected sequences were concatenated into a single FASTA file and clustered with CD-HIT-est (settings: -aS 1 -c 1 -r 1 -g 1 -p 0, [97]) to reduce the redundancy. This made it possible to remove shorter sequences that aligned with 100% of identity on a part of the larger sequences.

Step 2: TE classification

In the second step, sequences were automatically classified with PASTEC [72], following the Wicker et al.

classification system [5]. This tool was improved with custom databanks (described below). Three libraries were manually constructed with a “Russian doll” strategy in order to perform separate annotations (Additional file 1: Figure S1): a “potentially autonomous TEs library”, a “total TEs library” containing the potentially autonomous TEs and the non-autonomous TEs and a “repeated elements library” also containing the uncategorized repeated sequences. Sequences classified as LTR, LINE and TIR were manually sorted by superfamily (according to the evidence section produced by PASTEC). To facilitate their manual check, sequences belonging to the same putative superfamily were grouped into families with MCL (MCL_inflation: 1.5; MCL_coverage: 0). The percentage of identity between sequences belonging to the same family were checked with Blastn (-identity: 80%). We followed the 80–80–80 Wicker rules to form families [5]. Finally, larger sequences from each TE family were checked and selected for the “potentially autonomous TEs library” according to the presence of TE domains or similarities with Pfam (<http://pfam.xfam.org>), NCBI-BLASTx and Censor (<http://www.girinst.org/censor>). We defined as potentially autonomous LTR, sequences bearing at least a reverse transcriptase and an integrase domain and having similarity to known LTR elements. We defined as potentially autonomous LINE, sequences bearing at least a reverse transcriptase domain and sharing similarity to known LINE elements. We defined as potentially autonomous TIR, sequences with evidence of a transposase domain or similarity with known TIR elements.

No manual checks were performed for sequences classified as non-autonomous TEs. Sequences classified as SINE, MITE and TRIM were directly selected for the “total TEs library”. Only sequences classified as LARD, which were obtained with the repetitiveness-based approach of TE detections (TEde novo or Repeat Scout), were selected. Sequences detected by SINE-Finder and MITE-Hunter were also directly selected for the “total TEs library”. Finally, the sequences classified as noCat (uncategorized) and obtained with the repetitiveness-based approach at the TE detection step were selected for the “repeated elements library”.

Step 3: TE annotation

Three libraries were built (Additional file 1: Figure S1): 1) a “potentially autonomous TEs library” 2) a “total TEs library” and 3) a “repeated elements library”. A first run of TEannot ([68], default setting, steps 1, 2, 3, 7 and 8) was performed for each library to known sequences matching with a full-length size on the genome (FLC sequences) and to remove potential chimeric data. A second run of TEannot was performed with these FLC sequences for each of the final libraries (default setting,

steps 1, 2, 3, 4, 5, 7 and 8) and three annotations were obtained.

Proportion of TEs and repeated elements in *T. lutea*

From the annotation file obtained with the “potentially autonomous TEs library”, we manually selected 572 sequences and calculated their proportion in the genome of *T. lutea*. TEs. The different criteria used are detailed in Additional file 1: Method S1 and Table S2. An illustration of the position of these sequences on the *T. lutea* genome assembly was built with the tool Circos [98]. The annotations obtained with the “total TEs library” and the “repeated elements library” were used to estimate the total proportion of TEs and to calculate the proportion of uncategorized repeated elements in the genome of *T. lutea*. Details on the method are available in Additional file 1: Method S2 and Table S3. The proportion of simple repeats was calculated with the tool RepeatMasker (setting: -s -noint -no_is, [61]).

PiRATE databanks

Nucleotide and protein databanks

A nucleotide and a protein databank of TEs were built with sequences from Repbase (REPET version 20.05, <http://www.girinst.org/repbase>), the P-MITE database (<http://pmite.hzau.edu.cn>) and SINE base (<http://sines.eimb.ru>). Because algae originally arose from the predation of a cyanobacterial organism by a eukaryotic heterotrophic organism, cyanobacterial TE sequences were also added from the IS-finder database (<http://www-is.bio-toul.fr>) (Additional file 5). Moreover, we added non-inventoried TEs of microalgae and macroalgae, retrieved from the NCBI database (Additional file 5).

Profile HMMs databank

A homemade databank of profile HMMs was built with sequences of the protein databank. Multiple protein alignments were performed with Clustal Omega (<http://www.ebi.ac.uk/Tools/msa/clustalo/>). When possible, TE protein sequences from algae were favoured. 78 profile HMMs were performed with the HMMbuild tool of HMMER [92] for 62 TE categories displayed on the Browse Repbase tool (<http://www.girinst.org/repbase/update/browse.php>). This databank was used with TE-HMMER at the detection step. At the classification step, we combine this databank with the default databank of PASTEC (ProfilesBankForREPET_Pfam27.0_GypsyDB.hmm, <https://urgi.versailles.inra.fr/download/repet>).

Evaluation of PiRATE

The efficiency of the detection and classification steps of PiRATE were evaluated with genomic data of the model plant *A. thaliana*. We used the genome assembly TAIR10 available on the TAIR project (https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGenes%2FTAIR10_genome_release%2FTAIR10_chromosome_files) and the raw Illumina data available at the 1001 genome project <http://1001genomes.org/data/SLU/SLUHenning2014/releases/current/strains/Seattle-0>). These data of *A. thaliana* were submitted to the step 1 of PiRATE (RepeatMasker and TE-HMMER were used without data from *Arabidopsis* species in the databanks). The detected sequences were submitted to PASTEC [72] and compared to the 359 TE families described in *A. thaliana* and available on Repbase (<http://www.girinst.org/repbase>). We didn't include the terminal repeated sequences of the LTR TE families and the heterologous TE named DRL1. From the classification file, we selected each of the sequences matching to a TE consensus sequences of *A. thaliana*. Those covering less than 40% of the full-length of a consensus sequences were considered as too short to be efficiently classified and were not taken in account. We considered as a partial or complete detection the detected sequences covering at least 40% or 70% of the full-length of a consensus TE family of *A. thaliana*, respectively. For the comparison of the detection step of PiRATE with TEdenovo [68] (steps 1 to 4, with LTRharvest [75]), LTRharvest [75], RepeatScout [63], RepeatMasker [61], dnaPipeTE [64], RepeatExplorer [76] and RepARK [77], the number of detected TE families was calculated with the same method previously described for the evaluation of the PiRATE detection step. For the evaluation of the classification step of PiRATE, we used the longest detected sequences of the 292 TE families detected by PiRATE during the evaluation of the detection step as a control. These 292 sequences were classified with PASTEC using modified versions of the three PiRATE databanks (Nucleotide, protein and profile HMMs), without data from *Arabidopsis* sp. We calculated the percentages of correct classification, incorrect classification or uncategorized classification. Details on the impact of the genome assembly quality on the efficiency of the TE detection step of PiRATE are available in Additional file 1: Method S3.

arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGenes%2FTAIR10_genome_release%2FTAIR10_chromosome_files) and the raw Illumina data available at the 1001 genome project <http://1001genomes.org/data/SLU/SLUHenning2014/releases/current/strains/Seattle-0>). These data of *A. thaliana* were submitted to the step 1 of PiRATE (RepeatMasker and TE-HMMER were used without data from *Arabidopsis* species in the databanks). The detected sequences were submitted to PASTEC [72] and compared to the 359 TE families described in *A. thaliana* and available on Repbase (<http://www.girinst.org/repbase>). We didn't include the terminal repeated sequences of the LTR TE families and the heterologous TE named DRL1. From the classification file, we selected each of the sequences matching to a TE consensus sequences of *A. thaliana*. Those covering less than 40% of the full-length of a consensus sequences were considered as too short to be efficiently classified and were not taken in account. We considered as a partial or complete detection the detected sequences covering at least 40% or 70% of the full-length of a consensus TE family of *A. thaliana*, respectively. For the comparison of the detection step of PiRATE with TEdenovo [68] (steps 1 to 4, with LTRharvest [75]), LTRharvest [75], RepeatScout [63], RepeatMasker [61], dnaPipeTE [64], RepeatExplorer [76] and RepARK [77], the number of detected TE families was calculated with the same method previously described for the evaluation of the PiRATE detection step. For the evaluation of the classification step of PiRATE, we used the longest detected sequences of the 292 TE families detected by PiRATE during the evaluation of the detection step as a control. These 292 sequences were classified with PASTEC using modified versions of the three PiRATE databanks (Nucleotide, protein and profile HMMs), without data from *Arabidopsis* sp. We calculated the percentages of correct classification, incorrect classification or uncategorized classification. Details on the impact of the genome assembly quality on the efficiency of the TE detection step of PiRATE are available in Additional file 1: Method S3.

Transcriptomic and proteomic analyses

The expression analysis was performed using eight sets of previously published transcriptomics data [56, 58]. These data were concatenated and normalized using the tool `insilico_read_normalization.pl` of Trinity [99]. Reads were then mapped on the new genome assembly of *T. lutea* with TopHat [100] and crossed with the annotation of the potentially autonomous TEs. HTseqCount [101] was used to count the number of mapped reads for each potentially autonomous TEs. With a homemade script we retrieved the TE families with transcripts covering at least 90% of the annotated sequences. Sequences

of the TE copies of these expressed TE families were then compared with BLASTx to published proteomic data [58]. Sequence alignments of the peptides of the Mariner (*Luffy*) and hAT (*Ace*) elements on the predicted transposases were performed with ClustalOmega and visualized with Geneious (Additional file 4). With the global-alignment tool of Geneious [102] (setting: free end gaps), a mean pairwise identity was calculated for each expressed TE family having at least three annotated copies (Additional file 3).

PiRATE is automated through a stand alone Galaxy

All tools used in PiRATE are automated in a standalone Galaxy [103]. The PiRATE-Galaxy is available through a virtual machine at <https://doi.org/10.17882/51795>. A tutorial file can be download.

Additional files

Additional file 1: Additional supporting information. This file contains the additional supporting figures, tables, results and materials and methods. (PDF 594 kb)

Additional file 2: Percentage of identity between copies of the expressed TE families. This file lists the percentage of identity between the TE copies of the 17 expressed TE families identified in the genome of *Tisochrysis lutea*. (XLSX 19 kb)

Additional file 3: Sequences alignment of TE copies of the TIR/Mariner *Luffy* family. This file contains the sequence alignment of the copies belonging to the TIR/Mariner *Luffy* described in the genome of *Tisochrysis lutea*. (PDF 17427 kb)

Additional file 4: Sequences alignment of the peptides matching on the predicted TE proteins. This file contains the alignment of the peptides matching on the predicted proteins of the TIR/Mariner *Luffy* and the TIR/hAT *Ace*. (PDF 996 kb)

Additional file 5: List of non-inventoried sequences added to the databanks used by the pipeline PiRATE. This file lists the non-inventoried sequences added to the databanks used by the pipeline PiRATE, they belong to algae and cyanobacteria. (XLSX 23 kb)

Abbreviations

HMM: Hidden Markov Model; LARD: Large Retrotransposon Derivative; LINE: Long Interspersed Element; LTR: Long Terminal Repeat; MITE: Miniature Inverted-repeat Transposable Element; PiRATE: Pipeline to Retrieve and Annotate Transposable Element; PLE: Penelope; SINE: Short Interspersed Nuclear Element; TE: Transposable element; TIR: Terminal Inverted Repeat; TRIM: Terminal-repeat Retrotransposons In Miniature

Acknowledgements

This work was supported by the French Region of Pays de la Loire with the Atlantic Microalgae program and the French Research Institute for Exploitation of the Sea (IFREMER). We thank the platform Genotoul GeT-PlaGe for the genome sequencing of *T. lutea*. We also thank the URGI Team for their advice about REPET as well as Jonathan Filée and Etienne Bucher for their advice on this study. We thank Helen McCombie for the proofreading. The authors are grateful to the anonymous reviewers for their critical comments, which have greatly improved the manuscript.

Funding

This work was supported by the French region of Pays de la Loire and the French Research Institute for Exploitation of the Sea (IFREMER).

Availability of data and materials

Datasets relating to the identification of TEs, as well as the improved genome assembly of *T. lutea*, are available at <https://doi.org/10.17882/52231>. The virtual machine of the PiRATE-Galaxy and the tutorial are available at <https://doi.org/10.17882/51795>.

Authors' contributions

JB developed PiRATE, carried out the TE annotation and the expression analyses, participated in the genome assembly and drafted the manuscript. NC participated in the coordination and contributed to draft the manuscript. ND designed the pipeline for the genome assembly, contributed for the analysis of the genome assembly and contributed to draft the manuscript. VJ contributed to parameter the classification and the annotation steps of PiRATE (PASTEC and TEannot), contributed to interpret the classification and annotation results and contributed to draft the manuscript. BSJ participated in the coordination of the study, the expression analyses and contributed to draft the manuscript. GC coordinated the design of PiRATE, the control, the TE annotation and the expression analyses, participated in the genome assembly, realized the gene annotation and helped to draft the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

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

Author details

¹IFREMER, Physiology and Biotechnology of Algae Laboratory, rue de l'Île d'Yeu, 44311 Nantes, France. ²Mer Molécules Santé, EA 2160 IUML - FR 3473 CNRS, Le Mans University, Le Mans, France. ³Institut de Recherche en Horticulture et Semences, INRA of Angers, AGROCAMPUS-Ouest, SFR4207 QUASAV, Université d'Angers, Angers, France. ⁴Research Unit in Genomics-Info, INRA of Versailles, Versailles, France. ⁵Université Bretagne Loire, Angers, France.

Received: 8 December 2017 Accepted: 7 May 2018

Published online: 22 May 2018

References

- Adams M, Kerlavage A, Fleischmann R, Fuldner R, Bult C, Lee N, et al. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*. 1995;377:3–174.
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. *Science*. 2009;326:1112.
- Carr M, Bensasson D, Bergman CM. Evolutionary Genomics of Transposable Elements in *Saccharomyces cerevisiae*. Stajich JE, editor. *PLoS ONE*. 2012;7:e50978.
- McClintock B. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci*. 1950;36:334–55.
- Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capi P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 2007;8:973–82.
- Bureau TE, Wessler SR. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell*. 1994;6:907–16.
- Kidwell MG. Transposable elements and the evolution of genome size in eukaryotes. *Genetica*. 2002;115:49–63.
- Chénais B, Caruso A, Hiard S, Casse N. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene*. 2012;509:7–15.
- Levis RW, Ganesan R, Houtchens K, Tolar LA, Sheen F. Transposons in place of telomeric repeats at a *Drosophila* telomere. *Cell*. 1993;75:1083–93.
- Casacuberta E, González J. The impact of transposable elements in environmental adaptation. *Mol Ecol*. 2013;22:1503–17.
- Kazanian HH. Mobile elements: drivers of genome evolution. *Science*. 2004;303:1626–32.

12. Nekrutenko A, Li W-H. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet.* 2001;17:619–21.
13. Lisch D. How important are transposons for plant evolution? *Nat Rev Genet.* 2012;14:49–61.
14. Darboux I, Charles J-F, Pauchet Y, Warot S, Pauron D. Transposon-mediated resistance to *Bacillus sphaericus* in a field-evolved population of *Culex pipiens* (Diptera: Culicidae). *Cell Microbiol.* 2007;9:2022–9.
15. Maumus F, Fiston-Lavier A-S, Quesneville H. Impact of transposable elements on insect genomes and biology. *Current Opinion in Insect Science.* 2015;7:30–6.
16. Egue F, Chenais B, Tastard E, Marchand J, Hiard S, Gateau H, et al. Expression of the retrotransposons *Surcouf* and *Blackbeard* in the marine diatom *Phaeodactylum tricomutum* under thermal stress. *Phycologia.* 2015;54:617–27.
17. Helliwell KE, Collins S, Kazamia E, Purton S, Wheeler GL, Smith AG. Fundamental shift in vitamin B12 eco-physiology of a model alga demonstrated by experimental evolution. *The ISME journal.* 2015;9:1446–55.
18. Maumus F, Allen AE, Mhiri C, Hu H, Jabbari K, Vardi A, et al. Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC Genomics.* 2009;10:624.
19. Philippsen GS, Avaca-Crusca JS, Araujo APU, DeMarco R. Distribution patterns and impact of transposable elements in genes of green algae. *Gene.* 2016;594:151–9.
20. De Clerck O, Guiry MD, Leliaert F, Samyn Y, Verbruggen H. Algal taxonomy: a road to nowhere? *J Phycol.* 2013;49:215–25.
21. Sakai N, Sakamoto Y, Kishimoto N, Chihara M, Karube I. *Chlorella* strains from hot springs tolerant to high temperature and high CO₂. *Energy Convers Manag.* 1995;36:693–6.
22. Rott E, Cantonati M, Füreder L, Pfister P. Benthic algae in high altitude streams of the alps – a neglected component of the aquatic biota. *Hydrobiologia.* 2006;562:195–216.
23. Anesio AM, Laybourn-Parry J. Glaciers and ice sheets as a biome. *Trends Ecol Evol.* 2012;27:219–25.
24. Treves H, Raanan H, Finkel OM, Berkowicz SM, Keren N, Shotland Y, et al. A newly isolated *Chlorella* sp. from desert sand crusts exhibits a unique resistance to excess light intensity. *FEMS Microbiol Ecol.* 2013;36:373–80.
25. Bertheliet J, Schnitzler CE, Wood-Charlson EM, Poole AZ, Weis VM, Detournay O. Implication of the host TGFβ pathway in the onset of symbiosis between larvae of the coral *Fungia scutaria* and the dinoflagellate *Symbiodinium* sp. (clade C1f). *Coral Reefs.* 2017;36:1263–8.
26. Guiry MD. How many species of algae are there? *J Phycol.* 2012;48:1057–63.
27. Not F, Siano R, Kooistra WHCF, Simon N, Vulot D, Probert I. Diversity and Ecology of Eukaryotic Marine Phytoplankton. *Advances in Botanical Research* [Internet]. Elsevier; 2012 [cited 2015 Oct 29]. p. 1–53. Available from: <http://linkinghub.elsevier.com/retrieve/pii/B9780123914996000013>
28. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, et al. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science.* 2007;318:245–50.
29. Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, et al. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carterii*. *Science.* 2010;329:223–6.
30. Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, et al. The *Chlorella variabilis* NC64A genome reveals adaptation to Photosymbiosis, coevolution with viruses, and cryptic sex. *Plant Cell.* 2010;22:2943–55.
31. Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ, Robbens S, et al. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc Natl Acad Sci.* 2006;103:11647–52.
32. Palenik B, Grimwood J, Aerts A, Rouzé P, Salamov A, Putnam N, et al. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc Natl Acad Sci.* 2007;104:7705–10.
33. Worden AZ, Lee J-H, Mock T, Rouzé P, Simmons MP, Aerts AL, et al. Green evolution and dynamic adaptations revealed by genomes of the marine Picoeukaryotes *Ostreococcus*. *Science.* 2009;324:268.
34. Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, et al. The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biol.* 2012;13:R39.
35. Vieler A, Wu G, Tsai C-H, Bullard B, Cornish AJ, Harvey C, et al. Genome, Functional Gene Annotation, and Nuclear Transformation of the Heterokont Oleaginous Alga *Nannochloropsis oceanica* CCMP1779. *Bhattacharya D, editor. PLoS Genetics.* 2012;8:e1003064.
36. Moreau H, Verhelst B, Couloux A, Derelle E, Rombauts S, Grimsley N, et al. Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular specializations at the base of the green lineage. *Genome Biol.* 2012;13:R74.
37. Roth MS, Cokus SJ, Gallaher SD, Walter A, Lopez D, Erickson E, et al. Chromosome-level genome assembly and transcriptome of the green alga *Chromochloris zofingiensis* illuminates astaxanthin production. *Proc Natl Acad Sci.* 2017;114:E4296–305.
38. Nozaki H, Takano H, Misumi O, Terasawa K, Matsuzaki M, Maruyama S, et al. A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga *Cyanidioschyzon merolae*. *BMC Biol.* 2007;5:28.
39. Armbrust EV. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science.* 2004;306:79–86.
40. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, et al. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature.* 2008;456:239–44.
41. Tanaka T, Maeda Y, Veluchamy A, Tanaka M, Abida H, Maréchal E, et al. Oil accumulation by the oleaginous diatom *Fistulifera solaris* as revealed by the genome and transcriptome. *The Plant Cell Online.* 2015;27:162–76.
42. McEWAN M, Humayun R, Slamovits CH, Keeling PJ. Nuclear genome sequence survey of the dinoflagellate *Heterocapsa triquetra*. *J Eukaryot Microbiol.* 2008;55:530–5.
43. Jaekisch N, Yang I, Wohlrab S, Glöckner G, Kroymann J, Vogel H, et al. Comparative Genomic and Transcriptomic Characterization of the Toxicogenic Marine Dinoflagellate *Alexandrium ostenfeldii*. *Moustafa A. PLoS ONE.* 2011;6:e28012.
44. Shoguchi E, Shinzato C, Kawashima T, Gyoja F, Mungpakdee S, Koyanagi R, et al. Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. *Curr Biol.* 2013;23:1399–408.
45. Lin S, Cheng S, Song B, Zhong X, Lin X, Li W, et al. The *Symbiodinium kawagutii* genome illuminates dinoflagellate gene expression and coral symbiosis. *Science.* 2015;350:691–4.
46. Aranda M, Li Y, Liew YJ, Baumgarten S, Simakov O, Wilson MC, et al. Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle. *Scientific Reports* [Internet]. 2016 [cited 2018 Feb 16];6. Available from: <http://www.nature.com/articles/srep39734>
47. Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, et al. Pan genome of the phytoplankton *Emiliania* underpins its global distribution. *Nature.* 2013;499:209–13.
48. Hermann D. Caractérisation d'éléments transposables de type marinier chez les microalgues marines [Internet]. Université du Maine; 2011 [cited 2015 Nov 23]. Available from: <https://tel.archives-ouvertes.fr/tel-00732952/>
49. Nguyen DH. Caractérisation et expression de nouveaux éléments génétiques transposables de la superfamille Tc1-Mariner chez la microalgue marine *Amphora acutiuscula* (Bacillariophyta). 2014;
50. Chen JE, Cui G, Wang X, Liew YJ, Aranda M. Recent expansion of heat-activated retrotransposons in the coral symbiont *Symbiodinium microadriaticum*. *The ISME Journal.* 2017;
51. Carrier G, Baroukh C, Rouxel C, Duboscq-Bidot L, Schreiber N, Bougaran G. Draft genomes and phenotypic characterization of *Tisochrysis lutea* strains. Toward the production of domesticated strains with high added value. *Algal Res.* 2018;29:1–11.
52. Liu W, Pearce CM, McKinley RS, Forster IP. Nutritional value of selected species of microalgae for larvae and early post-set juveniles of the Pacific geoduck clam, *Panopea generosa*. *Aquaculture.* 2016;452:326–41.
53. Marchetti J, Bougaran G, Le Dean L, Mégrier C, Lukomska E, Kaas R, et al. Optimizing conditions for the continuous culture of *Isochrysis affinis galbana* relevant to commercial hatcheries. *Aquaculture.* 2012;326–329:106–15.
54. Sánchez Á, Maceiras R, Cancela Á, Pérez A. Culture aspects of *Isochrysis galbana* for biodiesel production. *Appl Energy.* 2013;101:192–7.
55. Bougaran G, Rouxel C, Dubois N, Kaas R, Grouas S, Lukomska E, et al. Enhancement of neutral lipid productivity in the microalga *Isochrysis affinis galbana* (T-Iso) by a mutation-selection procedure. *Biotechnol Bioeng.* 2012;109:2737–45.
56. Carrier G, Garnier M, Le Cunff L, Bougaran G, Probert I, De Vargas C, et al. Comparative transcriptome of wild type and selected strains of the microalgae *Tisochrysis lutea* provides insights into the genetic basis, lipid metabolism and the life cycle. *Abad-Grau MM. PLoS One.* 2014;9:e86889.
57. Charrier A, Bérard J-B, Bougaran G, Carrier G, Lukomska E, Schreiber N, et al. High-affinity nitrate/nitrite transporter genes (*Nrt2*) in *Tisochrysis lutea*: identification and expression analyses reveal some interesting specificities of Haptophyta microalgae. *Physiol Plant.* 2015;154:572–90.

58. Garnier M, Bougaran G, Pavlovic M, Berard J-B, Carrier G, Charrier A, et al. Use of a lipid rich strain reveals mechanisms of nitrogen limitation and carbon partitioning in the haptophyte *Tisochrysis lutea*. *Algal Res.* 2016;20: 229–48.
59. Thiriet-Rupert S, Carrier G, Chénais B, Trottier C, Bougaran G, Cadoret J-P, et al. Transcription factors in microalgae: genome-wide prediction and comparative analysis. *BMC Genomics.* 2016;17:282.
60. Thiriet-Rupert S, Carrier G, Trottier C, Eveillard D, Schoefs B, Bougaran G, et al. Identification of transcription factors involved in the phenotype of a domesticated oleaginous microalgae strain of *Tisochrysis lutea*. *Algal Res.* 2018;30:59–72.
61. Smit, A. F., Hubley, R., & Green, P. (1996). RepeatMasker. [Internet]. Available from: <http://www.repeatmasker.org>.
62. Han Y, Wessler SR. MITE-hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* 2010;38:e199–e199.
63. Price AL, Jones NC, Pezner PA. De novo identification of repeat families in large genomes. *Bioinformatics.* 2005;21:351–8.
64. Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M. De novo assembly and annotation of the Asian Tiger mosquito (*Aedes albopictus*) Repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (*Aedes aegypti*). *Genome Biology and Evolution.* 2015;7:1192–205.
65. Ragupathy R, You FM, Cloutier S. Arguments for standardizing transposable element annotation in plant genomes. *Trends Plant Sci.* 2013;18:367–76.
66. Hoen DR, Hickey G, Bourque G, Casacuberta J, Cordaux R, Feschotte C, et al. A call for benchmarking transposable element annotation methods. *Mobile DNA* [Internet]. 2015 [cited 2017 Jun 28];6. Available from: <http://www.mobilednajournal.com/content/6/1/13>
67. Arensburg P, Piégu B, Bigot Y. The future of transposable element annotation and their classification in the light of functional genomics - what we can learn from the fables of Jean de la Fontaine? *Mobile Genetic Elements.* 2016;6:e1256852.
68. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in De novo annotation approaches. *Xu Y. PLoS One.* 2011;6:e16526.
69. Smit, AF, Hubley, R. RepeatModeler Open-1.0 [Internet]. 2010. Available from: <http://www.repeatmasker.org>.
70. Hoen DR, Hickey G, Bourque G, Casacuberta J, Cordaux R, Feschotte C, et al. A call for benchmarking transposable element annotation methods. *Mob DNA.* 2015;6:13.
71. Kamoun C, Payen T, Hua-Van A, Filée J. Improving prokaryotic transposable elements identification using a combination of de novo and profile HMM methods. *BMC Genomics.* 2013;14:700.
72. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, et al. PASTEC: An Automatic Transposable Element Classification Tool. *Cordaux R. PLoS One.* 2014;9:e91929.
73. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research.* 2005;110:462–7.
74. Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA.* 2015;6:11.
75. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics.* 2008;9:18.
76. Novak P, Neumann P, Pech J, Steinhaisl J, Macas J. RepeatExplorer: a galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics.* 2013;29:792–3.
77. Koch P, Platzer M, Downie BR. RepARK—de novo creation of repeat libraries from whole-genome NGS reads. *Nucleic Acids Res.* 2014;42:e80–e80.
78. Garrett JE, Carroll D. Tx1: a transposable element from *Xenopus laevis* with some unusual properties. *Mol Cell Biol.* 1986;6:933–41.
79. Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet.* 2008;9:411–2.
80. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics.* 2011;13: nrg3117.
81. McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, et al. Illumina TruSeq synthetic long-reads empower De novo assembly and resolve complex, highly-repetitive transposable elements. *Singh N. PLoS One.* 2014;9:e106689.
82. 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. *bioRxiv.* 2017:071282.
83. Khost DE, Eickbush DG, Larracuente AM. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. *Genome Res.* 2017;27:709–21.
84. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of ion torrent, Pacific biosciences and Illumina MiSeq sequencers. *BMC Genomics.* 2012;13:341.
85. Phillippy AM. New advances in sequence assembly. *Genome Res.* 2017;27: xi–xiii.
86. Zimin AV, Puiu D, Luo M-C, 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;27:787–92.
87. 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. PLoS One.* 2014;9:e112963.
88. Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics.* 2001;17:847–8.
89. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 2007;18:188–96.
90. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC bioinformatics.* 2011;12:491.
91. Lomsadze A, Gemayel K, Tang S, Borodovsky M. Improved Prokaryotic Gene Prediction Yields Insights into Transcription and Translation Mechanisms on Whole Genome Scale. *bioRxiv.* 2017;193490.
92. Eddy SR. Others. Multiple alignment using hidden Markov models. *Ismb.* 1995;3:114–20.
93. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–10.
94. Yang L, Bennetzen JL. Structure-based discovery and description of plant and animal Helitrons. *Proc Natl Acad Sci.* 2009;106:12832–7.
95. Rho M, Tang H. MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. *Nucleic Acids Res.* 2009;37:e143–e143.
96. Wenke T, Dobel T, Sorensen TR, Junghans H, Weisshaar B, Schmidt T. Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *THE PLANT CELL ONLINE.* 2011;23:3117–28.
97. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics.* 2006;22:1658–9.
98. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* 2009;19:1639–45.
99. 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;29:644–52.
100. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* 2009;25:1105–11.
101. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics.* 2015;31:166–9.
102. Kearsley M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics.* 2012;28: 1647–9.
103. Giardine B, Riemer C, Hardison RC, Burhans R, Eltniski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 2005;15: 1451–5.

5. Résultats complémentaires

5.1. Identification de séquences similaire à des IS chez *T. lutea*

Lors de l'étape de détection des ETs par PiRATE, des séquences similaires à des ETs de bactéries, nommées IS ont été identifiées dans le génome de *T. lutea*. Les IS sont des ETs très répandus et abondants dans les génomes procaryotes (Aziz et al., 2010). Les IS sont classés en familles ayant des structures de séquences diverses. La majorité des IS code pour une transposase mais les différents groupes se déplacent avec des mécanismes différents (Siguier et al., 2014). Certains groupes d'IS possèdent des séquences terminales répétées inversées à leurs terminaisons, similaires à celles des TIR. Au même titre que les ETs chez les eucaryotes, les IS peuvent participer à l'évolution des génomes de bactéries en générant par exemple des expansions massives de copies. De même, ils peuvent impacter l'activité des gènes et également provoquer des réarrangements du génome hôte (Siguier et al., 2014).

Bien qu'étant d'origine bactérienne, la présence d'IS dans le génome des microalgues n'est pas fondamentalement surprenante compte tenu de l'origine évolutive de ces organismes. En effet, l'émergence des trois grandes lignées de microalgues (les Glaucophytes, les Rhodophytes et les Chlorophytes) résulte de la phagocytose d'une Cyanobactérie ancestrale par une cellule hétérotrophe primitive, il y a environ 1.8 milliards d'années (Delwiche, 1999). Suite à cet événement d'endosymbiose, le procaryote photosynthétique a été « domestiqué » par la cellule hôte pour donner naissance au chloroplaste, nouvel organe spécialisé dans la photosynthèse. Cette « domestication » a notamment donné lieu à des remaniements structuraux, ainsi qu'à des phénomènes d'échanges et de pertes de gènes. Ceci pour finalement aboutir au sein de la cellule hôte, à un organe photosynthétique fonctionnel : le chloroplaste.

Concernant *T. lutea*, des séquences similaires à des IS ont été identifiées par PiRATE. En effet, une première comparaison de ces séquences à la base de données NCBI a permis de distinguer des similitudes de séquences avec des IS. De même, certaines de ces séquences présentent également des similarités avec des séquences de génomes de phycodnavirus. L'origine et la présence des IS dans le génome de *T. lutea* ont été étudiées à l'occasion d'un

stage de Master 1 (Chedhomme et Badawi, 2018 mémoire de Master 1). Pour cela, des recherches de similarités avec des séquences d'IS connues ont tout d'abord été effectuées par différentes méthodes d'analyses (NCBI-BLASTn, NCBI-BLASTx et Interproscan) (Annexe 3). Ces approches ont permis de valider leur similarité avec des séquences d'IS connues répertoriés dans les banques de données NCBI et EMBL-EBI. Ces séquences ont ensuite pu être regroupées en trois groupes selon leur pourcentage d'identité. Un premier groupe est composé de cinq séquences qui partagent des similarités avec des transposons de la famille IS 5. L'utilisation de TEannot a permis d'identifier que ces séquences sont présentes en 2 578 copies dans l'assemblage du génome de *T. lutea*. Un second groupe est constitué de sept séquences similaires à des IS 5. Environ 2 142 copies de ces séquences ont été identifiées dans l'assemblage du génome de *T. lutea*. Enfin, le troisième groupe est composé d'une séquence qui partage des similarités avec les IS 605. Cette séquence est présente en 518 copies dans l'assemblage du génome de *T. lutea*. D'autre part, les séquences issues des groupes 2 et 3 présentent également des similarités avec des séquences des coccolithovirus 88 et 202, dont un hôte connu de ces virus est *E. huxleyi* (Nissimov et al., 2012a, 2012b). L'origine précise et l'éventuelle mobilité de ces séquences chez *T. lutea* restent encore à élucider.

La présence de ces séquences similaires à des IS suggère de probables anciens événements de transferts d'ADN. A la vue des informations dont nous disposons, il est prématuré de se prononcer sur l'origine précise de ces séquences dans le génome de *T. lutea*. En effet, plusieurs hypothèses peuvent être proposées pour expliquer la présence de ces séquences. Parmi celles-ci, les phénomènes de transferts de gènes durant les événements d'endosymbioses primaire et/ou secondaire ancestrales pourraient éventuellement être à l'origine de leur présence. Enfin, ces séquences pourraient également être directement issues d'événements de transferts horizontaux entre des génomes bactériens et le génome de la microalgue. Ces transferts auraient pu être orchestrés par un vecteur viral. Dans la littérature, les virus sont supposés être impliqués dans des transferts de gènes entre organismes (Gilbert et Feschotte, 2018). Dans le milieu marin, l'importance des consortiums entre microalgues, bactéries et virus permet de suggérer que ce type de transfert est probable.

5.2. Etude de l'expression des ETs à partir d'un nouvel assemblage transcriptomique de *T. lutea*

De récents travaux ont étudié l'expression des gènes de *T. lutea* sous une forte concentration en iode (Hernández Javier et al., 2018). Dans cette étude, les auteurs ont réalisé un transcriptome à partir de données de séquençage *short reads*. Ce nouveau transcriptome de *T. lutea* s'avère une bonne opportunité pour renforcer nos résultats. Ainsi, pour rechercher d'éventuels ETs actifs à partir de ces nouvelles données transcriptomiques, une requête BLASTn des séquences d'ETs autonomes de *T. lutea* a été réalisée contre ce nouveau transcriptome. Les transcrits identifiés qui partageant au minimum 70% de la taille de la séquence référente avec 98% de similarité, ont été sélectionnés. Cette analyse permet de confirmer l'activité des familles TIR Mariner *Luffy* et hAT *Ace*, ainsi que trois autres familles d'ETs qui avaient précédemment été identifiées comme exprimées (Berthelier et al., 2018) (cf. Chapitre 1 § 3). Ces données suggèrent que l'expression de ces familles d'ETs n'est pas spécifique à une condition de stress spécifique.

Tableau 5 : ETs identifiés dans un transcriptome de *T. lutea*, obtenu sous condition de forte concentration en iode.

Famille d'ETs	Nom des transcrits
hAT2_ <i>Ace</i>	isochrysis_v2.0_unigene4
Mariner7	isochrysis_v2.0_unigene2228
Mariner3_ <i>Luffy</i>	isochrysis_v2.0_unigene10267
Mariner6	isochrysis_v2.0_unigene23004
Harbinger3	isochrysis_v2.0_unigene4716

5.3. Distribution et amélioration du pipeline PiRATE

Afin de mettre le pipeline PiRATE à disposition de la communauté scientifique, les outils le composant ont été installés dans une machine virtuelle linux Kubuntu. L'utilisation de machines virtuelles permet de s'affranchir de la difficulté d'installation des programmes qui peut être complexe et chronophage. Cette machine virtuelle nommée PiRATE-VM est téléchargeable en ligne doi.org/10.17882/51795 et un Github a été créé github.com/JBerthelier/PiRATE afin d'interagir avec les utilisateurs et de connaître leurs propositions d'améliorations ou les éventuels problèmes rencontrés. Afin d'automatiser l'utilisation de ces différents outils, ceux-ci ont été intégrés dans une plateforme Galaxy. Celle-ci permet l'automatisation de programmes et rend plus facile l'accès de la bio-informatique, via une interface graphique web simple du type « clique-bouton ». La plateforme est nommée PiRATE-Galaxy (Figure 46). Un tutoriel d'utilisation de la PiRATE-VM et du PiRATE-Galaxy est proposé en ligne archimer.ifremer.fr/doc/00412/52373/.

Une seconde version de PiRATE est en cours de réalisation. Pour cette nouvelle version, un programme a été développé via l'encadrement de Tristan Frances à l'occasion de son stage de Master 2 (Frances et al., 2018 mémoire de Master 2). Le but de ce programme nommé *pirateClassif* a été dans un premier temps d'améliorer l'étape de classification de PiRATE et plus précisément la classification prédite par PASTEC, et dans un second temps, d'automatiser la construction des trois librairies d'ETs, utilisées lors de l'étape d'annotation de PiRATE (Figure 45). Ce programme permet de s'affranchir des étapes manuelles de la première version de PiRATE. Un schéma présentant les processus de *pirateClassif* est proposé en Annexe 4.

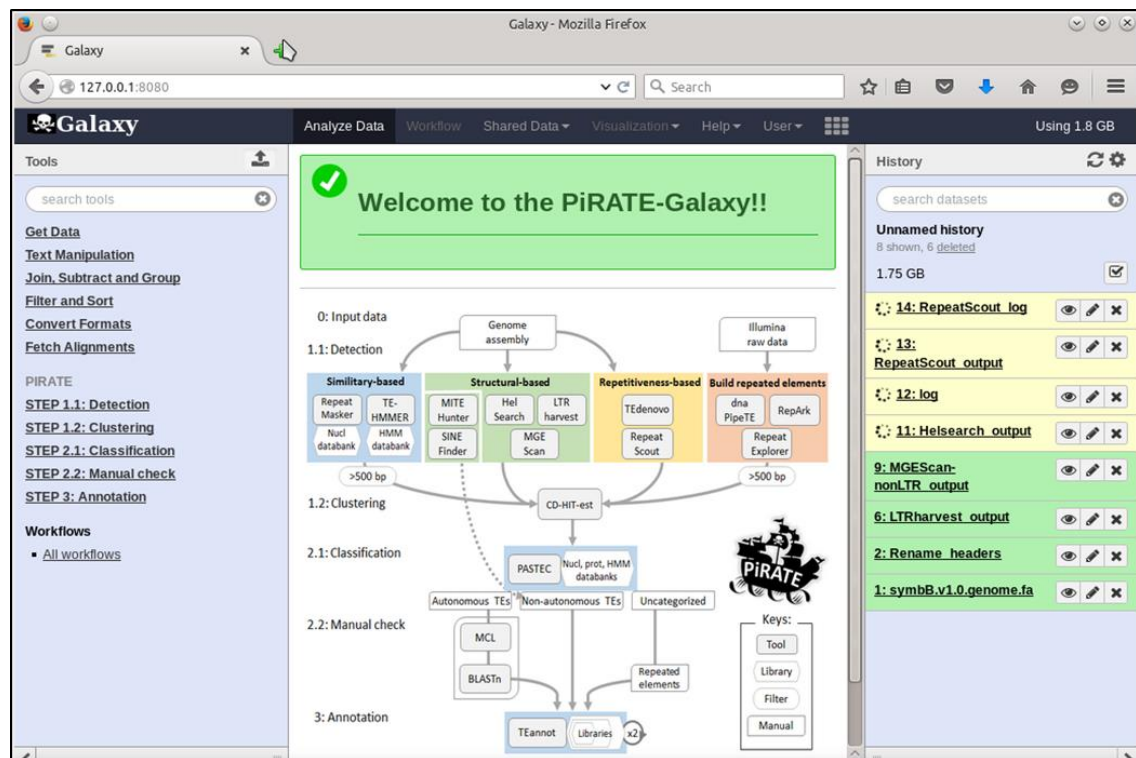


Figure 46 : Visualisation de l'interface du PiRATE-Galaxy (version 1).

Les différents outils du pipeline PiRATE sont automatisés dans cette plateforme Galaxy.

Pour améliorer la classification prédite par l'outil PASTEC lors de l'étape de classification de PiRATE, le programme `pirateClassif` commence par inventorier pour chaque séquence d'étude, le plus haut degré final de classification obtenu par PASTEC. Si aucun Ordre n'a été déterminé, la séquence est alors cataloguée, selon sa Classe prédite (e.i. « `ClassI_undefined` » ou « `ClassII_undefined` »). Au contraire, si un Ordre a été prédit par PASTEC, le programme `pirateClassif` cherche alors à prédire lui-même une Superfamille. En effet, l'Ordre est le plus haut niveau de classification proposé par PASTEC. Cependant, le fichier de classification de PASTEC indique des informations qui peuvent permettre de déterminer la Superfamille de l'élément. PASTEC recense le nom, la Classe, l'Ordre et la Superfamille des séquences référentes similaires à chacune des séquences d'étude (tBLASTx et BLASTx). PASTEC indique aussi le nom des domaines protéiques (profils HMMs) retrouvés dans les séquences d'études. Ces informations sont utilisées par `pirateClassif` pour prédire une Superfamille à la séquence d'étude (Annexe 5). Pour cela, un processus de comparaison fonctionnant à travers un compteur de points a été construit, afin de dissocier les séquences qui présentent des informations concordantes pour une Superfamille, de même que les séquences présentant des informations discordantes (Annexe 5). Si les informations issues de BLAST et de HMMER sont concordantes, une Superfamille peut alors être prédite pour l'élément (Annexe 5).

Après avoir prédit le plus haut niveau de classification possible pour chaque séquence (Classe ou Ordre ou Superfamille), `pirateClassif` construit un fichier résumé et construit de façon automatique les trois bibliothèques qui seront utilisées par PiRATE pour l'étape d'annotation (Annexe 4). Pour construire ces bibliothèques, les séquences sont sélectionnées selon leur degré de classification, leur taille et le logiciel qui a permis leur détection. Un fichier de configuration permet à l'utilisateur de configurer ces critères à façon (Annexe 6). Le programme `pirateClassif` est disponible sur Github github.com/francetr/pirateClassif.

Enfin, d'autres améliorations sont envisagées pour la seconde version de PiRATE. Celui-ci pourrait être intégré, en plus de Galaxy, dans d'autres systèmes de développement de pipelines, tels que « TOGGLE » ou « Snakemake » (Köster et Rahmann, 2012; Monat et al., 2015). En continuant à améliorer les différentes banques de données utilisées par PiRATE, ce pipeline pourra être couramment utilisé pour la recherche et l'annotation des ETs chez d'autres types d'organismes que les microalgues.

6. Conclusion

Annotation des ETs dans le génome de *T. lutea*

Le travail effectué dans cette thèse a abouti à la réalisation d'un nouvel assemblage du génome de *T. lutea*, ainsi qu'à la création du pipeline PiRATE adapté à la reconnaissance et à l'annotation des ETs chez les microalgues. L'outil PiRATE a permis d'annoter les ETs du génome de cette espèce de façon exhaustive. La proportion des ETs dans le génome de *T. lutea* est estimée à environ 20%.

Cependant, environ 18% du génome de *T. lutea* correspond à des éléments répétés qui n'ont pas pu être caractérisés lors de l'étape de classification (noCat). Un constat similaire a été précédemment établi lors de l'annotation des ETs dans le génome de l'Haptophyte *E. huxleyi* (Read et al., 2013). Pour cette espèce, les auteurs ont souligné que 30.9 % de son génome correspond à des éléments répétés qui n'ont pas pu être caractérisés. Ces éléments répétés inconnus présents dans les génomes de *T. lutea* et *E. huxleyi* pourraient correspondre à des ETs très différents de ceux présents dans les banques de données des outils de classification. Il serait ainsi intéressant d'analyser manuellement ces séquences. Ce procédé pourrait permettre la découverte de d'ETs atypiques pouvant constituer de nouvelles Superfamilles ou même de nouveaux Ordres.

Etude de l'activité des ETs de *T. lutea*

A l'aide de données transcriptomiques obtenues dans des conditions de limitations nutritives, l'expression de 17 familles d'ETs a pu être identifiée chez *T. lutea*. Parmi ces 17 familles, l'expression de cinq d'entre elles a également été observée dans un transcriptome de *T. lutea* obtenu en présence d'une forte concentration en iode (Hernández Javier et al., 2018). L'absence de données transcriptomiques « témoins » ne nous permet cependant pas d'émettre un lien entre ces différents stress et l'activité de ces éléments. Il serait donc intéressant de mesurer l'expression de ces éléments dans une condition « témoin » et dans différentes conditions de stress. Cependant, il est également envisageable que ces éléments soient exprimés de manière constitutive à un faible niveau d'expression (Oliver et al., 2010). Un même travail a été précédemment effectué pour la Diatomée *P. tricornutum*,

chez qui les auteurs ont mis en évidence la surexpression de rétrotransposons LTR en condition de stress azoté (Maumus et al., 2009) ou thermique (Egue et al., 2015; Chen et al., 2017a).

L'utilisation de données protéomiques a finalement permis de souligner que les transposases de deux familles d'ETs (Mariner *Luffy* et hAT *Ace*) sont synthétisées par la cellule en condition de stress azoté (Garnier et al. 2014). De plus, ces deux familles font partie des trois transcrits les plus exprimés dans des données transcriptomiques de *T. lutea* obtenues en condition de limitation en azote (Carrier et al., 2014; Thiriet-Rupert et al., 2018).

L'expression des ARNm et la synthèse des protéines de ces deux familles indiquent une forte probabilité de mobilité pour certains éléments autonomes de ces familles ou pour des éléments non-autonomes. Ces deux familles d'ETs représentent des candidates très intéressantes qui peuvent être impliquées dans l'évolution et la dynamique du génome de *T. lutea*.

Chapitre 2

Etude de la dynamique des ETs chez deux
souches domestiquées de *T. lutea*

Au cours du chapitre précédent, les ETs présents dans le génome de la microalgue *T. lutea* ont été annotés et deux familles nommées *Luffy* et *Ace* ont été désignées comme potentiellement actives et mobiles.

L'objectif de ce second chapitre est d'étudier la dynamique des ETs dans le génome de deux souches domestiquées de *T. lutea*. Dans un premier temps, les programmes d'amélioration à l'origine de ces deux souches domestiquées ont été décrits. Puis les données de séquençage disponibles pour les souches étudiées seront présentées. Ensuite, l'état de l'art des approches bio-informatiques permettant la prédiction d'indels associées aux ETs a été réalisé. En préambule aux paragraphes qui vont suivre, le terme « indel » est utilisé pour définir des insertions/délétions de séquences nucléotidiques. L'utilisation de plusieurs outils bio-informatiques a abouti à la prédiction d'indels associées aux ETs et spécifiques aux souches domestiquées. Les familles d'ETs dont la mobilité a été prépondérante ont été caractérisées.

1. Obtention de données génomiques pour des souches domestiquées et sauvage de *T. lutea*

1.1. Origine des souches

Le laboratoire PBA est pionnier dans l'amélioration des souches de microalgues. Il met en place depuis plus de dix ans divers programmes d'amélioration afin d'obtenir des souches domestiquées de *T. lutea* (Rouxel et al., 2011; Bougaran et al., 2012; Carrier et al., 2018). Concernant cette thèse, deux programmes d'amélioration différents ont permis d'obtenir deux souches domestiquées présentant des phénotypes distincts de la souche originelle. Cette dernière, nommée Tiso-Arg, est conservée à la station Ifremer (Institut Français de Recherche pour l'Exploitation de la Mer) d'Argenton. La souche Tiso-Arg est différente de la souche sauvage Tiso-CCAP927/14, qui a été choisie pour réaliser l'assemblage du génome de *T. lutea* (cf. Chapitre 1 § 1.2).

A partir de la souche (Tiso-Arg), l'application de deux programmes d'amélioration a permis d'obtenir deux souches domestiquées différentes (Figure 47) : la première souche, nommée Tiso-LipidePlus est caractérisée par une accumulation significative de lipides neutres alors que la seconde, nommée Tiso-LipideMoins, en présente une faible accumulation.

La souche domestiquée Tiso-LipidePlus a été obtenue à partir d'un programme d'amélioration avec pour objectif l'obtention d'une souche accumulant davantage de lipides neutres que la souche de départ (principalement sous la forme de triacylglycérol, TAG). Le programme de domestication appliqué se décline sous la forme de deux cycles de « mutation-sélection » (Figure 47). L'étape de « mutation » consiste à exposer la souche sauvage Tiso-Arg à un rayonnement UV afin de favoriser l'apparition de mutations et augmenter ainsi sa diversité génétique intra-souche. La deuxième étape, dite de « sélection », consiste à sélectionner dans la population mutagénisée, les 10% d'individus les plus riches en lipides neutres. Cette sélection est réalisée à l'aide d'un cytomètre en flux, après marquage spécifique des lipides neutres contenu dans les cellules algales (Rumin et al., 2015). Ce programme d'amélioration a conduit à l'obtention de la souche Tiso-LipidePlus, qui accumule plus de lipides neutres que la souche sauvage Tiso-Arg.

La deuxième souche domestiquée, nommée Tiso-LipideMoins, provient également de la souche Tiso-Arg. Son obtention a été réalisée par un programme d'amélioration composé de cinq étapes successives de sélection à l'aide d'un cytomètre en flux, après marquage spécifique des lipides neutres contenu dans cellules algales (Figure 47). La souche Tiso-LipideMoins accumule moins de lipides neutres par rapport à la souche d'origine Tiso-Arg.

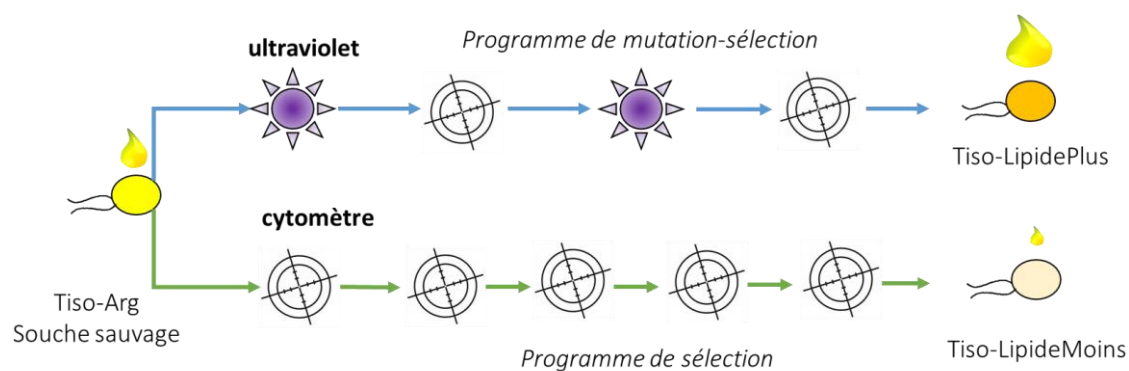


Figure 47 : Illustration des étapes des programmes d'amélioration utilisées pour obtenir les deux souches domestiquées de *T. lutea*.

Les souches domestiquées Tiso-LipidePlus et Tiso-LipideMoins ont été obtenues à partir de la souche sauvage Tiso-Arg. La souche Tiso-LipidePlus a été obtenue à la suite d'un programme de mutation-sélection et la souche Tiso-LipideMoins après cinq étapes de sélection.

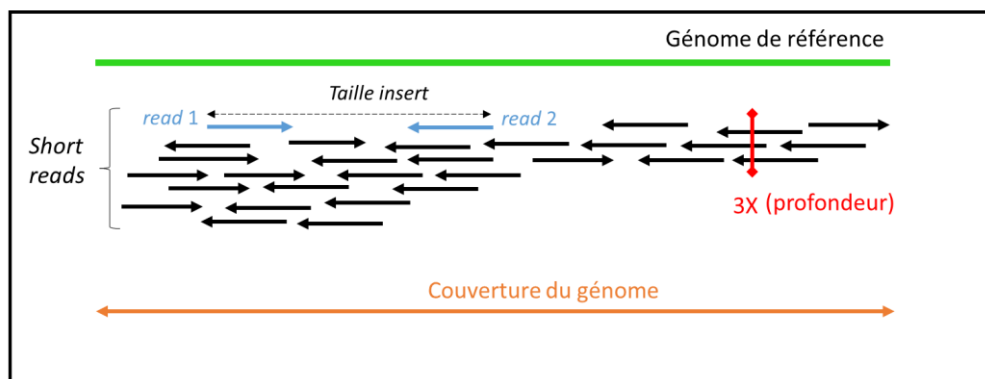
1.2. Séquençage des souches

L'objectif a été d'identifier les indels associées aux ETs dans les génomes respectifs des souches domestiquées, par rapport à celui de la souche sauvage. Pour cela, le séquençage du génome de ces trois souches (Tiso-Arg, Tiso-LipidePlus et Tiso-LipideMoins) a été réalisé à l'aide des technologies Illumina HiSeq 3000 (150 pb, *paired-end*) et Oxford Nanopore MinION. Tandis que la technologie Illumina est maintenant communément utilisée pour séquencer des génomes de microalgues (Annexe 1), la technologie Oxford Nanopore est très récente et demande encore des mises aux points techniques et méthodologiques. Actuellement, seuls deux génomes de microalgues (d'eau douce) *Galdieria sulphuraria* (Rhodophyte) et *Chlorella variabilis* (Chlorophytes) ont été séquencés avec cette technologie (Davis, 2016; Minei et al., 2018).

1.3. Alignement des données de séquençage sur le génome de référence

Après suppression des adaptateurs et des *reads* de mauvaise qualité, les données de séquençage *short reads* et *long reads* issues des génomes respectifs des trois souches ont été alignées indépendamment sur le génome de référence de *T. lutea* (Tiso-CCAP927/14 ; Bertheliet et al. 2018) avec l'outil bwa-mem (Li et Durbin, 2009). Avec le séquenceur Illumina HiSeq, au moins 33 millions de *short reads* ont été obtenues pour chacune des trois souches et ont été alignées indépendamment sur le génome référent (Tableau 6 et Figure 48a). Les profondeurs de séquençage obtenues sont respectivement de 58X, 59X et 68X pour les génomes des souches Tiso-Arg, Tiso-LipideMoins et Tiso-LipidePlus (Tableau 6 et Figure 48a). Ceci avec une couverture moyenne de 99% par rapport au génome de référence (Tableau 6 et Figure 48a). Ces résultats sont satisfaisants et correspondent à l'attente de cette technologie.

Illumina HiSeq (*short reads paired-end*)



Oxford Nanopore MinION (*long reads*)

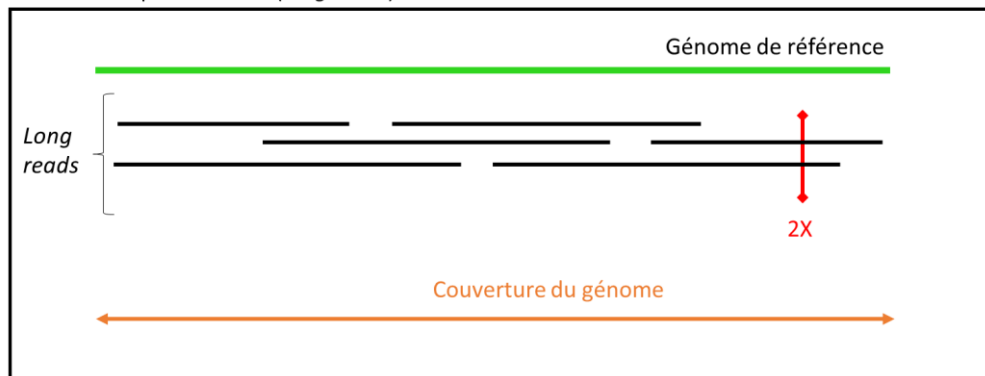


Figure 48 : Illustration de l'alignement de *short reads* en paires (*paired-end*) et de *long reads* sur un génome de référence.

a) Rappel des notions de profondeur de *reads*, de couverture de *reads* et de la taille d'insertion obtenue lors d'un alignement de *short reads* en paires (Illumina HiSeq) sur un génome de référence. Les *reads* 1 et 2, représentées en bleu, appartiennent à la même paire de *reads* et présentent une taille d'insert connue (en pb) qui doit en théorie les espacer. b) Rappel des notions de profondeur de *reads* et de couverture de *reads* obtenues lors d'un alignement de *long reads* sur un génome de référence (Oxford Nanopore MinION).

Tableau 6 : Résultats d'alignements des *short reads* et *long reads*.

Les données de séquençage sont issues des souches Tiso-Arg, Tiso-LipidePlus et Tiso-LipideMoins et ont été alignées sur le génome de référence de *T. lutea* (Bertheliet al. 2018).

	Nombre de <i>reads</i> alignées	% de <i>reads</i> alignées	Profondeur des <i>reads</i>	% du génome couvert par les <i>reads</i>
Oxford Nanopore MinION				
Tiso-Arg	76 412	83,6	7,9	78%
Tiso-LipidePlus	144 449	84,6	11,6	84%
Tiso-LipideMoins	52 088	91,7	5,1	71%
Illumina HiSeq 3000				
Tiso-Arg	33 032 147	92,0	58,3	99
Tiso-LipidePlus	38 639 744	90,6	68,2	99
Tiso-LipideMoins	33 422 550	97,8	59,0	99

Avec le séquenceur MinION, une quantité plus faible de *reads* a pu être obtenu et a été aligné sur le génome de référence (Tableau 6 et Figure 48b). Un nombre de 76 412 *reads* ont été obtenue pour Tiso-Arg, 144 449 pour Tiso-LipidePlus et 52 088 pour Tiso-LipideMoins.

En conséquence, de faibles profondeurs de 7,9X, 11,6X et 5,1X (soit en moyenne 1 Giga pb par séquençage) ont été acquises respectivement pour les génomes des souches Tiso-Arg, Tiso-LipidePlus et Tiso-LipideMoins (Tableau 6 et Figure 48b). Ces *reads* ne couvrent pas la totalité du génome de référence mais 71% pour Tiso-LipideMoins, 78% pour Tiso-Arg et 84% pour Tiso-LipidePlus (Tableau 6 et Figure 48b).

La quantité de séquences obtenue par le MinION a été inférieure au regard des chiffres annoncés par le constructeurs (10 Gpb). En 2017, ces résultats étaient globalement décevants.

Une hypothèse avancée pour expliquer ce faible résultat porte sur la pureté de l'ADN tot séquencé. En effet, l'extraction d'ADN à partir de matrices salines, telles que les cellules de microalgues marines, peut contenir de nombreuses impuretés (e.g. sels, pigments hydrosolubles) qui peuvent interférer sur l'efficacité des protocoles standards de biologie moléculaire. Par conséquent, il conviendrait d'optimiser les protocoles d'extraction afin de minimiser d'éventuelles interactions et améliorer la profondeur de séquençage de l'ADN.

2. Etat de l'art des approches visant la prédiction d'indels

Suite à l'obtention des données de séquençage (Oxford Nanopore et Illumina), l'objectif a été de sélectionner les outils *ad hoc* pour identifier les indels associées à des ETs, présentes dans les génomes respectifs des souches de *T. lutea* (Tiso-Arg, Tiso-LipidePlus et TisoLipideMoins). Pour cela, l'inventaire des différents outils bio-informatiques permettant l'identification de nouvelles indels associées à des ETs, à partir de données génomiques *short reads* et *long reads* a été réalisé. Environ, 90 outils bio-informatiques utilisant des données *short reads* ont été recensés (Annexe 7). En revanche, seuls trois outils utilisant des données de séquençage *long reads* sont disponibles (Annexe 7).

2.1. Approches permettant la prédiction d'indels à l'aide de données *short reads*

Comme mentionné précédemment, l'état de l'art des approches employées par les outils bio-informatiques permettant de détecter des indels associées à des ETs à l'aide de données de séquençage *short reads*, a été réalisé. Ces outils peuvent être regroupés en quatre approches différentes selon leur stratégie d'analyse respective (Tattini et al., 2015) : le comptage des *reads*, l'analyse des paires de *reads* discordantes, l'analyse des *reads* tronquées et enfin l'assemblage *de novo* (Figure 49). Une présentation de ces différentes approches est proposée ci-dessous.

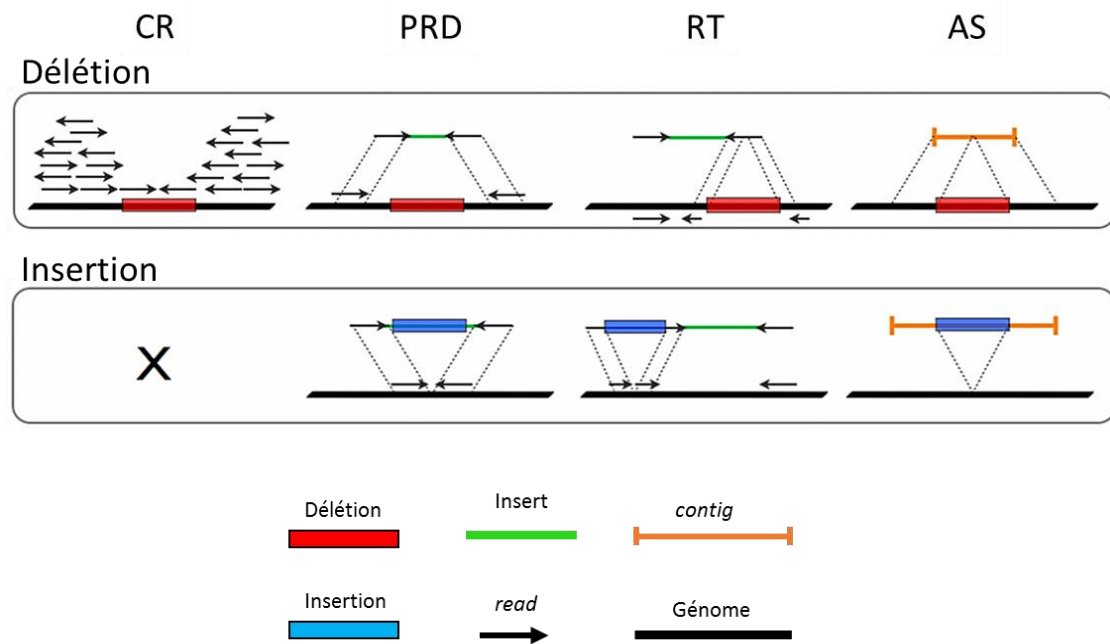


Figure 49 : Illustration des approches d'identification d'indels    l'aide de donn  es de s  quencage *short reads*.

Les diff  rentes approches indiqu  es sont: CR) par comptage de *reads*, PRD) par analyse de paires de *reads* discordantes, RT) par analyse des *reads* tronqu  es, AS) par l'assemblage de novo. Image adapt  e de Tattini et al. (2015).

Approche par comptage des reads

Cette stratégie suppose une distribution homogène des *short reads* alignées sur l'ensemble du génome de référence, à l'exception des régions sujettes aux délétions. Les outils qui utilisent cette stratégie recherchent des régions du génome présentant une distribution de *reads* différente par rapport aux autres régions du génome. Ainsi, une région du génome ayant une faible profondeur de *reads* correspondra potentiellement à un événement de délétion (Figure 49). Les outils utilisant cette méthode n'ont pas été retenus. En effet, le séquençage des régions d'un génome n'est pas uniforme pour de multiples raisons, notamment techniques. La difficulté d'amplification des régions riches en GC ou AT ainsi que le niveau de fragmentation de l'ADN lors de la préparation des échantillons avant séquençage engendrent des biais. Ceux-ci donnent lieu à une distribution non uniforme des *reads* alignées sur un génome de référence (Benjamini et Speed, 2012; Poptsova et al., 2015; Tattini et al., 2015).

Approche d'analyse de paires de *short reads* discordantes

Cette approche utilise des données de séquençage constituées de paires de *short reads* espacées d'une centaine de bp (Figure 49). En effet, les *short reads* composant une paire sont séquencées dans une orientation précise et avec un espacement défini, dit taille d'insert (Figure 48a). Les paires de *reads* discordantes qui sont alignées sur le génome de référence avec une taille d'insertion plus grande que celle attendue sont ainsi associées à des événements potentiels de délétions (Figure 49). Au contraire, les paires de *reads* discordantes présentant une taille d'insert plus petite que celle attendue seront utilisées pour prédire des événements d'insertions (Figure 49) (Tattini et al., 2015). Cette stratégie est employée et conseillée pour l'identification de délétions associées aux ETs et a été sélectionnée dans cette étude (Chen et al., 2017b; Lu et al., 2017; Rajaby et Sung, 2018; Goerner-Potvin et Bourque, 2018). Au contraire, cette approche n'a pas été choisie pour identifier de nouvelles insertions d'ETs, en raison de son manque de précision et de preuve concernant la prédiction de larges insertions (Ewing, 2015 ; Tattini et al., 2015).

Analyse des *short reads* tronquées

Cette méthode utilise également des *short reads* en paires pour identifier des indels. Cette approche exploite les paires de *reads* pour lesquelles l'une des deux *reads* est alignée correctement sur le génome de référence et pour qui la seconde *read* est tronquée. La portion tronquée de la seconde *read* ne s'alignant pas correctement sur le génome de référence peut être utilisée pour prédire la position et l'origine de la séquence absente ou insérée. Cette méthode est recommandée dans la littérature (Ewing, 2015; Goerner-Potvin et Bourque, 2018) et a été choisie pour détecter de nouvelles indels associées à des ETs. Cette approche est particulièrement pertinente pour l'identification de nouvelles insertions d'ETs. En effet, les outils peuvent identifier des similarités avec des extrémités d'ETs connus, ce qui rend possible l'identification de la famille de l'ET potentiellement inséré (Ewing, 2015). Cependant, cette stratégie ne permet pas de dissocier si une séquence d'ETs insérée appartient à un élément autonome ou non-autonome.

Approche par assemblage des *short reads* discordantes

Enfin, des outils ont été élaborés afin de reconstruire les indels grâce à l'assemblage des *short reads*. Lors de l'alignement des *short reads* sur un génome de référence, les points de cassure des *reads* tronquées peuvent être une conséquence de l'insertion ou de la délétion d'ETs (Figure 49). Au niveau de ces points de cassure, les outils de cette approche extraient l'ensemble des *short reads* discordantes. Ils assemblent ensuite ces *short reads* afin de reconstruire et identifier la séquence à l'origine de l'indel. Cette méthode n'a pas été sélectionnée en raison de la difficulté de reconstruction des séquences d'ETs à l'aide de *short reads* (Tattini et al., 2015).

2.2. Outils sélectionnés pour prédire des indels associées à des ETs à l'aide de données *short reads*

Parmi les différents outils existants (Annexe 7), Pindel (Ye et al., 2009), Breakdancer (Chen et al., 2009) et Mobster (Thung et al., 2014) ont été sélectionnés.

Les outils Pindel et Breakdancer sont fortement employés dans la littérature pour conduire une recherche *de novo* de variations génétiques et en particulier des délétions de séquences nucléotidiques (Lu et al., 2017; Rajaby et Sung, 2018). Ils ont été régulièrement améliorés par leurs concepteurs et sont complémentaires. Ces outils combinent les deux approches choisies pour prédire des délétions associées à des ETs. Le programme Pindel met en œuvre la stratégie d'analyse basée sur « l'analyse de *reads* tronquées ». Le programme Breakdancer utilise la stratégie « d'analyse des paires de *reads* discordantes ».

Pour identifier des événements d'insertions à partir de données *short reads*, le logiciel Mobster a été sélectionné. Celui-ci a été spécifiquement conçu pour identifier des insertions d'ETs entre un génome de référence et des données *short reads* en paires. Il utilise la méthode « d'analyse de *reads* tronquées » et/ou une approche « d'analyse des paires de *reads* ». Comme indiqué précédemment, l'approche d'analyse de paires de *reads* n'a pas été choisie pour détecter des insertions associées aux ETs. Ainsi, seule l'approche d'analyse de *reads* tronquées a été employée avec cet outil.

Pour ces trois outils (Pindel, Breakdancer et Mobster), seules les indels associées à des ETs, présentant une fréquence allélique supérieure à 10% et supportées par un minimum de 6 *reads* ont été considérées.

2.3. Outil sélectionné pour prédire des indels associées à des ETs à l'aide de données *long reads*

Seulement trois outils ayant pour but la prédiction d'indels à l'aide de séquences *long reads* étaient disponibles lors de cette thèse : LoRTE (Disdero et Filée, 2017), SMRT-SV (Huddleston et al., 2017) et SNIFFLES (Sedlazeck et al., 2018). SMRT-SV et SNIFFLES ont été développés pour identifier plusieurs types de variations génétiques, tandis que LoRTE a été conçu pour identifier des indels associées à des ETs.

SMRT-SV identifie des indels en utilisant l'assemblage d'un génome référent pour générer un consensus à chaque région composée de variations. SNIFFLES identifie les *reads* tronquées ou qui sont très fortement discordantes lors de leur alignement sur le génome de référence (Goerner-Potvin et Bourque, 2018).

Dans cette analyse, l'outil LoRTE a été choisi en raison de sa spécificité vis-à-vis des ETs. Cet outil utilise le génome référent de l'espèce d'étude, des données *long reads*, ainsi qu'une librairie des séquences référentes d'ETs de l'espèce d'étude et leur annotation.

LoRTE utilise l'information des ETs connus (leurs séquences et leur annotation) afin d'identifier la présence d'indels d'ETs dans les données *long reads* par rapport au génome de référence. Cet outil est constitué de deux modules. Un premier permet de vérifier dans les données *long reads* la présence ou l'absence d'ETs aux positions indiquées dans l'annotation du génome de référence. Ceci sert à identifier des événements de délétions. Un second module permet l'identification de nouvelles insertions dans les *long reads* par rapport au génome de référence, en prenant en compte l'annotation des ETs et des séquences référentes présentes dans la banque de données *ad hoc*. LoRTE est particulièrement complémentaire à Mobster, car contrairement à ce dernier, il permet de reconnaître si l'ET inséré est un élément autonome ou non-autonome. Par contre, en raison du fort taux d'erreur que présentent pour le moment les données de séquençage *longs reads*, il ne permet pas de connaître la séquence exacte de l'élément inséré.

Finalement, quatre outils ont été sélectionnés pour prédire des indels chez les souches domestiquées. Un aperçu de leurs approches, limites et avantages est proposé dans le Tableau 7. Un tableau récapitulatif des outils sélectionnés dans ce chapitre est également présenté en Annexe 2.

Tableau 7 : Synthèse des caractéristiques des outils choisies pour conduire les prédictions d'indels associées aux ETs.

Outils	Données de séquençage	Prédictions	Approches	Avantages	Limites
Pindel	<i>Short reads (paired-end)</i>	Délétions de séquences	Analyse de <i>reads tronquées</i>	Recherche de novo des délétions	Prédictions à vérifier
Break dancer	<i>Short reads (paired-end)</i>	Délétions de séquences	Analyse de paires de <i>reads discordantes</i>	Recherche <i>de novo</i> des délétions	Prédictions à vérifier
Mobster	<i>Short reads (paired-end)</i>	Insertions d'ETs	Analyse de <i>reads tronquées</i>	Dédié aux insertions d'ETs	Prédictions à vérifier
LoRTE	<i>Long reads</i>	Insertions et délétions d'ETS	Analyse par similarités	Dédié aux d'ETs et permet de connaître la nature des indels	Le taux d'erreur des <i>long reads</i>

3. Prédiction d'indels associées aux ETs et spécifiques aux souches domestiquées de *T. lutea*

3.1. Prédiction d'indels associées à des ETs avec des outils utilisant des données *short reads*

Dans un premier temps, des événements d'indels ont été recherchés dans les données génomiques de chacune des souches (Tiso-Arg, Tiso-LipidePlus, Tiso-LipideMoins) avec les outils Pindel, Breakdancer et Mobster, par rapport au génome de référence de l'espèce (Berthelier et al. 2018). Puis, les indels potentiellement spécifiques aux souches domestiquées ont été identifiés. La démarche est illustrée sous la forme d'un pipeline Figure 50.

Prédiction de délétions associées à des ETs

Les outils Pindel et Breakdancer ont permis d'identifier 1441 et 72 délétions chez la souche Tiso-Arg, 1485 et 84 délétions chez la souche Tiso-LipidePlus et 1450 et 82 délétions chez la souche Tiso-LipideMoins. Les délétions communes entre les génomes respectifs des souches domestiquées et celui de la souche sauvage ont ensuite été retirées (Figure 50). Avec Pindel et Breakdancer, 553 et 34 délétions spécifiques ont été prédites pour le génome de la souche Tiso-LipidePlus (Tableau 8). Ensuite, 393 et 26 délétions spécifiques ont été prédites pour le génome de la souche Tiso-LipideMoins (Tableau 8).

Tableau 8 : Résultats obtenus avec Pindel et Breakdancer pour les génomes des souches Tiso-Arg, Tiso-LipidePlus et TisoLipideMoins.

Nombre de délétions :	Pindel			Breakdancer		
	Total	Spécifique inf. 150 bp	Spécifique sup. à 150 bp	Total	Spécifique inf. 150 bp	Spécifique sup. à 150 pb
Tiso-Arg	1441	/	/	72	/	/
Tiso-LipidePlus	1485	541	12	84	5	29
Tiso-LipideMoins	1450	382	11	82	5	21

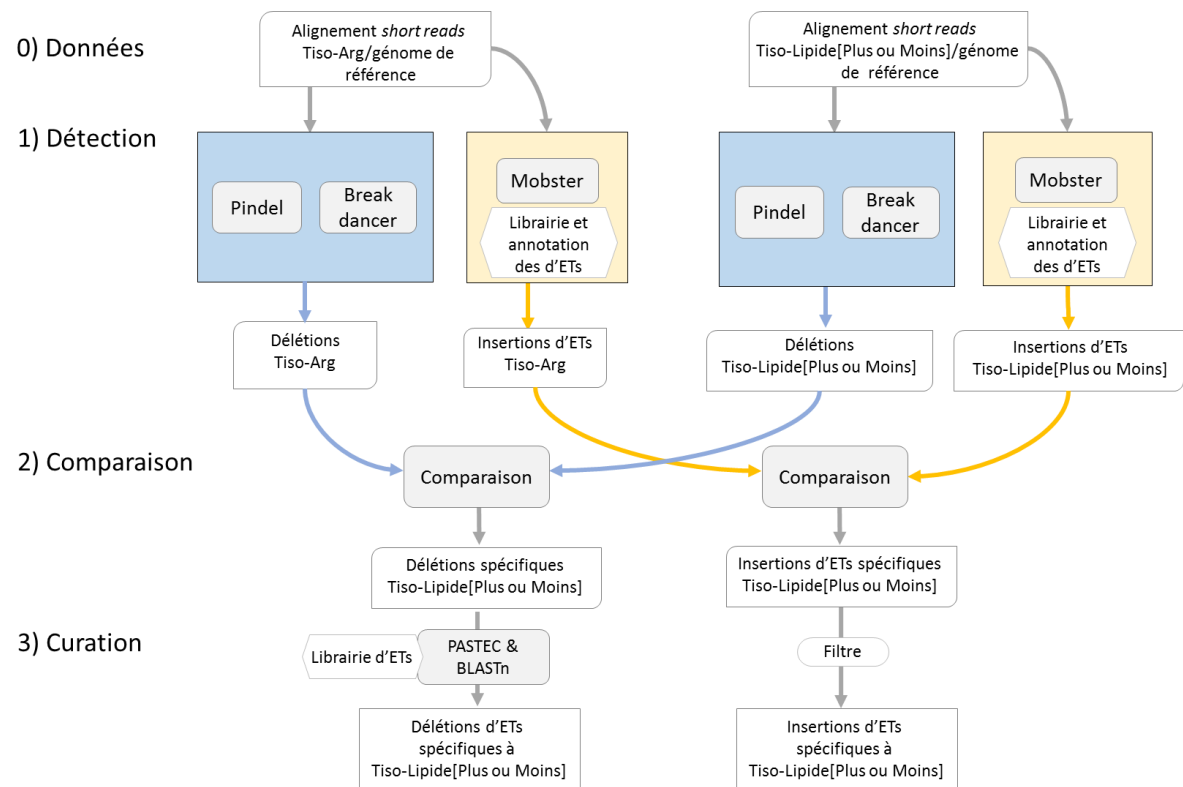


Figure 50 : Illustration du pipeline employé pour prédire des indels associées à des ETs qui sont spécifiques aux souches domestiquées.

0) L'alignement des *short reads* de la souche sauvage (Tiso-Arg) et d'une des deux souche domestiquées (Tiso-LipidePlus ou TisoLipideMoins) est réalisé indépendamment sur le génome de référence. 1) Etape de détection : Les alignements sont soumis aux outils Pindels et Breakdancer pour identifier des délétions et à Mobster pour identifier des insertions. 2) Etape de comparaison : Les prédictions obtenues par Pindel et Breakdancer sont combinées. Les prédictions de délétions et d'insertions identifiées chez les deux souches ont été comparées (positions et tailles) et seuls les événements de délétions spécifiques aux souches domestiquées ont été gardés. 3) Etape de curation : Les séquences associées à une délétion et spécifiques à la souche domestiquée ont été comparées à la librairie d'ETs autonomes de *T. lutea* avec BLASTn, puis soumises à PASTEC, afin d'identifier celles correspondant à des ETs. Finalement, des prédictions d'indels associées aux ETs et spécifiques à la souche domestiquée ont été obtenues.

Ensuite, seuls les événements d'une taille de séquence supérieure à 150 pb ont été considérés comme pouvant être associés à des ETs (Tableau 8). Parmi ces prédictions, deux délétions spécifiques à la souche Tiso-LipidePlus ont pu être identifiées comme des séquences d'ETs. Ces délétions présentent des terminaisons TIRs distinctes à leurs extrémités. La première séquence est un ET non-autonome de la famille *Ace* (famille caractérisée dans le chapitre 1) et la seconde correspond à un élément non-autonome d'une nouvelle famille encore non-caractérisée et nommée *Shanks* (Tableau 9). Cette nouvelle famille a été caractérisée plus en détail dans une autre partie de ce chapitre (cf. Chapitre 2 § 4.3). Concernant la souche Tiso-LipideMoins, aucune délétion spécifique au génome de cette souche n'a été détectée.

Tableau 9 : Prédictions d'événements de délétions spécifiques à la souche Tiso-LipidePlus.

Contig	Début	Fin	Longueur (pb)	Familles ETs	Outils	Fréquence allélique (%)
Contig_142	46892	48002	1110	<i>Shanks</i>	Pindel & Breakdancer	27
Contig_093	1613198	1614060	862	<i>Ace</i>	Pindel & Breakdancer	25

Prédiction d'insertions associées à des ETs

Les nouvelles insertions d'ETs ont été prédites avec l'outil Mobster. Celui-ci a été employé avec une librairie de séquence des ETs autonomes de *T. lutea* identifiés dans le Chapitre 1 et leur annotation dans le génome. Il est à noter que les informations relatives aux éléments appartenant à la famille *Shanks* ont été ajoutées dans la librairie et l'annotation. Ainsi, de la même façon que pour les délétions, des événements d'insertions ont pu être identifiés dans les génomes respectifs des souches Tiso-Arg, Tiso-LipidePlus et Tiso-LipideMoins, par rapport au génome référence. Un total de 623 événements d'insertions a été identifié pour le génome de Tiso-Arg, 755 pour le génome de Tiso-LipidePlus et 825 pour le génome de Tiso-LipideMoins.

Les événements d'insertions spécifiques aux souches Tiso-LipidePlus et Tiso-LipideMoins ont ensuite été recherchés en enlevant ceux étant communs à la souche Tiso-Arg, comme énoncé précédemment (Figure 50). Avec cette méthodologie, 144 événements d'insertions ont été détectés pour la souche Tiso-LipidePlus (Annexe 8) et 32 pour la souche Tiso-LipideMoins (Annexe 9). Parmi ces insertions 141/144 (98%) et 26/32 (81%) sont prédites

pour les familles d'ETs *Ace*, *Luffy* et *Shanks* (Annexes 8 et 9). La méthodologie de Mobster ne permet pas de savoir si ces événements sont dus à la mobilité d'ETs autonomes ou non-autonomes (Thung et al., 2014). Malgré les précautions et filtres utilisés, ces analyses sont sujettes à des faux positifs.

Ainsi, les prédictions d'indels obtenues par ces trois outils seront validées par des méthodes de PCR et/ou de re-séquençage.

3.2. Prédiction d'indels associées à des ETs avec un outil utilisant des données *long reads*

L'outil LoRTE a été appliqué afin d'identifier des indels associées à des ETs à partir de données *long reads*, et confirmer les prédictions obtenues par Pindel, Breakdancer et Mobster. Dans un premier temps, il a été utilisé avec l'annotation des ETs potentiellement autonomes et la librairie des séquences correspondantes, en incorporant les informations obtenues pour la famille *Shanks*. Dans un second temps, il a été employé avec une nouvelle librairie contenant des potentielles séquences non-autonomes des familles *Ace* et *Shanks*, ainsi que leur annotation dans le génome (cf. Chapitre 2 § 4.3).

Malgré l'identification de potentielles indels, l'utilisation de LoRTE n'a pas été concluante. Tout d'abord, aucune des deux délétions identifiées par Pindel et Breakdancer n'a été identifiée par LoRTE à partir des données *long reads* de la souche Tiso-LipidePlus. Ensuite, parmi les 144 événements d'insertions prédits par Mobster pour la souche Tiso-LipidePlus, LoRTE n'a identifié que quatre de ces insertions (Tableau 10). Concernant la souche Tiso-LipideMoins, aucune des 32 insertions prédites par Mobster n'a été détectée par LoRTE.

Ces résultats suggèrent que l'utilisation de LoRTE ne semble pas appropriée pour ce cas d'étude. Il est plausible que la majorité des indels ne soient pas présentes dans ces données *long reads*, en raison de leur faible fréquence dans la population et de la faible profondeur des *long reads* obtenues. En effet, les événements d'insertions d'ETs prédites par l'outil Mobster présentent en moyenne une faible fréquence allélique estimée à 26% et 19% respectivement, pour les souches Tiso-LipidePlus et Tiso-LipideMoins. Ce résultat indique que ces variations génétiques sont présentes en minorité dans la population d'individus qui constituent chacune de ces souches. Ainsi, l'absence d'une profondeur de séquençage

suffisante (7X en moyenne) reste l'hypothèse la plus probable pour expliquer l'échec de l'outil LoRTE dans notre analyse. Cette hypothèse est d'autant plus crédible à la vue des résultats d'efficacité de LoRTE par ses concepteurs (Disdero et Filée, 2017). Dans cette étude réalisée sur l'espèce *D. melanogaster*, les auteurs ont montré que pour une profondeur de 9X, la plupart des indels d'ETs homozygotes sont identifiées par LoRTE (5/6 délétions et 12/14 des insertions). En revanche, pour cette même profondeur de séquençage, très peu d'indels d'ETs hétérozygotes ont pu être identifiées (2/3 délétions et 6/22 insertions).

Les résultats obtenus par l'utilisation de LoRTE sont décevants car les résultats de cet outil pouvaient être complémentaires à ceux de Mobster, Pindel et Breakdancer. En effet, LoRTE aurait pu nous permettre de confirmer les résultats issus des autres outils utilisés, mais également d'identifier de nouvelles indels.

Parmi les indels détectées par LoRTE, quatre événements d'insertions sont similaires aux prédictions de Mobster pour le génome de la souche Tiso-LipidePlus. Ces insertions correspondent à des ETs non-autonomes des familles *Ace* et *Shanks* et à un élément autonome de la famille *Luffy*. Bien que ce résultat indique un poids supplémentaire dans la dynamique de ces familles d'ETs, les résultats de LoRTE ne peuvent pas confirmer si ces insertions sont spécifiques à cette souche domestiquée. Ceci devra être validé ultérieurement par des méthodes de PCR et/ou de re-séquençage.

Tableau 10 : Prédiction d'événements d'insertions consensus entre LoRTE et Mobster pour la souche Tiso-LipidePlus.

<i>Contigs</i>	Position Insertions	Taille insertions	Famille ETs	Nombre de <i>reads</i> soutenant l'insertion	
				LoRTE	Mobster
Contig_139	1732637	2966	<i>Luffy</i>	1	8
Contig_173	1833115	1214	<i>Ace</i>	1	14
Contig_162	330582	3317	<i>Shanks</i>	2	5
Contig_041	117447	3310	<i>Shanks</i>	2	12

4. Identification des ETs les plus dynamiques dans les génomes des souches domestiquées de *T. lutea*

A partir des prédictions d'indels obtenues par Mobster, Breakdancer et Pindel, la proportion des différentes familles d'ETs potentiellement mobiles a été établie pour les deux souches domestiquées Tiso-LipidePlus et Tiso-LipideMoins.

4.1. Proportion des familles d'ETs associées à des indels

Pour le génome de la souche Tiso-LipidePlus, 146 indels spécifiques (144 insertions et 2 délétions) ont été prédites et 99% (144 indels) sont associées à des ETs de Classe II (Figure 51 et Annexe 8). La majorité des indels correspondent à la famille TIR hAT *Ace* (82% soit 119 indels). De potentielles insertions ont aussi été attribuées à la famille *Shanks* (10% soit 15 indels), à la famille TIR Mariner *Luffy* (6% soit 9 insertions) et une insertion serait liée à la famille de TIR *Harbinger7* (1% soit 1 insertion). Enfin, deux familles de rétrotransposons LTR, *Gypsy1* et *Gypsy2* sont associées à 1% des insertions (1 insertion détectée respectivement pour *Gypsy1* et *Gypsy2*).

Chez le génome de la souche Tiso-LipideMoins 32 potentielles insertions spécifiques ont été détectées. Elles sont aussi pour la plupart attribuées à des éléments de Classe II (81% soit 26 insertions) (Figure 51 et Annexe 9). La majorité de ces potentielles insertions correspondent à des éléments appartenant à la famille TIR hAT *Ace* (66% soit 21 insertions). Des insertions sont aussi prédites pour la famille d'éléments TIR Mariner *Luffy* (13% soit 4 insertions) et seulement une insertion est associée à la famille TIR hAT *Shanks*. Des insertions correspondantes aux familles de Classe I LTR *Gypsy1* (3 insertions), *Gypsy2* (2 insertions) et *Copia6* (3%) ont aussi été prédites.

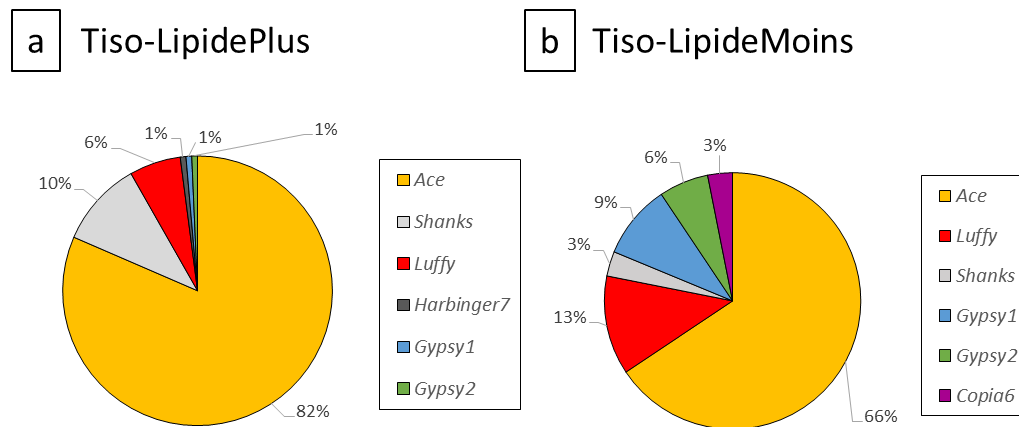


Figure 51 : Pourcentage d'indels associées aux familles d'ETs dans le génome des deux souches domestiquées de *T. lutea*.

Pourcentage d'indels associées aux familles d'ETs, a) chez la souche domestiquée Tiso-LipidePlus, b) chez la souche domestiquée Tiso-LipideMoins.

4.2. Activité prépondérante de trois familles d'ETs

Au sein des indels prédites chez les souches domestiquées, la majorité est issue d'éléments de la Classe II (95% Classe II vs 5% Classe I). Celles-ci sont principalement attribuées à trois familles de TIR: *Ace*, *Shanks* et *Luffy* (Figure 51). L'activité de *Luffy* et *Ace* a été précédemment identifiée chez *T. lutea*, à partir de données transcriptomiques et protéomiques obtenues en conditions d'un stress nutritif en azote (Bertheliet al 2018). Ainsi, l'identification d'indels associées à ces familles d'ETs dans les génomes respectifs de ces deux souches domestiquées soutient la potentielle mobilité d'éléments associés à ces familles, comme suggéré dans le Chapitre 1. Dans ce second chapitre, une nouvelle famille d'ETs a été identifiée chez *T. lutea* nommée *Shanks*. Celle-ci semble aussi particulièrement active.

Un plus grand nombre d'insertions a été prédit pour ces éléments TIR par rapport au nombre de délétions. Cette différence pourrait suggérer que les transpositions de ces éléments sont réalisées lors d'événements de réplication et/ou réparation de l'ADN. En effet, il est décrit dans la littérature que lorsque la transposition d'un ET appartenant à l'Ordre des TIR est réalisée à partir d'une chromatide nouvellement répliquée vers un site non répliqué, cela permet le gain d'une copie de l'élément (Feschotte et Pritham, 2007). Ce comportement a été documenté pour l'élément hAT *Ac* chez le maïs. Il s'explique par la liaison préférentielle de la transposase *Ac* aux sites de liaisons hémi-méthylées (Ros et Kunze, 2001). Enfin, il est également possible qu'à la suite de l'excision d'un élément TIR, une copie de celui-ci soit générée au niveau de sa position d'origine par l'intervention des mécanismes de réparation de l'ADN. En effet, si l'ET qui a été excisé sur l'un des deux chromosomes est présent sur le chromosome homologue, la réparation de l'excision par recombinaison homologue peut entraîner la réintroduction du transposon sur le site donneur (Feschotte et Pritham, 2007). Chez *D. melanogaster*, les transposons *P* ont leur nombre de copies qui augmente avec ce type de mécanisme (Engels et al., 1990). Enfin, il est probable qu'une grande partie de ces mutations soit causée par des amplifications de MITE. Pour exemple, le génome du riz contient des milliers de copies de MITE qui peuvent être déplacées en *trans* par des éléments autonomes (Feschotte, 2005).

En raison de leur activité prépondérante, une caractérisation détaillée des familles *Luffy*, *Ace* et *Shanks* a été réalisée.

4.3. Caractérisation des familles d'ETs les plus dynamiques

La famille *Luffy*

Dans le génome de référence de *T. lutea*, 20 copies d'éléments potentiellement autonomes appartenant à la famille *Luffy* ont été identifiées. Parmi elles, 12 copies sont identiques et possèdent une taille de 2 568 pb. Des éléments non-autonomes associés à cette famille ont été recherchés avec l'outil TEannot (Quesneville et al., 2005), mais aucune séquence n'a été identifiée. En revanche, un élément de grande taille (5 129 pb) a été détecté. Celui-ci est en réalité composé de deux copies juxtaposées. Cette chimère peut découler d'un événement de duplication d'une des copies ou d'une éventuelle transposition enchâssée. Concernant les terminaisons TIRs de la famille *Luffy*, celles-ci sont constituées de 14 pb. Chacune des terminaisons est bordée par des TSD « TA », ce qui suggère que cette famille d'ET appartient à la Superfamille TIR Mariner comme proposé précédemment (Berthelier et al., 2018). Un ORF unique de 463 AA a été identifié pour les ETs autonomes de cette famille (ORFfinder, www.ncbi.nlm.nih.gov/orffinder/) (Figure 52). En revanche, aucun domaine protéique n'a été identifié avec Pfam pour cet ORF (Finn et al., 2011). L'outil HHpred (Söding, 2005; Zimmermann et al., 2018) apporte, quant à lui, une prédiction de 97% pour la transposase du TIR Mariner *Mos* apparentée au génome de l'organisme *Drosophila mauritiana* (Annexe 10) (Medhora et al., 1991). Cet ORF a ensuite été comparé à la base de données Repbase avec l'outil Censor (Kohany et al., 2006). Ce dernier partage une similarité de 35% avec l'élément *Mariner-12_HM* appartenant à l'espèce *Hydra vulgaris* (Annexe 10). Une analyse d'alignement global indique que les transposases de ces deux éléments partagent un faible pourcentage d'identité de 22% (Annexe 10) (Geneious, Kearse et al., 2012). Enfin, la séquence protéique de la transposase de *Luffy* a été comparée avec tBLASTx à des assemblages de génomes d'Haptophytes disponibles (*E. huxleyi*, *Pavlova* sp. et *C. tobin*, Annexe 1). Dix-huit séquences similaires ont été détectées dans le génome de *Pavlova* sp., ainsi qu'une séquence similaire dans le génome de *E. huxleyi* (Annexe 11).

La famille Ace

La famille *Ace* est composée de trois copies potentiellement autonomes dans le génome de *T. lutea*, partageant 99% d'identité et d'une taille moyenne de 2 589 pb. Des éléments non-autonomes associés à cette famille ont été recherchés à l'aide de l'outil TEannot et 92 séquences potentielles ont été identifiées. Ces séquences ont des tailles comprises entre 243 et 1 631 pb et deux terminaisons TIR similaires à la famille *Ace*. Les terminaisons TIRs des éléments de cette famille ont une longueur de 12 pb et sont bordées de TSD de taille de 8 bp. Ainsi, cette famille d'ETs appartient à la Superfamille TIR hAT comme proposé précédemment dans le Chapitre 1 (Berthelier et al., 2018). Les éléments autonomes de la famille *Ace* codent pour un unique ORF de 713 AA (Figure 52) dans lequel deux domaines protéiques (PF04937 et PF05699) ont été identifiés par l'outil Pfam (Annexe 10). Ces deux domaines sont associés à la Superfamille des TIR hAT. Enfin, le programme HHpred a permis d'obtenir une prédiction de 100% pour une transposase de l'élément TIR hAT *Hermes* de l'organisme *Musca domestica* (Annexe 10) (Warren et al., 1994). Cet ORF a ensuite été comparé à Repbase avec l'outil Censor et une similarité de 29% a été identifiée pour l'élément *hAT-7_ATr* appartenant à l'espèce *Amborella trichopoda* (Annexe 10). L'analyse de l'alignement global indique que les transposases de ces éléments partagent entre elles un pourcentage d'identité de 14% (Annexe 10). Enfin, la séquence protéique de la transposase de *Ace* a été comparée à des assemblages de génomes d'Haptophytes, comme indiqué précédemment pour *Luffy*. Un élément potentiel similaire à *Ace* a été identifié dans les génomes respectifs des Haptophytes *C. tobin* et *E. huxleyi* (Annexe 11).

La famille *Shanks*

Dans le génome de *T. lutea*, 30 séquences de cette famille ont été identifiées comme potentiellement autonomes. Celles-ci partagent une identité de 99.9% et une taille moyenne de 6 791 pb. A l'aide de l'outil TEannot, 85 potentiels éléments non-autonomes associés à cette famille ont été détectés. La taille de ces éléments varie de 462 à 3 049 pb. Ces séquences présentent des terminaisons TIRs similaires aux éléments autonomes. Les éléments qui constituent cette famille présentent des terminaisons TIRs d'une taille de 16 pb et une taille de TSD de 7 pb. Les copies potentiellement autonomes de cette famille codent respectivement pour deux ORF de 796 AA et 580 AA (Figure 52). Ces deux ORF ont été analysés avec les outils HHpred et Pfam. Aucun résultat n'a été obtenu avec Pfam. En revanche, HHpred donne une prédiction de 93% pour l'ORF1 par rapport à la transposase de l'élément TIR hAT *Hermes* présent chez *M. domestica* (Annexe 10) (Warren et al., 1994). Les deux ORF de *Shanks* ont ensuite été comparés à Repbase. Cette analyse montre que ces ORF présentent tous deux une similarité d'environ 29% avec l'élément *hATw-1_BF* de l'organisme *Branchiostoma floridae* (Annexe 10). Des alignements globaux ont été réalisés et les ORF 1 et 2 présentent tous les deux un faible pourcentage d'identité d'environ 15% par rapport à la transposase de l'élément *hATw-1_BF*. L'élément *hATw-1_BF* correspond à un sous-groupe de hAT dans la base de données Repbase, nommé hATw (Annexe 10). Ce sous-groupe de hAT est composé de 17 éléments ayant la particularité de générer des TSD de 7 pb, tout comme la famille *Shanks*. Il est ainsi possible que la famille *Shanks* soit, d'une certaine façon, liée à ce sous-groupe atypique. Les séquences protéiques des deux ORF de *Shanks* ont été comparées à des assemblages de génomes d'Haptophytes, comme indiqué précédemment. Cette analyse a permis d'identifier une séquence similaire dans le génome de *C. tobin* et 12 copies d'une potentielle famille semblable à *Shanks* dans le génome de *E. huxleyi* (Annexe 11). Les séquences identifiées chez *E. huxleyi* sont présentes sur NCBI et ne possèdent aucune fonction prédite. Leur nature était jusqu'alors probablement inconnue.

Bilan de la caractérisation des familles d'ETs *Luffy*, *Ace* et *Shanks*

L'ensemble de ces résultats confirme que les éléments autonomes des familles *Luffy*, *Ace* et *Shanks* présentent bien des ORF codant pour des transposases et différentes signatures structurales confirmant leur appartenance à l'Ordre des TIR et aux Superfamilles Mariner et hAT, respectivement. Le faible pourcentage de similarité de leur transposase respective, par rapport à celles des TIR Mariner ou hAT qui sont répertoriées dans la banque de données Repbase, confirme la divergence de ces éléments par rapport à ceux appartenant à d'autres phylum. Au contraire, des séquences semblables, correspondant probablement à des d'ETs, ont pu être détectées dans les génomes d'autres Haptophytes.

Cette caractérisation a également permis d'identifier un grand nombre d'éléments non-autonomes pour les familles *Ace* et *Shanks*. Ces éléments pourraient avoir participé de manière importante à la dynamique du génome des souches domestiquées de *T. lutea*. En effet, les deux délétions prédites par les outils Breakdancer et Pindels pour la souche Tiso-LipidePlus correspondent à des ETs non-autonomes de ces familles. De même, trois événements d'insertions identifiés par LoRTE et communs aux prédictions de Mobster correspondent à des éléments non-autonomes de *Ace* et *Shanks* (Tableau 10).

Pour la famille *Luffy*, la recherche d'éléments non-autonomes avec le programme d'annotation TEannot n'a pas été concluante. Ceci suggère que les indels prédites pour cette famille sont causées uniquement par des ETs autonomes. De même, LoRTE a prédit un événement d'insertion commun à Mobster pour la famille *Luffy* et celui-ci est associé à un élément autonome (Tableau 10). L'ensemble de ces hypothèses reste cependant à être validé par PCR et/ou re-séquençage.

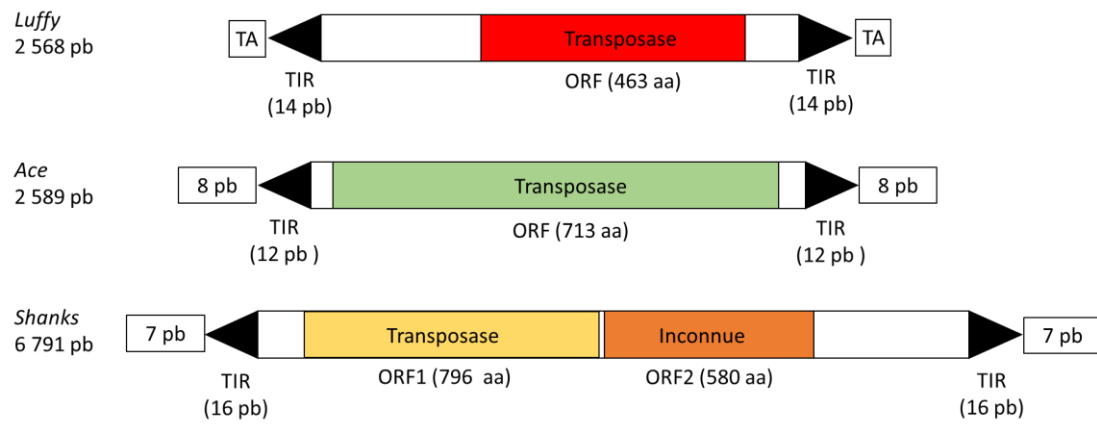


Figure 52 : Représentation schématique des séquences autonomes des familles *Luffy*, *Ace* et *Shanks*.

5. Existe-t-il un lien entre les indels associées aux ETs et les phénotypes des souches domestiquées ?

Cette étude a permis de prédire plusieurs événements d'indels associées à des ETs, spécifiques aux souches Tiso-LipidePlus (146 indels) et Tiso-LipideMoins (32 insertions). Il est maintenant bien établi que les d'ETs peuvent impacter l'expression et/ou la fonction des gènes, par exemple en s'insérant à l'intérieur ou à proximité de ceux-ci (Casacuberta et González, 2013). Afin d'établir une éventuelle relation entre la position des différentes indels et les phénotypes observés (Tiso-LipidePlus et Tiso-LipideMoins) une analyse de l'ensemble des gènes pouvant être impactés par ces variations génétiques a été réalisée.

5.1. Recherche de gènes potentiellement impactés par les indels

Les gènes situés à proximité d'indels (situées entre + 1000 et - 1000 pb de la position d'un gène) ont été recherchés dans le génome de référence de *T. lutea* (Bertheliet et al. 2018). Parmi les 146 événements d'indels prédites chez la souche Tiso-LipidePlus, 70 insertions et 2 délétions sont situées à proximité d'un gène. Parmi ces gènes, 41 ont une annotation fonctionnelle automatique qui est associée à une ontologie (identifiée par le programme BLAST2GO, Conesa et al., 2005; Bertheliet et al., 2018). Une analyse similaire a été conduite chez la souche Tiso-LipideMoins et parmi les 32 potentielles insertions détectées, 10 sont situées à proximité d'un gène et possèdent une fonction prédite.

Différents termes GO (ontologie de gène) correspondant à deux processus biologiques ont été recherchés (Tableaux 11 et 12) afin d'identifier les gènes candidats dont la fonction peut être liée aux phénotypes lipidiques affichés par ces souches :

- le terme « *lipid* » en raison du caractère spécifique en lipides neutres de ces deux souches domestiquées.
- le terme « *Catabolic* » car une accélération ou une diminution du catabolisme des lipides neutres pourrait expliquer les caractères lipidiques des souches domestiquées (Carrier et al. 2014 ; Garnier et al. 2016).

Tableau 11 : Liste des gènes candidats pour la souche Tiso-LipidePlus.

Les enzymes prédites pour ces gènes sont associées aux GO (ontologie de gène) « *Lipid* » et « *Catabolic* ». Les prédictions concernant les noms de l'enzyme et les GO ont été réalisées par BLAST2GO.

Famille d'ETs	indels	Fréquence allélique (%)	Contigs	Position indels	Gène ID	Enzyme prédite	GO	Position indel/gène
<i>Shanks</i>	Del	27	142	46892-48002	15006	Lipase Class 3	<i>Lipid</i>	Amont (-50 pb)
<i>Shanks</i>	In	23	16	260041	13322	Arylsulfatase	<i>Lipid, Catabolic</i>	Intérieur
<i>Shanks</i>	In	28	86	172454	10131	Cyclin-dependent kinase A-1	<i>Lipid, Catabolic</i>	Amont (-511 pb)
<i>Ace</i>	In	29	86	411209	10190	Acyl-lipid (8-3)-desaturase	<i>Lipid</i>	Amont (-819 pb)
<i>Ace</i>	In	27	162	603564	08455	Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyl transferase	<i>Lipid</i>	Amont (-789 pb)
<i>Ace</i>	In	26	68	75062	11876	Sphingolipid delta(4)-desaturase	<i>Lipid</i>	Amont (-954 pb)
<i>Ace</i>	In	64	170	217539	13807	40S ribosomal S8	<i>Catabolic</i>	Aval (+819 pb)
<i>Shanks</i>	In	47	127	894448	03165	Proteasome subunit beta type-2	<i>Catabolic</i>	Aval (+729 pb)
<i>Ace</i>	In	38	189	830328	06146	60S ribosomal L10-2	<i>Catabolic</i>	Amont (-324 pb)
<i>Luffy</i>	In	37	179	341961	09731	Cystathionine gamma-synthase	<i>Catabolic</i>	Amont (-27 pb)
<i>Ace</i>	In	27	97	1811743	02964	na	<i>Catabolic</i>	Aval (+853 pb)

Tableau 12 : Liste des gènes candidats pour la souche Tiso-LipideMoins.

Les enzymes prédites pour ces gènes sont associées aux GO (ontologie de gène) « *Lipid* », « *Catabolic* ». Les prédictions des enzymes et de GO ont été réalisées par BLAST2GO.

Famille d'ETs	Indels	Fréquence allélique (%)	Contigs	Position Indels	Gene ID	Enzyme prédite	GO	Position indel/gène
<i>Shanks</i>	In	20	142	157577	15024	1-acylglycerol-3-phosphate O-acyltransferase	<i>Lipide</i>	Aval (+935 pb)
<i>Luffy</i>	In	23	37	500851	10464	1-phosphatidylinositol-4-phosphate 5-kinase	<i>Lipide</i>	Aval (+124 pb)
<i>Gypsy1</i>	In	11	89	35380	09514	60S ribosomal L27a	<i>Catabolic</i>	Amont (-697 pb)

5.2. Analyse des gènes candidats associés au métabolisme des lipides neutres

Chez la souche domestiquée Tiso-LipidePlus, onze gènes candidats codent pour des protéines liées au métabolisme des lipides et aux processus cataboliques. Après vérification, quatre gènes sont impliqués dans la synthèse des lipides membranaires (prédictions : *Arylsulfatase*, *Cyclin-dependent kinase A-1*, *Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyl transferase* et *Sphingolipid delta(4)-desaturase* ; Tableau 11) et deux peuvent être associés à la synthèse des lipides neutres : une lipase (prédiction : *Lipase Class 3*) et une désaturase d'acide gras (prédiction : *Acyl-lipid (8-3)-desaturase*). Concernant la souche Tiso-LipideMoins, l'analyse n'a pas été concluante car aucun gène ne participe à la synthèse des lipides neutres. Deux gènes participent à la synthèse des lipides membranaires (prédictions : *1-acylglycerol-3-phosphate O-acyltransferase* et *1-phosphatidylinositol-4-phosphate 5-kinase* ; Tableau 12).

Les deux gènes candidats identifiés dans le génome de la souche domestiquée Tiso-LipidePlus ont été examinés plus en détail, afin d'estimer quel impact leur modification pourrait causer au niveau du métabolisme des lipides neutres de *T. lutea*.

Gène candidat 1: lipase

Une délétion d'un ET non-autonome de la famille *Shanks* a été prédite par Pindel et Breakdancer, en amont d'un gène codant pour une lipase (Tableau 11). Les lipases sont associées au métabolisme des lipides neutres car elles participent à l'hydrolyse de ces molécules. En effet, elles réalisent la libération de trois molécules d'acides gras ainsi que d'une molécule de glycérol qui compose les TAG. Les lipides neutres sont majoritairement constitués de TAG. Ainsi, chez l'homme et la drosophile, la déficience de l'expression de gènes codants pour des lipases est associée à des phénotypes accumulant plus de TAG par rapport aux organismes contrôles (Grönke et al., 2005; Zimmermann et al., 2018).

Le rôle des lipases a également été étudié chez les microalgues (Nomaguchi et al., 2018). Chez les Diatomée *T. pseudonana* et *P. tricornutum*, des études ont montrées que la répression de gènes codant pour des lipases conduit à une augmentation de l'accumulation de lipides neutres (Trentacoste et al., 2013; Barka et al., 2016).

Chez les Haptophytes, l'expression de gènes codants pour des lipases a précédemment été rapportée pour les espèces *Isochrysis galbana* et *T. lutea* (Godet et al., 2012; Carrier et al., 2014). Dans cette dernière étude, les auteurs ont comparé le nombre de transcrits respectifs pour les gènes d'une souche domestiquée de *T. lutea* (accumulant plus de lipide) et la souche sauvage d'origine (Tiso-CCAP927/14). Ils ont identifié la présence de trois gènes codant pour des lipases et ont montré que l'un d'entre eux est sous-exprimé chez la souche domestiquée de *T. lutea*. Ceci a été confirmé par RT-qPCR (Carrier et al., 2014).

A partir des données transcriptomiques de cette précédente étude, il a été possible d'identifier que le gène détecté ici n'est pas le même que celui identifié par Carrier et al. (2014). Cependant, ce gène est aussi exprimé dans un transcriptome issue de la souche sauvage Tiso-CCAP927/14 (Carrier et al., 2014).

Ainsi, la délétion de *Shanks* à proximité d'un gène codant pour une lipase pourrait réprimer l'expression de celui-ci, donnant lieu à une diminution de l'hydrolyse des TAG et conduisant à leur suraccumulation dans la souche Tiso-LipidePlus.

Gene candidat 2 : désaturase

L'insertion d'un ETs associé à la famille *Ace* a été prédite en amont d'un gène codant pour une désaturase d'acide gras (Tableau 11). Après avoir comparé ce gène sur NCBI, celui-ci a été précédemment caractérisé chez *Isochrysis galbana*, *Isochrysis sphaerica* et *T. lutea* (BLASTx, 100% identique à AFJ74710.1). Ce gène code pour une enzyme désaturase delta-4 qui participe à la conversion des acides gras polyinsaturés DPA en DHA (Pereira et al., 2004; Shi et al., 2012; Guo et al., 2013; Huerlimann et al., 2014). L'insertion d'un ET en amont de ce gène pourrait éventuellement changer l'expression de celui-ci et participer ainsi à la diminution ou à l'augmentation de la synthèse de DHA.

Chez la Chlorophyte *C. reinhardtii*, une analyse de génétique comparative a été réalisée entre une souche sauvage et une souche mutante, présentant une réduction d'accumulation

de 65% d'acides gras omega-3 par rapport à la souche sauvage dont elle est issue (Nguyen et al., 2013). Après introspection, les auteurs ont identifié l'insertion d'un ET au niveau de la région promotrice d'un gène codant pour une désaturase. La restauration du promoteur de celui-ci a permis de rétablir un profil d'acide gras équivalent à celui de la souche sauvage (Nguyen et al., 2013). L'insertion d'un ET en amont d'un gène codant pour une désaturase pourrait ainsi impacter l'accumulation des acide gras chez la souche Tiso-LipidePlus. Cependant, le lien entre cette mutation et le phénotype de suraccumulation lipidique n'est pas manifeste et nécessite des expériences complémentaires.

En raison d'un manque de temps, la confirmation de la spécificité de ces deux événements d'indels associées à des ETs et l'analyse de l'expression différentielle de ces deux gènes candidats n'a pu être achevée lors de l'écriture de ce manuscrit. Cependant, cette expérimentation est actuellement en cours de réalisation au laboratoire PBA.

6. Conclusion

Les ETs participent à la dynamique du génome des souches domestiquées de *T. lutea*

Cette étude a permis de mettre en évidence la dynamique d'ETs dans le génome des souches domestiquées par rapport au génome de la souche sauvage dont elles sont issues. En effet, 146 potentielles indels ont été identifiées chez la souche Tiso-LipidePlus et 32 chez la souche Tiso-LipideMoins. Pour ces deux souches, trois familles appartenant à l'Ordre des TIR présentent une dynamique dominante : les familles Mariner *Luffy*, hAT *Ace* et hAT *Shanks*. L'ensemble de ces analyses supportent ainsi les résultats du premier chapitre de cette thèse. En effet, les familles *Ace* et *Luffy* ont précédemment été identifiées comme potentiellement actives (Bertheliet al., 2018). Ce résultat indique que la méthode employée dans le Chapitre 1 est bien adaptée. Ensuite, une nouvelle famille d'ETs nommée *Shanks*, a été découverte dans ce Chapitre 2 et n'avait pas été caractérisée dans le Chapitre 1. Ceci confirme la complexité d'identifier de façon automatique les ETs dans le génome d'un organisme non-modèle, ainsi que le manque important de séquences d'ETs dans les banques de données, en particulier pour les Phylum peu étudiés comme celui des Haptophytes. Il est probable que d'autres éléments atypiques n'aient pas encore été caractérisés dans le génome de *T. lutea*. Enfin, la confirmation de ces événements d'indels reste tout de même à valider à l'aide d'une amplification par PCR et/ou le re-séquençage des régions prédites.

Existe-t-il un lien entre les indels et les phénotypes des souches domestiquées ?

Dans l'hypothèse où les prédictions d'indels associées aux ETs seraient responsables du caractère amélioré de ces souches, les gènes proches de leurs positions ont été recensés. Ceux dont la fonction prédite est impliquée dans la voie métabolique des lipides neutres ont été examinés. Tandis que aucun gène candidat n'a été identifié pour la souche Tiso-LipideMoins, deux potentiels gènes candidats, une lipase et une desaturase, ont été identifiés pour la souche Tiso-LipidePlus.

Existe-t-il un lien entre l'origine des souches domestiquées et le nombre d'indels associées aux ETs ?

Quatre fois plus d'indels associées aux ETs ont été prédits dans le génome de la souche Tiso-LipidePlus par rapport à celui de la souche Tiso-LipideMoins. Ce différentiel pourrait être lié à l'origine des souches. En effet, tandis que la souche Tiso-LipideMoins a été obtenue à l'issue d'un programme d'amélioration composé uniquement d'étapes de sélection, la souche Tiso-LipidePlus a été obtenue par un programme comportant deux étapes de sélection et deux étapes de mutagenèse par rayons ultraviolets. Les rayons UV sont connus pour induire l'expression des ETs chez des organismes eucaryotes multicellulaires tels que le maïs, la drosophile ou le melon *Cucumis melo* (Walbot, 1992; Ramallo et al., 2008; Jardim et al., 2015), mais également chez l'organismes unicellulaires *Saccharomyces cerevisiae* (Rolfe et al., 1986). Il est ainsi probable qu'une partie des indels associées aux ETs dans le génome de la souche Tiso-LipidePlus ait été généré lors des étapes de mutagenèse par rayonnements UV. Néanmoins, les données de séquençage à disposition proviennent des souches actuellement en collection au laboratoire PBA. Des prélèvements d'échantillons n'ont pas été réalisés avant et après les étapes de mutagenèse. Cela permettrait d'attester cette hypothèse. La réalisation d'une nouvelle expérimentation serait donc nécessaire pour le vérifier.

Conclusion générale et perspectives

Dans la littérature, la microalgue *T. lutea* tend à devenir une espèce modèle des Haptophytes. En effet, de nombreuses études éco-physiologiques, biochimiques et « omiques » s'accumulent de plus en plus à son sujet.

L'objectif de ce travail était d'initier la première étude de la dynamique du génome de l'haptophyte *Tisochrysis lutea*, à travers l'étude des ETs.

Chez les Haptophytes, peu de travaux ont été réalisés sur ces séquences d'ADN mobiles (Read et al., 2013). Ainsi, ce travail compilant la diversité, l'activité et la dynamique des ETs dans le génome de la microalgue *T. lutea* ouvre à de futures recherches fondamentales, appliquées et environnementales.

Le génome de *T. lutea* abrite une grande variété de familles d'ETs

Le premier objectif de cette thèse a été de réaliser une annotation exhaustive des ETs présents dans le génome de la microalgue *T. lutea*, et en particulier, d'identifier des éléments potentiellement autonomes. A cette occasion, un nouvel assemblage du génome de *T. lutea* a été établi et un nouveau pipeline d'analyse nommé PiRATE a été créé. Ce pipeline a été conçu dans l'idée d'identifier le plus exhaustivement possible les ETs chez des espèces non modèles et d'identifier en priorité les éléments autonomes complets. Le pipeline PiRATE a été publié et présenté dans plusieurs congrès, notamment lors de la session « *Digital Tools* » du *Plant and Animal Genomes Conference XXVI* et a été cité dans le « meeting report » de celui-ci (Beiki et al., 2018). Enfin, une seconde version incorporant une amélioration de son étape de classification est actuellement en préparation et sera prochainement publiée.

Dans le premier chapitre de cette thèse, une grande variété de familles d'ETs appartenant à la Classe I (LTR, LINE et SINE) et à la Classe II (TIR) ont été détectées dans le génome de *T. lutea*. Suite à l'annotation des séquences répétées, nous avons estimé que 20% du génome de *T. lutea* est constitué d'ETs bien caractérisés et que 18% est constitué d'éléments répétés encore non caractérisés (Bertheliet et al., 2018). La caractérisation de la famille *Shanks* dans le Chapitre 2 en est un bon exemple. Considérant le faible nombre d'études visant l'identification et la classification des ETs chez les microalgues, il est plausible que les génomes de microalgues abritent des ETs atypiques.

Il serait ainsi pertinent de conduire une étude à grande échelle visant l'identification des ETs chez les espèces de microalgues. Une centaine d'assemblage de génomes d'espèces différentes sont actuellement disponibles (Annexe 1). Cette analyse permettrait de comparer la diversité et la proportion des ETs dans les génomes d'espèces de différents phylums. L'obtention d'une banque de données d'ETs et de séquences répétées spécifiques aux microalgues permettrait d'améliorer l'efficacité des outils de classification des pipelines existants.

Ensuite, l'étude de la diversité des ETs chez les microalgues pourrait aider à mettre en évidence la présence de potentiels événements de transferts horizontaux d'ETs entre les génomes de microalgues. De plus, considérant que celles-ci vivent en consortium avec des virus, il serait intéressant d'explorer si de tels transferts pourraient subvenir par l'intermédiaire de ces derniers.

Découverte d'éléments actifs chez *T. lutea*

Le second objectif de la thèse a été d'étudier l'activité des ETs dans le génome de *T. lutea*.

L'activité potentielle de 17 familles d'éléments de Classe I et Classe II a tout d'abord été révélée en condition de culture limitée en azote (Chapitre 1). Parmi ces éléments, les familles *Luffy* et *Ace* ont été identifiées parmi les trois transcrits les plus abondants chez *T. lutea*, suggérant la présence de plusieurs copies actives. Leur expression respective a été également détectée en condition de forte concentration en iode (Hernández Javier et al., 2018). Ensuite, les transposases respectives de ces deux familles d'éléments ont été identifiées dans un protéome de *T. lutea* obtenu en condition de limitation en azote. Ces résultats confirment la capacité d'expression de ces éléments et la synthèse de leur transposase par la machinerie cellulaire de cette microalgue.

L'expression de ces deux familles d'éléments (*Luffy* et *Ace*) indique que celles-ci pourraient être actives lors de conditions de stress (limitation en azote et/ou iode). Cependant, le manque de conditions contrôles ne permet pas de confirmer ces hypothèses. Il serait ainsi intéressant de comparer l'expression de ces éléments par RT-qPCR lors de conditions de stress vis-à-vis d'une condition contrôle (Maumus et al., 2009; Egue et al., 2015; Chen et al., 2017a).

Les ETs participent à la dynamique du génome de *T. lutea*

Dans cette étude, la dynamique des ETs dans deux souches domestiquées de *T. lutea* nomées Tiso-LipidePlus et TisoLipideMoins a été analysée par la comparaison de leur génome respectif à celui de leur souche sauvage d'origine. Cette exploration a conduit à la découverte d'un grand nombre de potentielles indels (146 et 32) associées à des ETs de Classe I et Classe II spécifiques à ces deux souches. Néanmoins, l'implication prépondérante de trois familles de Classe II a été distinguée. Celle des familles *Luffy* et *Ace*, qui ont été identifiées lors du Chapitre 1, et celle d'une nouvelle famille d'ETs nommée *Shanks*, caractérisée lors du Chapitre 2.

Considérant le grand nombre d'éléments non-autonomes associés aux familles *Ace* et *Shanks*, il est probable qu'une part importante des indels prédites dans le génome de ces souches ait pour origine l'activité de ce type d'éléments.

Il apparait donc que les familles *Luffy*, *Ace* et *Shanks* participent fortement à la dynamique des génomes des souches domestiquées et de manière générale à la plasticité du génome de l'espèce *T. lutea*.

Quel est l'importance de l'activité des ETs dans un contexte d'amélioration des microalgues ?

Il est envisageable que les indels associées aux ETs et qui sont spécifiques aux deux souches domestiquées de *T. lutea* puissent avoir un lien avec les différents programmes d'amélioration appliqués à la souche d'origine. Deux scénarios sont envisageables :

Dans un premier scénario, il est probable que les ETs aient été mobilisés durant l'application des programmes d'amélioration. En effet, les indels associées aux ETs qui ont été prédites dans le génome de la souche Tiso-LipidePlus issue d'un programme d'amélioration de « mutation-sélection » sont quatre fois plus nombreuses par rapport au génome de la souche Tiso-LipideMoins issue d'un programme d'amélioration par « sélection ». Ce constat pourrait suggérer un effet important de l'étape de « mutation » opérée par des radiations UV sur l'activité des ETs. L'induction de l'activité des ETs par les UV est connue chez d'autres organismes (Rolfe et al., 1986; Ramallo et al., 2008). L'UV pourrait constituer un stress pertinent pour induire l'activité des ETs de *T. lutea* et ainsi favoriser l'apparition d'un large panel de mutations. Ensuite, étant donné que le caractère phénotypique amélioré des souches domestiquées est conservé depuis leur obtention (sept et quatre ans pour la souche Tiso-LipidePlus et Tiso-LipideMoins respectivement), il est probable que le génome de celles-ci soit relativement stable et que la majorité des transpositions aient eu lieu lors du programme d'amélioration. Ces indels auraient ainsi pu favoriser l'apparition de caractères lipidiques d'intérêt chez des individus, qui auraient ensuite été sélectionnés avec le cytomètre en flux. Des populations de ces individus, présentant de la diversité génétique, pourraient aujourd'hui constituer la souche domestiquée Tiso-LipidePlus. Une insertion d'ET a déjà été identifiée comme la cause de l'apparition d'un phénotype lipidique mutant chez la microalgue *C. reinhardtii* (Nguyen et al., 2013).

Pour étudier les effets de ces deux programmes d'amélioration (« mutation-sélection » et « sélection ») sur l'activité des ETs, il serait nécessaire de les réitérer en collectant des échantillons voués au séquençage. Ceci pourrait être conduit à partir d'une souche clonale

afin de faciliter l'identification des indels. Des échantillons pourraient être recueillis avant et après la réalisation de ces programmes d'amélioration.

Dans un second scénario, il est possible que les souches domestiquées présentent des familles d'ETs constitutivement actives et mobiles et donc un génome plus dynamique. Celui-ci paraît moins plausible considérant la stabilité du caractère lipidique de ces souches (sept ans pour la souche Tiso-LipidePlus et quatre ans pour la souche Tiso-LipideMoins). Ce scénario ne serait pas souhaitable dans un contexte de domestication de souches promises à des applications biotechnologiques. A l'heure actuelle, le taux de mutations spontanées de *T. lutea* est inconnu. Il serait donc intéressant de comparer le niveau d'expression de ces éléments chez ces différentes souches dans différentes conditions de cultures. Il pourrait également être pertinent de conduire des expériences d'évolution dirigée sur des centaines de générations afin d'estimer le taux de mutations associées aux ETs dans le génome des souches sauvages et domestiquées. Précédemment, des études mesurant les taux de mutations ont été réalisées chez les Chlorophytes (Ness et al., 2012; Sung et al., 2012; Ness et al., 2015; Krasovec et al., 2017, 2018). Néanmoins, ces études se sont uniquement intéressées à l'étude des SNPs et des petites indels. Au vu de l'importance de la dynamique des ETs soulignée ici, leur participation à la dynamique des génomes de microalgues devrait être davantage prise en compte dans le futur.

L'activité des ETs pourraient t'elle être employé pour l'amélioration des microalgues ?

Dans la littérature, des études commencent à souligner plusieurs aspects problématiques de l'utilisation d'agents mutagènes UV et chimiques lors de programmes d'amélioration de type « sélection-mutation » (Krasovec et al., 2018). Ces agents pourraient favoriser l'augmentation du taux de mutations délétères et la sélection indirecte de cellules résistantes à ces mutagène. De même, les souches de microalgues obtenues par ces agents sont maintenant considérées comme des OGM (organismes génétiquement modifiés) et pourront donc être soumises à de nombreuses conditions commerciales et politiques (Spicer et Molnar, 2018).

Une idée alternative à la mutagenèse aléatoire pourrait être l'utilisation d'étapes de « mutation » n'utilisant pas d'agents mutagènes, mais des stress pouvant être considérés comme plus « naturels » comme des chocs thermiques. En effet, dans l'hypothèse où les ETs représentent une part importante des mutations induites par les agents mutagènes, leur

induction par de tels stress pourrait être suffisante pour créer de la variabilité génétique et faire apparaître des phénotypes améliorés. Par le passé, des conditions de stress thermiques ont par exemple été employés pour induire la mobilité d'ETs et générer des individus mutants de la microalgue *V. carteri* (Miller et al., 1993; Ueki et Nishii, 2008).

Quelle est l'importance de l'activité des ETs dans les mécanismes d'adaptation évolutive des microalgues ?

Tout comme chez la plupart des organismes, l'activité des ETs est de plus en plus suggérée comme un processus important dans l'évolution adaptative des microalgues face aux pressions environnementales (Maumus et al., 2009; Lescot et al., 2016; Chen et al., 2017a). En effet, les microalgues ont pour caractéristiques d'avoir un temps de génération court, d'être majoritairement unicellulaires, avec un cycle de vie souvent haploïde. Ainsi, l'activité et la mobilité des ETs pourraient rapidement conduire à la sélection de nouveaux phénotypes, présentant des mutations associées aux ETs et leur procurant un avantage adaptatif par rapport à une pression extérieure (Chen et al., 2017a).

En plus de leur capacité à créer un large panel de mutations lors de leur déplacement (Casacuberta et González, 2013), les ETs pourraient participer à l'adaptation évolutive des microalgues à travers d'autres mécanismes comme la rétro-positions (Tan et al., 2016). Pour illustrer ce propos, la présence de très nombreux rétro-gènes chez les génomes des espèces de Dinophytes *Symbiodinium* (22% et 23% des gènes totaux chez *S. mimitum* et *S. kawagutii*, respectivement) suggère l'implication d'ETs de Classe I dans ce processus (Shoguchi et al., 2013; Lin et al., 2015; Tan et al., 2016). Une étude de génomique comparative, montre que deux épisodes majeurs d'explosion de rétro-gènes coïncident avec deux périodes dramatiques de changement climatique sur terre (il y a ~60 et 6 millions d'années)(Song et al., 2018). Cela suggère que le mécanisme de rétro-position aurait facilité l'adaptation de ces microalgues face à ces épisodes de changements climatiques (Song et al., 2017, 2018). Ces spéculations font ainsi écho au contexte actuel de changement climatique. Bien que celles-ci restent à être considérées avec prudence, elles peuvent suggérer un rôle important des ETs dans les capacités adaptatives des microalgues face aux changements actuels rencontrés dans les océans, telles que l'augmentation de la température, l'acidification ou l'eutrophisation.

Pour explorer l'importance des ETs, une perspective d'étude serait d'étudier l'activité de ces éléments chez différents Phylum de microalgues lors d'une efflorescence microalgale. Ces phénomènes, telles que les marées rouges, ont lieu en raison de différents critères naturels, mais leur nombre semble augmenter à cause d'impacts anthropiques comme le déversement d'éléments nutritifs dans les eaux côtières (Almeda et al., 2018). Ces phénomènes sont ainsi de plus en plus fréquents dans le monde et peuvent se répéter dans des zones à risques. Il serait intéressant d'étudier l'impact de ces stress sur l'activité des ETs des microalgues à l'aide d'un échantillonnage et d'un séquençage méta-transcriptomique (Lescot et al., 2016). La comparaison des méta-transcriptomes avant et lors de différentes phases de l'efflorescence pourrait permettre d'identifier l'induction de l'activité de familles d'ETs chez différentes espèces de microalgues. Ceci pourrait mettre en évidence le rôle *in natura* de ces éléments dans l'évolution des génomes de microalgues et leur possible implication dans l'adaptation évolutive de ces espèces lors de stress environnementaux.

Références

-
- Adams, M.D. (2000). The Genome Sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195.
- Alkan, C., Sajjadian, S., and Eichler, E.E. (2011). Limitations of next-generation genome sequence assembly. *Nature Methods* 8, 61–65.
- Alkhamis, Y., and Qin, J.G. (2015). Comparison of pigment and proximate compositions of *Tisochrysis lutea* in phototrophic and mixotrophic cultures. *Journal of Applied Phycology*.
- Almeda, R., Cosgrove, S., and Buskey, E.J. (2018). Oil Spills and Dispersants Can Cause the Initiation of Potentially Harmful Dinoflagellate Blooms (“Red Tides”). *Environmental Science & Technology* 52, 5718–5724.
- Alonso, D.L., del Castillo, C.I.S., Sánchez, J.L.G., Sánchez Pérez, J.A., and Camacho, F.G. (1994). Quantitative Genetics of Fatty Acid Variation in *Isochrysis Galbana* (prymnesiophyceae) and *Phaeodactylum Tricornutum* (bacillariophyceae)1. *Journal of Phycology* 30, 553–558.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215, 403–410.
- Alzohairy, A.M., Gyulai, G., Jansen, R.K., and Bahieldin, A. (2013). Transposable elements domesticated and neofunctionalized by eukaryotic genomes. *Plasmid* 69, 1–15.
- Amato, A., Sabatino, V., Nylund, G.M., Bergkvist, J., Basu, S., Andersson, M.X., Sanges, R., Godhe, A., Kjørboe, T., Selander, E., et al. (2018). Grazer-induced transcriptomic and metabolomic response of the chain-forming diatom *Skeletonema marinoi*. *The ISME Journal* 12, 1594–1604.
- Andersen, R.A., Kim, J.I., Tittley, I., and Yoon, H.S. (2014). A re-investigation of *Chrysotila* (Prymnesiophyceae) using material collected from the type locality. *Phycologia* 53, 463–473.
- Anesio, A.M., and Laybourn-Parry, J. (2012). Glaciers and ice sheets as a biome. *Trends in Ecology & Evolution* 27, 219–225.
- Aristote (343AD). *Historia Animalium*.
- Armbrust, E.V. (2004). The Genome of the Diatom *Thalassiosira Pseudonana*: Ecology, Evolution, and Metabolism. *Science* 306, 79–86.
- Aziz, R.K., Breitbart, M., and Edwards, R.A. (2010). Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Research* 38, 4207–4217.
- Bao, Z., and Eddy, S.R. (2002). Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. *Genome Research* 12, 1269–1276.
- Barka, F., Angstenberger, M., Ahrendt, T., Lorenzen, W., Bode, H.B., and Büchel, C. (2016). Identification of a triacylglycerol lipase in the diatom *Phaeodactylum tricornutum*. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1861, 239–248.
- Beiki, H., Eveland, A.L., and Tuggle, C.K. (2018). Recent advances in plant and animal genomics are taking agriculture to new heights. *Genome Biology* 19.

Belon (1551). L'Histoire naturelle des estranges poissons marins avec la vraie peinture et description du Dauphin et de plusieurs autres de son espèce observée par Pierre Belon du Mans.

Bendif, E.M., Probert, I., Hervé, A., Billard, C., Goux, D., Lelong, C., Cadoret, J.-P., and Véron, B. (2011). Integrative Taxonomy of the Pavlovophyceae (Haptophyta): A Reassessment. *Protist* 162, 738–761.

Bendif, E.M., Probert, I., Schroeder, D.C., and de Vargas, C. (2013). On the description of *Tisochrysis lutea* gen. nov. sp. nov. and *Isochrysis nuda* sp. nov. in the Isochrysidales, and the transfer of Dicrateria to the Prymnesiales (Haptophyta). *Journal of Applied Phycology* 25, 1763–1776.

Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research* 40, e72–e72.

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

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

Berthelie, J., Schnitzler, C.E., Wood-Charlson, E.M., Poole, A.Z., Weis, V.M., and Detournay, O. (2017). Implication of the host TGF β pathway in the onset of symbiosis between larvae of the coral *Fungia scutaria* and the dinoflagellate *Symbiodinium* sp. (clade C1f). *Coral Reefs* 36, 1263–1268.

Berthelie, J., Casse, N., Daccord, N., Jamilloux, V., Saint-Jean, B., and Carrier, G. (2018). A transposable element annotation pipeline and expression analysis reveal potentially active elements in the microalga *Tisochrysis lutea*. *BMC Genomics* 19, 378.

Biemont, C. (2010). A Brief History of the Status of Transposable Elements: From Junk DNA to Major Players in Evolution. *Genetics* 186, 1085–1093.

Blanc-Mathieu, R., Krasovec, M., Hebrard, M., Yau, S., Desgranges, E., Martin, J., Schackwitz, W., Kuo, A., Salin, G., Donnadiou, C., et al. (2017). Population genomics of picophytoplankton unveils novel chromosome hypervariability. *Science Advances* 3, e1700239.

Bonnefond, H., Grimaud, G., Rumin, J., Bougaran, G., Talec, A., Gachelin, M., Boutoute, M., Pruvost, E., Bernard, O., and Sciandra, A. (2017). Continuous selection pressure to improve temperature acclimation of *Tisochrysis lutea*. *PLOS ONE* 12, e0183547.

Bopp, L., Bowler, C., Guidi, L., Karsenti, E., and de Vargas, C. (2002). L'océan, pompe à carbone.

Bougaran, G., Rouxel, C., Dubois, N., Kaas, R., Grouas, S., Lukomska, E., Le Coz, J.-R., and Cadoret, J.-P. (2012). Enhancement of neutral lipid productivity in the microalga *Isochrysis affinis Galbana* (T-Iso) by a mutation-selection procedure. *Biotechnology and Bioengineering* 109, 2737–2745.

Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R.P., et al. (2008). The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* 456, 239–244.

Brownlee, C., Wheeler, G.L., and Taylor, A.R. (2015). Coccolithophore biomineralization: New questions, new answers. *Seminars in Cell & Developmental Biology* 46, 11–16.

-
- Brune, D.E., Lundquist, T. J., and Benemann, J. R. (2009). Microalgal Biomass for Greenhouse Gas Reductions: Potential for Replacement of Fossil Fuels and Animal Feeds. *Journal of Environmental Engineering* 135, 1136–1144.
- Cadoret, J.-P., Garnier, M., and Saint-Jean, B. (2012). Microalgae, Functional Genomics and Biotechnology. In *Advances in Botanical Research*, (Elsevier), pp. 285–341.
- Capy, P. (2005). Classification and nomenclature of retrotransposable elements. *Cytogenetic and Genome Research* 110, 457–461.
- Capy, P., Gasperi, G., Biémont, C., and Bazin, C. (2000). Stress and transposable elements: co-evolution or useful parasites? *Heredity* 85, 101–106.
- Capy P, Bazin C, Higuete D, and Langin T (1997). Dynamic and Evolution of Transposable Elements (RG Landes Company, Austin 1997b). RG Landes Company, Austin.
- Carr, M., Bensasson, D., and Bergman, C.M. (2012). Evolutionary Genomics of Transposable Elements in *Saccharomyces cerevisiae*. *PLoS ONE* 7, e50978.
- Carrier, G., Garnier, M., Le Cunff, L., Bougaran, G., Probert, I., De Vargas, C., Corre, E., Cadoret, J.-P., and Saint-Jean, B. (2014). Comparative Transcriptome of Wild Type and Selected Strains of the Microalgae *Tisochrysis lutea* Provides Insights into the Genetic Basis, Lipid Metabolism and the Life Cycle. *PLoS ONE* 9, e86889.
- Carrier, G., Baroukh, C., Rouxel, C., Duboscq-Bidot, L., Schreiber, N., and Bougaran, G. (2018). Draft genomes and phenotypic characterization of *Tisochrysis lutea* strains. Toward the production of domesticated strains with high added value. *Algal Research* 29, 1–11.
- Casacuberta, E., and González, J. (2013). The impact of transposable elements in environmental adaptation. *Molecular Ecology* 22, 1503–1517.
- Charrier, A., Bérard, J.-B., Bougaran, G., Carrier, G., Lukomska, E., Schreiber, N., Fournier, F., Charrier, A.F., Rouxel, C., Garnier, M., et al. (2015). High-affinity nitrate/nitrite transporter genes (*Nrt2*) in *Tisochrysis lutea* : identification and expression analyses reveal some interesting specificities of Haptophyta microalgae. *Physiologia Plantarum* 154, 572–590.
- Chaturvedi, R., and Fujita, Y. (2006). Isolation of enhanced eicosapentaenoic acid producing mutants of *Nannochloropsis oculata* ST-6 using ethyl methane sulfonate induced mutagenesis techniques and their characterization at mRNA transcript level. *Phycological Research* 54, 208–219.
- Chedhomme, M., and Badawi (2018). Éluclation de l'origine d'éléments génétiques répétés de micro-algue. Mémoire de master 1.
- Chen, J.E., Cui, G., Wang, X., Liew, Y.J., and Aranda, M. (2017a). Recent expansion of heat-activated retrotransposons in the coral symbiont *Symbiodinium microadriaticum*. *The ISME Journal*.
- Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* 6, 677–681.
- Chen, L., Chamberlain, A.J., Reich, C.M., Daetwyler, H.D., and Hayes, B.J. (2017b). Detection and validation of structural variations in bovine whole-genome sequence data. *Genetics Selection Evolution* 49.
-

Chen, Y., Li, D., Lu, W., Xing, J., Hui, B., and Han, Y. (2003). Screening and characterization of astaxanthin-hyperproducing mutants of *Haematococcus pluvialis*. *Biotechnology Letters* 25, 527–529.

Chénais, B. (2015). Transposable elements in cancer and other human diseases. *Current Cancer Drug Targets* 15, 227–242.

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

Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* 10, 563–569.

Chu, C., Nielsen, R., and Wu, Y. (2016). REPdenovo: Inferring De Novo Repeat Motifs from Short Sequence Reads. *PLOS ONE* 11, e0150719.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676.

Conte, M.H., Sicre, M.-A., Rühlemann, C., Weber, J.C., Schulte, S., Schulz-Bull, D., and Blanz, T. (2006). Global temperature calibration of the alkenone unsaturation index ($U^{K'_{37}}$) in surface waters and comparison with surface sediments: ALKENONE UNSATURATION INDEX. *Geochemistry, Geophysics, Geosystems* 7, n/a-n/a.

Cordero, B.F., Obratsova, I., Couso, I., Leon, R., Vargas, M.A., and Rodriguez, H. (2011). Enhancement of Lutein Production in *Chlorella sorokiniana* (Chlorophyta) by Improvement of Culture Conditions and Random Mutagenesis. *Marine Drugs* 9, 1607–1624.

Corliss, J.O. (1975). Three centuries of protozoology: a brief tribute to its founding father, A. van Leeuwenhoek of Delft. *The Journal of Protozoology* 22, 3–7.

Corner, E.D.S., and Davies, A.G. (1971). Plankton as a factor in the nitrogen and phosphorus cycles in the sea. In *Advances in Marine Biology*, (Elsevier), pp. 101–204.

Cowan, R.K., Hoen, D.R., Schoen, D.J., and Bureau, T.E. (2005). MUSTANG Is a Novel Family of Domesticated Transposase Genes Found in Diverse Angiosperms. *Molecular Biology and Evolution* 22, 2084–2089.

Curtis, B.A., Tanifuji, G., Burki, F., Gruber, A., Irimia, M., Maruyama, S., Arias, M.C., Ball, S.G., Gile, G.H., Hirakawa, Y., et al. (2012). Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* 492, 59–65.

Custódio, L., Soares, F., Pereira, H., Barreira, L., Vizetto-Duarte, C., Rodrigues, M.J., Rauter, A.P., Alberício, F., and Varela, J. (2014). Fatty acid composition and biological activities of *Isochrysis galbana* T-ISO, *Tetraselmis* sp. and *Scenedesmus* sp.: possible application in the pharmaceutical and functional food industries. *Journal of Applied Phycology* 26, 151–161.

Cuvelier, M.L., Allen, A.E., Monier, A., McCrow, J.P., Messie, M., Tringe, S.G., Woyke, T., Welsh, R.M., Ishoey, T., Lee, J.-H., et al. (2010). Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proceedings of the National Academy of Sciences* 107, 14679–14684.

Daccord, N., Celton, J.-M., Linsmith, G., Becker, C., Choisine, N., Schijlen, E., van de Geest, H., Bianco, L., Micheletti, D., Velasco, R., et al. (2017). High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nature Genetics* 49, 1099–1106.

Daniels, S.B., Peterson, K.R., Strausbaugh, L.D., Kidwell, M.G., and Chovnick, A. (1990). Evidence for horizontal transmission of the P transposable element between *Drosophila* species. *Genetics* 124, 339–355.

Darboux, I., Charles, J.-F., Pauchet, Y., Warot, S., and Pauron, D. (2007). Transposon-mediated resistance to *Bacillus sphaericus* in a field-evolved population of *Culex pipiens* (Diptera: Culicidae). *Cellular Microbiology* 9, 2022–2029.

von Dassow, P., Ogata, H., Probert, I., Wincker, P., Da Silva, C., Audic, S., Claverie, J.-M., and de Vargas, C. (2009). Transcriptome analysis of functional differentiation between haploid and diploid cells of *Emiliania huxleyi*, a globally significant photosynthetic calcifying cell. *Genome Biology* 10, R114.

Davis, S.J. (2016). Using MinION nanopore sequencing to generate a de novo eukaryotic draft genome: preliminary physiological and genomic description of the extremophilic red alga *Galdieria sulphuraria* strain SAG 107.79.

Daxinger, L., and Whitelaw, E. (2012). Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nature Reviews Genetics* 13, 153.

Day, A., and Rochaix, J.-D. (1991). Structure and inheritance of sense and anti-sense transcripts from a transposon in the green alga *Chlamydomonas reinhardtii*. *Journal of Molecular Biology* 218, 273–291.

Day, A., Schirmer-Rahire, M., Kuchka, M.R., Mayfield, S.P., and Rochaix, J.-D. (1988). A transposon with an unusual arrangement of long terminal repeats in the green alga *Chlamydomonas reinhardtii*. *The EMBO Journal* 7, 1917.

Decelle, J., Probert, I., Bittner, L., Desdevises, Y., Colin, S., de Vargas, C., Galí, M., Simó, R., and Not, F. (2012). An original mode of symbiosis in open ocean plankton. *Proceedings of the National Academy of Sciences* 109, 18000–18005.

Delbrut, A., Albina, P., Lapierre, T., Pradelles, R., and Dubreucq, E. (2018). Fucoxanthin and Polyunsaturated Fatty Acids Co-Extraction by a Green Process. *Molecules* 23.

Delwiche, C.F. (1999). Tracing the Thread of Plastid Diversity through the Tapestry of Life. *Am Nat* 154.

Derelle, E., Ferraz, C., Rombauts, S., Rouzé, P., Worden, A.Z., Robbens, S., Partensky, F., Degroeve, S., Echeynié, S., Cooke, R., et al. (2006). Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proceedings of the National Academy of Sciences* 103, 11647–11652.

Derelle, E., Yau, S., Moreau, H., and Grimsley, N.H. (2018). Prasinovirus Attack of *Ostreococcus* Is Furtive by Day but Savage by Night. *Journal of Virology* 92, 16.

Disdero, E., and Filée, J. (2017). LoRTE: Detecting transposon-induced genomic variants using low coverage PacBio long read sequences. *Mobile DNA* 8.

Doan, T.T.Y., and Obbard, J.P. (2012). Enhanced intracellular lipid in *Nannochloropsis* sp. via random mutagenesis and flow cytometric cell sorting. *Algal Research* 1, 17–21.

-
- Doolittle, W.F., and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284, 601.
- Doré, C., and Varoquaux, F. (2006). Histoire et amélioration de cinquante plantes cultivées.
- Douglas, S., Zauner, S., Fraunholz, M., Beaton, M., Penny, S., Deng, L.-T., Wu, X., Reith, M., Cavalier-Smith, T., and Maier, U.-G. (2001). The highly reduced genome of an enslaved algal nucleus. *Nature* 410, 1091.
- Driver, T., Bajhaiya, A., and Pittman, J.K. (2014). Potential of Bioenergy Production from Microalgae. *Current Sustainable/Renewable Energy Reports* 1, 94–103.
- Duarte, C.M., and Cebrian, J. (1996). The fate of marine autotrophic production. *Limnology and Oceanography* 41, 1758–1766.
- Eddy, S.R., and others (1995). Multiple alignment using hidden Markov models. *Ismb* 3, 114–120.
- Edgar, R.C., and Myers, E.W. (2005). PILER: identification and classification of genomic repeats. *Bioinformatics* 21, i152–i158.
- Edwardsen, B., Egge, E.S., and Vaulot, D. (2016). Diversity and distribution of haptophytes revealed by environmental sequencing and metabarcoding – a review. *Perspectives in Phycology* 3, 77–91.
- Efimova, K.V., Orlova, T.Y., and Brykov, V.A. (2016). Molecular genetic identification of a new *Tisochrysis lutea* (Bendif et Probert, 2013) strain isolated from the Russian coastal waters of the Sea of Japan. *Microbiology* 85, 325–332.
- Egue, F., Chenais, B., Tastard, E., Marchand, J., Hiard, S., Gateau, H., Hermann, D., Morant-Manceau, A., Casse, N., and Caruso, A. (2015). Expression of the retrotransposons *Surcouf* and *Blackbeard* in the marine diatom *Phaeodactylum tricornutum* under thermal stress. *Phycologia* 54, 617–627.
- Eikrem, W., Medlin, L.K., Henderiks, J., Rokitta, S., Rost, B., Probert, I., Throndsen, J., and Edwardsen, B. (2016). Haptophyta. In *Handbook of the Protists*, J.M. Archibald, A.G.B. Simpson, C.H. Slamovits, L. Margulis, M. Melkonian, D.J. Chapman, and J.O. Corliss, eds. (Cham: Springer International Publishing), pp. 1–61.
- Elbarbary, R.A., Lucas, B.A., and Maquat, L.E. (2016). Retrotransposons as regulators of gene expression. *Science* 351, aac7247–aac7247.
- Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 9, 18.
- Endo, H., Hanawa, Y., Araie, H., Suzuki, I., and Shiraiwa, Y. (2018). Overexpression of *Tisochrysis lutea* Akd1 identifies a key cold-induced alkenone desaturase enzyme. *Scientific Reports* 8.
- Engels, W.R., Johnson-Schlitz, D.M., Eggleson, W.B., and Sved, J. (1990). High-frequency P element loss in *Drosophila* is homolog dependent. *Cell* 62, 515–525.
- Ewing, A.D. (2015). Transposable element detection from whole genome sequence data. *Mobile DNA* 6.
- Falkowski, P.G. (2002). The ocean’s invisible forest. *Scientific American* 287, 54–61.

Fan, W.-H., Woelfle, M.A., and Mosig, G. (1995). Two copies of a DNA element, 'Wendy', in the chloroplast chromosome of *Chlamydomonas reinhardtii* between rearranged gene clusters. *Plant Molecular Biology* 29, 63–80.

Feder, M.E., and Hofmann, G.E. (1999). HEAT-SHOCK PROTEINS, MOLECULAR CHAPERONES, AND THE STRESS RESPONSE: Evolutionary and Ecological Physiology. *Annual Review of Physiology* 61, 243–282.

Ferris, P.J. (1989). Characterization of a *Chlamydomonas* transposon, Gulliver, resembling those in higher plants. *Genetics* 122, 363–377.

Ferris, P.J., Woessner, J.P., and Goodenough, U.W. (1996). A sex recognition glycoprotein is encoded by the plus mating-type gene *fus1* of *Chlamydomonas reinhardtii*. *Molecular Biology of the Cell* 7, 1235–1248.

Feschotte, C. (2005). DNA-binding specificity of rice mariner-like transposases and interactions with Stowaway MITEs. *Nucleic Acids Research* 33, 2153–2165.

Feschotte, C., and Mouches, C. (2000). Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a pogo-like DNA transposon. *Molecular Biology and Evolution* 17, 730–737.

Feschotte, C., and Pritham, E.J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annual Review of Genetics* 41, 331.

Feschotte, C., Jiang, N., and Wessler, S.R. (2002). Plant transposable elements: where genetics meets genomics. *Nature Reviews Genetics* 3, 329–341.

Field, C.B., Behrenfeld, M.J., Randerson, J.T., and Falkowski, P. (1998). Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* 281, 237–240.

Finlay, B.J. (2001). Exploring Leeuwenhoek's legacy: the abundance and diversity of protozoa. 9.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39.

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

Flowers, J.M., Hazzouri, K.M., Pham, G.M., Rosas, U., Bahmani, T., Khraiweh, B., Nelson, D.R., Jijakli, K., Abdrabu, R., Harris, E.H., et al. (2015). Whole-Genome Resequencing Reveals Extensive Natural Variation in the Model Green Alga *Chlamydomonas reinhardtii*. *The Plant Cell* 27, 2353–2369.

Flutre, T., Permal, E., and Quesneville, H. (2012). Transposable Element Annotation in Completely Sequenced Eukaryote Genomes. In *Plant Transposable Elements*, M.-A. Grandbastien, and J.M. Casacuberta, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 17–39.

Frances, T., Casse, N., Badawi, M., Carrier, G., and Berthelie, J. (2018). Automatisation d'un pipeline d'analyse d'éléments transposables dans les génomes séquencés. *Mémoire de Master* 2.

Gallaher, S.D., Fitz-Gibbon, S.T., Glaesener, A.G., Pellegrini, M., and Merchant, S.S. (2015). *Chlamydomonas* Genome Resource for Laboratory Strains Reveals a Mosaic of Sequence

Variation, Identifies True Strain Histories, and Enables Strain-Specific Studies. *The Plant Cell* 27, 2335–2352.

García, J.L., de Vicente, M., and Galán, B. (2017). Microalgae, old sustainable food and fashion nutraceuticals. *Microbial Biotechnology* 10, 1017–1024.

Garnier, M., Carrier, G., Rogniaux, H., Nicolau, E., Bougaran, G., Saint-Jean, B., and Cadoret, J.P. (2014). Comparative proteomics reveals proteins impacted by nitrogen deprivation in wild-type and high lipid-accumulating mutant strains of *Tisochrysis lutea*. *Journal of Proteomics* 105, 107–120.

Gast, R.J., McDonnell, T.A., and Caron, D.A. (2000). srDna-based taxonomic affinities of algal symbionts from a planktonic foraminifer and a solitary radiolarian. *Journal of Phycology* 36, 172–177.

Geider, R.J., Delucia, E.H., Falkowski, P.G., Finzi, A.C., Grime, J.P., Grace, J., Kana, T.M., La Roche, J., Long, S.P., and Osborne, B.A. (2001). Primary productivity of planet earth: biological determinants and physical constraints in terrestrial and aquatic habitats. *Global Change Biology* 7, 849–882.

George, J.A., Traverse, K.L., DeBaryshe, P.G., Kelley, K.J., and Pardue, M.-L. (2010). Evolution of diverse mechanisms for protecting chromosome ends by *Drosophila* TART telomere retrotransposons. *Proc Natl Acad Sci USA* 107, 21052.

Gilbert, C., and Feschotte, C. (2018). Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. *Current Opinion in Genetics & Development* 49, 15–24.

Godet, S., Hérault, J., Pencreac'h, G., Ergon, F., and Loiseau, C. (2012). Isolation and analysis of a gene from the marine microalga *Isochrysis galbana* that encodes a lipase-like protein. *Journal of Applied Phycology* 24, 1547–1553.

Godos, I. de, Blanco, S., García-Encina, P.A., Becares, E., and Muñoz, R. (2009). Long-term operation of high rate algal ponds for the bioremediation of piggery wastewaters at high loading rates. *Bioresource Technology* 100, 4332–4339.

Goerner-Potvin, P., and Bourque, G. (2018). Computational tools to unmask transposable elements. *Nature Reviews Genetics*.

Goubert, C., Modolo, L., Vieira, C., ValienteMoro, C., Mavingui, P., and Boulesteix, M. (2015). De Novo Assembly and Annotation of the Asian Tiger Mosquito (*Aedes albopictus*) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (*Aedes aegypti*). *Genome Biology and Evolution* 7, 1192–1205.

Gould, S.B., Waller, R.F., and McFadden, G.I. (2008). Plastid Evolution. *Annual Review of Plant Biology* 59, 491–517.

Graham, I.E., Spanier, J.G., and Jarvik, J.W. (1995). Isolation and characterization of Pioneer1, a novel *Chlamydomonas* transposable element. *Current Genetics* 28, 429–436.

Graham, L.A., Loughheed, S.C., Ewart, K.V., and Davies, P.L. (2008). Lateral Transfer of a Lectin-Like Antifreeze Protein Gene in Fishes. *PLoS ONE* 3, 11.

Gray, A., Lucas, I., Seed, R., and Richardson, C. (1999). *Mytilus edulis chilensis* infested with *Coccomyxa parasitica* (Chlorococcales, Coccomyxaceae). *Journal of Molluscan Studies* 65, 289–294.

-
- Green, J. t, and Parke, M. (1975). New observations upon members of the genus *Chrysotila* Anand, with remarks upon their relationships within the Haptophyceae. *Journal of the Marine Biological Association of the United Kingdom* 55, 109–121.
- Grima, E.M., Sánchez Pérez, J.A.S., Medina, A.R., and E M (1995). The Production of Polyunsaturated Fatty Acids by Microalgae: from Strain Selection to Product Purification. 9.
- Grönke, S., Mildner, A., Fellert, S., Tennagels, N., Petry, S., Müller, G., Jäckle, H., and Kühnlein, R.P. (2005). Brummer lipase is an evolutionary conserved fat storage regulator in *Drosophila*. *Cell Metabolism* 1, 323–330.
- Guedes, A., and Malcata, F. (2012). Nutritional Value and Uses of Microalgae in Aquaculture. In *Aquaculture*, Z. Muchlisin, ed. (InTech), p.
- Guiry, M.D. (2012). HOW MANY SPECIES OF ALGAE ARE THERE? *Journal of Phycology* 48, 1057–1063.
- Guo, Y., and Levin, H.L. (2010). High-throughput sequencing of retrotransposon integration provides a saturated profile of target activity in *Schizosaccharomyces pombe*. *Genome Research* 20, 239–248.
- Guo, B., Jiang, M.L., Wan, X., Gong, Y.M., Liang, Z., and Hu, C.J. (2013). Identification and Heterologous Expression of a $\Delta 4$ -Fatty Acid Desaturase Gene from *Isochrysis sphaerica*. *Journal of Microbiology and Biotechnology* 23, 1413–1421.
- Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8, 1494.
- Han, Y., and Wessler, S.R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research* 38, e199–e199.
- Hayward, A. (2017). Origin of the retroviruses: when, where, and how? *Current Opinion in Virology* 25, 23–27.
- Helliwell, K.E., Collins, S., Kazamia, E., Purton, S., Wheeler, G.L., and Smith, A.G. (2015). Fundamental shift in vitamin B12 eco-physiology of a model alga demonstrated by experimental evolution. *The ISME Journal* 9, 1446–1455.
- Hemaiswarya, S., Raja, R., Ravi Kumar, R., Ganesan, V., and Anbazhagan, C. (2011). Microalgae: a sustainable feed source for aquaculture. *World Journal of Microbiology and Biotechnology* 27, 1737–1746.
- Hermann, D. (2011). Caractérisation d'éléments transposables de type mariner chez les microalgues marines. Université du Maine.
- Hernández Javier, L., Benzekri, H., Gut, M., Claros, M.G., van Bergeijk, S., Cañavate, J.P., and Manchado, M. (2018). Characterization of Iodine-Related Molecular Processes in the Marine Microalga *Tisochrysis lutea* (Haptophyta). *Frontiers in Marine Science* 5.
- Higashiyama, T., Maki, S., and Yamada, T. (1995a). Molecular organization of *Chlorella vulgaris* chromosome I: presence of telomeric repeats that are conserved in higher plants. *Molecular and General Genetics MGG* 246, 29–36.

-
- Higashiyama, T., Noutoshi, Y., Akiba, M., and Yamada, T. (1995b). Telomere and LINE-like elements at the termini of the *Chlorella* chromosome I. pp. 71–72.
- Higashiyama, T., Noutoshi, Y., Fujie, M., and Yamada, T. (1997). Zepp, a LINE-like retrotransposon accumulated in the *Chlorella* telomeric region. *The EMBO Journal* 16, 3715–3723.
- Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., and Quesneville, H. (2014). PASTEC: An Automatic Transposable Element Classification Tool. *PLoS ONE* 9, e91929.
- Hoen, D.R., Hickey, G., Bourque, G., Casacuberta, J., Cordaux, R., Feschotte, C., Fiston-Lavier, A.-S., Hua-Van, A., Hubley, R., and Kapusta, A. (2015). A call for benchmarking transposable element annotation methods. *Mobile DNA* 6, 13.
- Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12, 491.
- Houdan, A., Billard, C., Marie, D., Not, F., Sáez, A.G., Young, J.R., and Probert, I. (2004). Holococcolithophore-heterococcolithophore (Haptophyta) life cycles: Flow cytometric analysis of relative ploidy levels. *Systematics and Biodiversity* 1, 453–465.
- Hovde, B.T., Deodato, C.R., Hunsperger, H.M., Ryken, S.A., Yost, W., Jha, R.K., Patterson, J., Monnat, R.J., Barlow, S.B., Starkenburg, S.R., et al. (2015). Genome Sequence and Transcriptome Analyses of *Chrysochromulina tobin*: Metabolic Tools for Enhanced Algal Fitness in the Prominent Order Prymnesiales (Haptophyceae). *PLOS Genetics* 11, e1005469.
- Huddleston, J., Chaisson, M.J., Steinberg, K.M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T.A., Munson, K.M., Kronenberg, Z.N., Vives, L., et al. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research* 27, 677–685.
- Huerlimann, R., Steinig, E.J., Loxton, H., Zenger, K.R., Jerry, D.R., and Heimann, K. (2014). Effects of growth phase and nitrogen starvation on expression of fatty acid desaturases and fatty acid composition of *Isochrysis aff. galbana* (TISO). *Gene* 545, 36–44.
- Hull, R. (2001). Classifying reverse transcribing elements: a proposal and a challenge to the ICTV. *Archives of Virology* 146, 2255–2261.
- Iglesias-Rodríguez, M.D., Brown, C.W., Doney, S.C., Kleypas, J., Kolber, D., Kolber, Z., Hayes, P.K., and Falkowski, P.G. (2002). Representing key phytoplankton functional groups in ocean carbon cycle models: Coccolithophorids: COCCOLITHOPHORIDS IN OCEAN CARBON CYCLE MODELS. *Global Biogeochemical Cycles* 16, 47-1-47–20.
- Iskandarov, U., Khozin-Goldberg, I., and Cohen, Z. (2010). Selection of a DGLA-producing mutant of the microalga *Parietochloris incisa*: I. Identification of mutation site and expression of VLC-PUFA biosynthesis genes. *Appl Microbiol Biotechnol* 90, 249–256.
- Islam, M.A., Heimann, K., and Brown, R.J. (2017). Microalgae biodiesel: Current status and future needs for engine performance and emissions. *Renewable and Sustainable Energy Reviews* 79, 1160–1170.
- Jaekisch, N., Yang, I., Wohlrab, S., Glöckner, G., Kroymann, J., Vogel, H., Cembella, A., and John, U. (2011). Comparative Genomic and Transcriptomic Characterization of the Toxigenic Marine Dinoflagellate *Alexandrium ostenfeldii*. *PLoS ONE* 6, e28012.
-

-
- Jaeger, L. de, Verbeek, R.E., Draaisma, R.B., Martens, D.E., Springer, J., Eggink, G., and Wijffels, R.H. (2014). Superior triacylglycerol (TAG) accumulation in starchless mutants of *Scenedesmus obliquus*: (I) mutant generation and characterization. *Biotechnology for Biofuels* 7, 69.
- Jardillier, L., Zubkov, M.V., Pearman, J., and Scanlan, D.J. (2010). Significant CO₂ fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. *The ISME Journal* 4, 1180–1192.
- Jardim, S.S., Schuch, A.P., Pereira, C.M., and Loreto, E.L.S. (2015). Effects of heat and UV radiation on the mobilization of transposon mariner-Mos1. *Cell Stress and Chaperones* 20, 843–851.
- Jiang, N., Bao, Z., Zhang, X., Eddy, S.R., and Wessler, S.R. (2004). Pack-MULE transposable elements mediate gene evolution in plants. *Nature* 431, 569.
- Jordan, R.W., Cros, L., and Young, J.R. (2004). A revised classification scheme for living haptophytes. *Micropaleontology* 55–79.
- Kajihara, D., Godoy, F. de, Hamaji, T.A., Blanco, S.R., Van Sluys, M.-A., and Rossi, M. (2012). Functional characterization of sugarcane mustang domesticated transposases and comparative diversity in sugarcane, rice, maize and sorghum. *Genetics and Molecular Biology* 35, 632–639.
- Kalendar, R., Vicient, C.M., Peleg, O., Ananthawat-Jonsson, K., Bolshoy, A., and Schulman, A.H. (2004). Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics* 166, 1437–1450.
- Kapitonov, V.V., and Jurka, J. (2007). Helitrons on a roll: eukaryotic rolling-circle transposons. *TRENDS in Genetics* 23, 521–529.
- Kapitonov, V.V., and Jurka, J. (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews Genetics* 9, 411–412.
- Kapusta, A., Suh, A., and Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences of the United States of America* 114, E1460–E1469.
- Kawachi, M., Inouye, I., Maeda, O., and Chihara, M. (1991). The haptonema as a food-capturing device: observations on *Chrysochromulina hirta* (Prymnesiophyceae). *Phycologia* 30, 563–573.
- Kazazian, H.H. (2004). Mobile Elements: Drivers of Genome Evolution. *Science* 303, 1626–1632.
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649.
- Keeling, P.J. (2010). The endosymbiotic origin, diversification and fate of plastids. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365, 729–748.
- Keeling, P.J. (2013). The Number, Speed, and Impact of Plastid Endosymbioses in Eukaryotic Evolution. *Annual Review of Plant Biology* 64, 583–607.
-

-
- Keeling, P.J., and Palmer, J.D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics* 9, 605–618.
- Khan, H., Kozera, C., Curtis, B.A., Bussey, J.T., Theophilou, S., Bowman, S., and Archibald, J.M. (2007). Retrotransposons and Tandem Repeat Sequences in the Nuclear Genomes of Cryptomonad Algae. *Journal of Molecular Evolution* 64, 223–236.
- Kidwell, M.G. (1979). Hybrid dysgenesis in *Drosophila melanogaster*: the relationship between the P–M and I–R interaction systems. *Genetical Research* 33, 205.
- Kidwell, M.G., and Lisch, D.R. (2001). Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution* 55, 1–24.
- Kim, K.-S. (2006). Natural History of Transposition in the Green Alga *Chlamydomonas reinhardtii*: Use of the AMT4 Locus as an Experimental System. *Genetics* 173, 2005–2019.
- Kirk, M.M., Stark, K., Miller, S.M., Muller, W., Taillon, B.E., Gruber, H., Schmitt, R., and Kirk, D.L. (1999). *regA*, a *Volvox* gene that plays a central role in germ-soma differentiation, encodes a novel regulatory protein. *Development* 126, 639–647.
- Kirkham, A.R., Jardillier, L.E., Holland, R., Zubkov, M.V., and Scanlan, D.J. (2011). Analysis of photosynthetic picoeukaryote community structure along an extended Ellett Line transect in the northern North Atlantic reveals a dominance of novel prymnesiophyte and prasinophyte phylotypes. *Deep Sea Research Part I: Oceanographic Research Papers* 58, 733–744.
- Kobayashi, S., Goto-Yamamoto, N., and Hirochika, H. (2004). Retrotransposon-induced mutations in grape skin color. *Science* 304, 982–982.
- Koch, P., Platzer, M., and Downie, B.R. (2014). RepARK--de novo creation of repeat libraries from whole-genome NGS reads. *Nucleic Acids Research* 42, e80–e80.
- Koester, J.A., Berthiaume, C.T., Hiranuma, N., Parker, M.S., Iverson, V., Morales, R., Ruzzo, W.L., and Armbrust, E.V. (2018). Sexual ancestors generated an obligate asexual and globally dispersed clone within the model diatom species *Thalassiosira pseudonana*. *Scientific Reports* 8.
- Koga, A., Inagaki, H., Bessho, Y., and Hori, H. (1995). Insertion of a novel transposable element in the tyrosinase gene is responsible for an albino mutation in the medaka fish, *Oryzias latipes*. *MGG Molecular & General Genetics* 249.
- Kohany, O., Gentles, A.J., Hankus, L., and Jurka, J. (2006). Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 7.
- Kojima, K.K. (2005). An extraordinary retrotransposon family encoding dual endonucleases. *Genome Research* 15, 1106–1117.
- Kolpakov, R. (2003). mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Research* 31, 3672–3678.
- Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *BioRxiv* 071282.
- Köster, J., and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 28, 2520–2522.
-

-
- Krasovec, M., Eyre-Walker, A., Sanchez-Ferandin, S., and Piganeau, G. (2017). Spontaneous Mutation Rate in the Smallest Photosynthetic Eukaryotes. *Molecular Biology and Evolution* 34, 1770–1779.
- Krasovec, M., Sanchez-Brosseau, S., Grimsley, N., and Piganeau, G. (2018). Spontaneous mutation rate as a source of diversity for improving desirable traits in cultured microalgae. *Algal Research* 35, 85–90.
- Kumekawa, N., Ohtsubo, E., and Ohtsubo, H. (1999). Identification and phylogenetic analysis of gypsy-type retrotransposons in the plant kingdom. *Genes & Genetic Systems* 74, 299–307.
- Kützing, F.T. (1833). *Algarum aquae dulcis germanicarum* (In commissis CA Schwertschkii).
- Lacour, T., Sciandra, A., Talec, A., Mayzaud, P., and Bernard, O. (2012). DIEL VARIATIONS OF CARBOHYDRATES AND NEUTRAL LIPIDS IN NITROGEN-SUFFICIENT AND NITROGEN-STARVED CYCLOSTAT CULTURES OF ISOCHRYYSIS SP.1: C RESERVES OF ISOCHRYYSIS SP. UNDER L:D ILLUMINATION. *Journal of Phycology* 48, 966–975.
- LaJeunesse, T.C., Lambert, G., Andersen, R.A., Coffroth, M.A., and Galbraith, D.W. (2005). SYMBIODINIUM (PYRRHOPHYTA) GENOME SIZES (DNA CONTENT) ARE SMALLEST AMONG DINOFLAGELLATES1. *Journal of Phycology* 41, 880–886.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
- Lang-Unnasch, N., Reith, M.E., Munholland, J., and Barta, J.R. (1998). Plastids are widespread and ancient in parasites of the phylum Apicomplexafn1. *International Journal for Parasitology* 28, 1743–1754.
- Larkum, A.W.D., Ross, I.L., Kruse, O., and Hankamer, B. (2012). Selection, breeding and engineering of microalgae for bioenergy and biofuel production. *Trends in Biotechnology* 30, 198–205.
- Larson, G., Piperno, D.R., Allaby, R.G., Purugganan, M.D., Andersson, L., Arroyo-Kalin, M., Barton, L., Climer Vigueira, C., Denham, T., Dobney, K., et al. (2014). Current perspectives and the future of domestication studies. *Proc Natl Acad Sci USA* 111, 6139.
- Lefebvre, P.A., and Silflow, C.D. (1999). Chlamydomonas: the cell and its genomes. *Genetics* 151, 9–14.
- Lerat, E., and Capy, P. (1999). Retrotransposons and retroviruses: analysis of the envelope gene. *Molecular Biology and Evolution* 16, 1198–1207.
- Lerat, E., Rizzon, C., and Biémont, C. (2003). Sequence Divergence Within Transposable Element Families in the Drosophila melanogaster Genome. *Genome Research*.
- Lerat, E., Burlet, N., Biémont, C., and Vieira, C. (2011). Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. *Gene* 473, 100–109.
- Lescot, M., Hingamp, P., Kojima, K.K., Villar, E., Romac, S., Veluchamy, A., Boccara, M., Jaillon, O., Iudicone, D., Bowler, C., et al. (2016). Reverse transcriptase genes are highly abundant and transcriptionally active in marine plankton assemblages. *The ISME Journal* 10, 1134–1146.
-

-
- Levin, H.L., and Moran, J.V. (2011). Dynamic interactions between transposable elements and their hosts. *Nature Reviews Genetics* 12, 615.
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.
- Li, W., Jaroszewski, L., and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 17, 282–283.
- Lian, M., Huang, H., Ren, L., Ji, X., Zhu, J., and Jin, L. (2010). Increase of Docosahexaenoic Acid Production by *Schizochytrium* sp. Through Mutagenesis and Enzyme Assay. *Applied Biochemistry and Biotechnology* 162, 935–941.
- Liao, G., Rehm, E.J., and Rubin, G.M. (2000). Insertion site preferences of the P transposable element in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences* 97, 3347–3351.
- Lin, S. (2011). Genomic understanding of dinoflagellates. *Research in Microbiology* 162, 551–569.
- Lin, S., Cheng, S., Song, B., Zhong, X., Lin, X., Li, W., Li, L., Zhang, Y., Zhang, H., and Ji, Z. (2015). The *Symbiodinium kawagutii* genome illuminates dinoflagellate gene expression and coral symbiosis. *Science* 350, 691–694.
- Lipman, D., and Pearson, W. (1985). Rapid and sensitive protein similarity searches. *Science* 227, 1435.
- Lisch, D. (2012). How important are transposons for plant evolution? *Nature Reviews Genetics* 14, 49–61.
- Liu, H., Probert, I., Uitz, J., Claustre, H., Aris-Brosou, S., Frada, M., Not, F., and de Vargas, C. (2009a). Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans. *Proceedings of the National Academy of Sciences* 106, 12803–12808.
- Liu, J., Sommerfeld, M., and Hu, Q. (2013). Screening and characterization of *Isochrysis* strains and optimization of culture conditions for docosahexaenoic acid production. *Applied Microbiology and Biotechnology* 97, 4785–4798.
- Liu, S., Yeh, C.-T., Ji, T., Ying, K., Wu, H., Tang, H.M., Fu, Y., Nettleton, D., and Schnable, P.S. (2009b). Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS Genetics* 5, e1000733.
- Lommer, M., Specht, M., Roy, A.-S., Kraemer, L., Andreson, R., Gutowska, M.A., Wolf, J., Bergner, S.V., Schilhabel, M.B., Klostermeier, U.C., et al. (2012). Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome Biology* 13, R66.
- Lu, L., Chen, J., Robb, S.M.C., Okumoto, Y., Stajich, J.E., and Wessler, S.R. (2017). Tracking the genome-wide outcomes of a transposable element burst over decades of amplification. *Proceedings of the National Academy of Sciences* 114, E10550–E10559.
- Ma, Y., Wang, Z., Zhu, M., Yu, C., Cao, Y., Zhang, D., and Zhou, G. (2013). Increased lipid productivity and TAG content in *Nannochloropsis* by heavy-ion irradiation mutagenesis. *Bioresource Technology* 136, 360–367.
-

-
- Mager, D.L., and Stoye, J.P. (2015). Mammalian endogenous retroviruses. In *Mobile DNA III*, (American Society of Microbiology), pp. 1079–1100.
- Marchant, H.J., and Thomsen, H. (1994). Haptophytes in polar waters. *SYSTEMATICS ASSOCIATION SPECIAL VOLUME 51*, 209–209.
- Marchetti, J., Bougaran, G., Le Dean, L., Mégrier, C., Lukomska, E., Kaas, R., Olivo, E., Baron, R., Robert, R., and Cadoret, J.P. (2012). Optimizing conditions for the continuous culture of *Isochrysis affinis galbana* relevant to commercial hatcheries. *Aquaculture* 326–329, 106–115.
- Marchetti, J., Bougaran, G., Jauffrais, T., Lefebvre, S., Rouxel, C., Saint-Jean, B., Lukomska, E., Robert, R., and Cadoret, J.P. (2013). Effects of blue light on the biochemical composition and photosynthetic activity of *Isochrysis* sp. (T-iso). *Journal of Applied Phycology* 25, 109–119.
- Marchetti, J., da Costa, F., Bougaran, G., Quéré, C., Soudant, P., and Robert, R. (2018). The combined effects of blue light and dilution rate on lipid class and fatty acid composition of *Tisochrysis lutea*. *Journal of Applied Phycology* 30, 1483–1494.
- Margulis, L. (1975). Symbiotic theory of the origin of eukaryotic organelles; criteria for proof. pp. 21–38.
- Marlowe, I.T., Green, J.C., Neal, A.C., Brassell, S.C., Eglinton, G., and Course, P.A. (1984). Long chain (n -C₃₇ –C₃₉) alkenones in the Prymnesiophyceae. Distribution of alkenones and other lipids and their taxonomic significance. *British Phycological Journal* 19, 203–216.
- Martínez, M., Sánchez, S., Jiménez, J., El Yousfi, F., and Muñoz, L. (2000). Nitrogen and phosphorus removal from urban wastewater by the microalga *Scenedesmus obliquus*. *Bioresource Technology* 73, 263–272.
- Matsuzaki, M., Misumi, O., Shin-i, T., Maruyama, S., Takahara, M., Miyagishima, S., Mori, T., Nishida, K., Yagisawa, F., Nishida, K., et al. (2004). Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. 428, 5.
- Maurus, F., Allen, A.E., Mhiri, C., Hu, H., Jabbari, K., Vardi, A., Grandbastien, M.-A., and Bowler, C. (2009). Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC Genomics* 10, 624.
- Maxam, A.M., and Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences* 74, 560–564.
- McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences* 36, 334–355.
- McClintock, B. (1984). The significance of responses of the genome to challenge. *Science* 226, 792–801.
- McCoy, R.C., Taylor, R.W., Blauwkamp, T.A., Kelley, J.L., Kertesz, M., Pushkarev, D., Petrov, D.A., and Fiston-Lavier, A.-S. (2014). Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements. *PLoS ONE* 9, e106689.
- Medhora, M., Maruyama, K., and Hartl, D. (1991). Molecular and functional analysis of the mariner mutator element *Mos1* in *Drosophila*. *Genetics* 128, 311–318.
-

-
- Meireles, L.A., Guedes, A.C., and Malcata, F.X. (2003). Increase of the yields of eicosapentaenoic and docosahexaenoic acids by the microalga *Pavlova lutheri* following random mutagenesis. *Biotechnol. Bioeng.* *81*, 50–55.
- Mendel, G. (1865). Versuche uber pflanzen-hybriden. Vorgelegt in Den Sitzungen.
- Mendoza, H., de la Jara, A., Freijanes, K., Carmona, L., Ramos, A.A., de Sousa Duarte, V., and Serafim Varela, J.C. (2008). Characterization of *Dunaliella salina* strains by flow cytometry: a new approach to select carotenoid hyperproducing strains. *Electronic Journal of Biotechnology* *11*, 0–0.
- Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Marechal-Drouard, L., et al. (2007). The *Chlamydomonas* Genome Reveals the Evolution of Key Animal and Plant Functions. *Science* *318*, 245–250.
- Mereschkowsky, C. (1910). Theorie der zwei Plasmaarten als Grundlage der Symbiogenesis, einer neuen Lehre von der Entstehung der Organismen. *Biologisches Centralblatt* *30*, 278–288.
- Meyer, R.S., and Purugganan, M.D. (2013). Evolution of crop species: genetics of domestication and diversification. *Nature Reviews Genetics* *14*, 840–852.
- Mijnendonckx, K., Provoost, A., Monsieurs, P., Leys, N., Mergeay, M., Mahillon, J., and Van Houdt, R. (2011). Insertion sequence elements in *Cupriavidus metallidurans* CH34: Distribution and role in adaptation. *Plasmid* *65*, 193–203.
- Milano, J., Ong, H.C., Masjuki, H.H., Chong, W.T., Lam, M.K., Loh, P.K., and Vellayan, V. (2016). Microalgae biofuels as an alternative to fossil fuel for power generation. *Renewable and Sustainable Energy Reviews* *58*, 180–197.
- Milledge, J.J. (2011). Commercial application of microalgae other than as biofuels: a brief review. *Reviews in Environmental Science and Bio/Technology* *10*, 31–41.
- Miller, S.M., and Kirk, D.L. (1999). *glsA*, a *Volvox* gene required for asymmetric division and germ cell specification, encodes a chaperone-like protein. *Development* *126*, 649–658.
- Miller, S.M., Schmitt, R., and Kirk, D.L. (1993). *Jordan*, an active *Volvox* transposable element similar to higher plant transposons. *The Plant Cell* *5*, 1125–1138.
- Mills, R.E., Bennett, E.A., Iskow, R.C., and Devine, S.E. (2007). Which transposable elements are active in the human genome? *Trends in Genetics* *23*, 183–191.
- Mimouni, V., Ulmann, L., Haimeur, A., Guéno, F., Meskini, N., and Tremblin, G. (2015). Marine microalgae used as food supplements and their implication in preventing cardiovascular diseases. *OCL* *22*, D409.
- Minei, R., Hoshina, R., and Ogura, A. (2018). De novo assembly of middle-sized genome using MinION and Illumina sequencers. *BMC Genomics* *19*.
- Monat, C., Tranchant-Dubreuil, C., Kougbéadjó, A., Farcy, C., Ortega-Abboud, E., Amanzougarene, S., Ravel, S., Agbessi, M., Orjuela-Bouniol, J., Summo, M., et al. (2015). TOGGLE: toolbox for generic NGS analyses. *BMC Bioinformatics* *16*, 374.
- Moon-van der Staay, S.Y., De Wachter, R., and Vaultot, D. (2001). Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* *409*, 607.
- Muller-Feuga, A. (2000). The role of microalgae in aquaculture: situation and trends. *22*.
-

-
- Nabout, J.C., da Silva Rocha, B., Carneiro, F.M., and Sant'Anna, C.L. (2013). How many species of Cyanobacteria are there? Using a discovery curve to predict the species number. *Biodiversity and Conservation* 22, 2907–2918.
- Nagarajan, D., Lee, D.-J., Kondo, A., and Chang, J.-S. (2017). Recent insights into biohydrogen production by microalgae – From biophotolysis to dark fermentation. *Bioresource Technology* 227, 373–387.
- Ness, R.W., Morgan, A.D., Colegrave, N., and Keightley, P.D. (2012). Estimate of the Spontaneous Mutation Rate in *Chlamydomonas reinhardtii*. *Genetics* 192, 1447–1454.
- Ness, R.W., Morgan, A.D., Vasanthakrishnan, R.B., Colegrave, N., and Keightley, P.D. (2015). Extensive de novo mutation rate variation between individuals and across the genome of *Chlamydomonas reinhardtii*. *Genome Research* 25, 1739–1749.
- Nguyen, D.H. (2014). Caractérisation et expression de nouveaux éléments génétiques transposables de la superfamille Tcl-Mariner chez la microalgue marine *Amphora acutiuscula* (Bacillariophyta).
- Nguyen, H.M., Cuine, S., Beyly-Adriano, A., Legeret, B., Billon, E., Auroy, P., Beisson, F., Peltier, G., and Li-Beisson, Y. (2013). The Green Microalga *Chlamydomonas reinhardtii* Has a Single -3 Fatty Acid Desaturase That Localizes to the Chloroplast and Impacts Both Plastidic and Extraplastidic Membrane Lipids. *PLANT PHYSIOLOGY* 163, 914–928.
- Nishii, I., Ogihara, S., and Kirk, D.L. (2003). A kinesin, InvA, plays an essential role in *Volvox* morphogenesis. *Cell* 113, 743–753.
- Nissimov, J.I., Worthy, C.A., Rooks, P., Napier, J.A., Kimmance, S.A., Henn, M.R., Ogata, H., and Allen, M.J. (2012a). Draft Genome Sequence of Four Coccolithoviruses: *Emiliana huxleyi* Virus EhV-88, EhV-201, EhV-207, and EhV-208. *Journal of Virology* 86, 2896–2897.
- Nissimov, J.I., Worthy, C.A., Rooks, P., Napier, J.A., Kimmance, S.A., Henn, M.R., Ogata, H., and Allen, M.J. (2012b). Draft Genome Sequence of the Coccolithovirus *Emiliana huxleyi* Virus 202. *Journal of Virology* 86, 2380–2381.
- Nomaguchi, T., Maeda, Y., Liang, Y., Yoshino, T., Asahi, T., and Tanaka, T. (2018). Comprehensive analysis of triacylglycerol lipases in the oleaginous diatom *Fistulifera solaris* JPCc DA0580 with transcriptomics under lipid degradation. *Journal of Bioscience and Bioengineering* 126, 258–265.
- Norden-Krichmar, T.M., Allen, A.E., Gaasterland, T., and Hildebrand, M. (2011). Characterization of the Small RNA Transcriptome of the Diatom, *Thalassiosira pseudonana*. *PLoS ONE* 6, e22870.
- Not, F., Siano, R., Kooistra, W.H., Simon, N., Vaultot, D., and Probert, I. (2012). Diversity and ecology of eukaryotic marine phytoplankton. *Advances in Botanical Research* 64, 1–53.
- Noutoshi, Y., Arai, R., Fujie, M., and Yamada, T. (1998). Structure of the *Chlorella* Zepp retrotransposon: nested Zepp clusters in the genome. *Molecular and General Genetics MGG* 259, 256–263.
- Novák, P., Neumann, P., and Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* 11, 378.

Novak, P., Neumann, P., Pech, J., Steinhaisl, J., and Macas, J. (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* 29, 792–793.

Oliver, M.J., Schofield, O., and Bidle, K. (2010). Density dependent expression of a diatom retrotransposon. *Marine Genomics* 3, 145–150.

Ota, S., Matsuda, T., Takeshita, T., Yamazaki, T., Kazama, Y., Abe, T., and Kawano, S. (2013a). Phenotypic spectrum of *Parachlorella kessleri* (Chlorophyta) mutants produced by heavy-ion irradiation. *Bioresource Technology* 149, 432–438.

Ota, S., Matsuda, T., Takeshita, T., Yamazaki, T., Kazama, Y., Abe, T., and Kawano, S. (2013b). Phenotypic spectrum of *Parachlorella kessleri* (Chlorophyta) mutants produced by heavy-ion irradiation. *Bioresource Technology* 149, 432–438.

Palenik, B., Grimwood, J., Aerts, A., Rouzé, P., Salamov, A., Putnam, N., Dupont, C., Jorgensen, R., Derelle, E., Rombauts, S., et al. (2007). The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proceedings of the National Academy of Sciences* 104, 7705–7710.

Partensky, F., Hess, W.R., and Vaulot, D. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiology and Molecular Biology Reviews* 63, 106–127.

Pawlowski, J., Christen, R., Lecroq, B., Bachar, D., Shahbazkia, H.R., Amaral-Zettler, L., and Guillou, L. (2011). Eukaryotic Richness in the Abyss: Insights from Pyrotag Sequencing. *PLoS ONE* 6, e18169.

Pereira, S.L., Huang, Y.-S., Bobik, E.G., Kinney, A.J., Stecca, K.L., Packer, J.C.L., and Mukerji, P. (2004). A novel ω 3-fatty acid desaturase involved in the biosynthesis of eicosapentaenoic acid. 7.

Perez-Alegre, M., Dubus, A., and Fernandez, E. (2005). REM1, a New Type of Long Terminal Repeat Retrotransposon in *Chlamydomonas reinhardtii*. *Molecular and Cellular Biology* 25, 10628–10638.

Perrineau, M.-M., Gross, J., Zelzion, E., Price, D.C., Levitan, O., Boyd, J., and Bhattacharya, D. (2014). Using Natural Selection to Explore the Adaptive Potential of *Chlamydomonas reinhardtii*. *PLoS ONE* 9, e92533.

Petrov, D.A., Schutzman, J.L., Hartl, D.L., and Lozovskaya, E.R. (1995). Diverse transposable elements are mobilized in hybrid dysgenesis in *Drosophila virilis*. *Proc Natl Acad Sci USA* 92, 8050.

Phillippy, A.M. (2017). New advances in sequence assembly. *Genome Research* 27, xi–xiii.

Picard, G., Bregliano, J.C., Bucheton, A., Lavige, J.M., Pelisson, A., and Kidwell, M.G. (1978). Non-mendelian female sterility and hybrid dysgenesis in *Drosophila melanogaster*. *Genetical Research* 32, 275.

Poole, A.M., and Neumann, N. (2011). Reconciling an archaeal origin of eukaryotes with engulfment: a biologically plausible update of the Eocyte hypothesis. *Research in Microbiology* 162, 71–76.

Poptsova, M.S., Il'icheva, I.A., Nechipurenko, D.Y., Panchenko, L.A., Khodikov, M.V., Oparina, N.Y., Polozov, R.V., Nechipurenko, Y.D., and Grokhovsky, S.L. (2015). Non-random DNA fragmentation in next-generation sequencing. *Scientific Reports* 4.

Price, D.C., Chan, C.X., Yoon, H.S., Yang, E.C., Qiu, H., Weber, A.P.M., Schwacke, R., Gross, J., Blouin, N.A., Lane, C., et al. (2012). *Cyanophora paradoxa* Genome Elucidates Origin of Photosynthesis in Algae and Plants. *Science* 335, 843–847.

Quesneville, H., Nouaud, D., and Anxolabéhère, D. (2003). Detection of New Transposable Element Families in *Drosophila melanogaster* and *Anopheles gambiae* Genomes. *Journal of Molecular Evolution* 57, S50–S59.

Quesneville, H., Bergman, C.M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., and Anxolabehere, D. (2005). Combined Evidence Annotation of Transposable Elements in Genome Sequences. *PLoS Computational Biology* 1, e22.

Rajaby, R., and Sung, W.-K. (2018). TranSurVeyor: an improved database-free algorithm for finding non-reference transpositions in high-throughput sequencing data. 11.

Ramallo, E., Kalendar, R., Schulman, A.H., and Martínez-Izquierdo, J.A. (2008). Reme1, a Copia retrotransposon in melon, is transcriptionally induced by UV light. *Plant Molecular Biology* 66, 137–150.

Rasul, I., Azeem, F., Siddique, M.H., Muzammil, S., Rasul, A., Munawar, A., Afzal, M., Ali, M.A., and Nadeem, H. (2017). Chapter 8 - Algae Biotechnology: A Green Light for Engineered Algae. In *Algae Based Polymers, Blends, and Composites*, K.M. Zia, M. Zuber, and M. Ali, eds. (Elsevier), pp. 301–334.

Read, B.A., Kegel, J., Klute, M.J., Kuo, A., Lefebvre, S.C., Maumus, F., Mayer, C., Miller, J., Monier, A., Salamov, A., et al. (2013). Pan genome of the phytoplankton *Emiliana underpins* its global distribution. *Nature* 499, 209–213.

Rechka, J.A., and Maxwell, J.R. (1988). Characterisation of alkenone temperature indicators in sediments and organisms. *Organic Geochemistry* 13, 727–734.

Reinar, W.B. (2016). In silico explorations of TE activity, diversity and abundance across 74 teleost fish genomes.

Renuka, N., Sood, A., Prasanna, R., and Ahluwalia, A.S. (2015). Phycoremediation of wastewaters: a synergistic approach using microalgae for bioremediation and biomass generation. *International Journal of Environmental Science and Technology* 12, 1443–1460.

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

Rolfe, M., Spanos, A., and Banks, G. (1986). Induction of yeast Ty element transcription by ultraviolet light. *Nature* 319, 339.

Rondelet (1554). De piscibus marinis, libri XVIII, in quibus veræ piscium effigies expressæ sunt.

Rondelet (1555). Universæ aquatiliū historiæ pars altera, cum veris ipsorum imaginibus.

Ros, F., and Kunze, R. (2001). Regulation of Activator/Dissociation Transposition by Replication and DNA Methylation. 11.

-
- Rott, E., Cantonati, M., Füreder, L., and Pfister, P. (2006). Benthic Algae in High Altitude Streams of the Alps – a Neglected Component of the Aquatic Biota. *Hydrobiologia* 562, 195–216.
- Rouxel, C., Bougaran, G., Doulin-Grouas, S., Dubois, N., and Cadoret, J. (2011). Novel *Isochrysis* sp tahitian clone and uses therefore.
- Rumin, J., Bonnefond, H., Saint-Jean, B., Rouxel, C., Sciandra, A., Bernard, O., Cadoret, J.-P., and Bougaran, G. (2015). The use of fluorescent Nile red and BODIPY for lipid measurement in microalgae. *Biotechnology for Biofuels* 8, 42.
- Ryckebosch, E., Bruneel, C., Termote-Verhalle, R., Goiris, K., Muylaert, K., and Foubert, I. (2014). Nutritional evaluation of microalgae oils rich in omega-3 long chain polyunsaturated fatty acids as an alternative for fish oil. *Food Chemistry* 160, 393–400.
- Sabot, F., and Schulman, A.H. (2006). Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. *Heredity* 97, 381–388.
- Sabot, F., Sourdille, P., Chantret, N., and Bernard, M. (2006). Morgane, a new LTR retrotransposon group, and its subfamilies in wheats. *Genetica* 128, 439–447.
- Saint-Vanne, J., Berthelie, J., and Carrier, G. (2017). Identification et reconstruction des haplotypes d'une population de microalgues à partir d'un séquençage 3ème génération. Mémoire de Master 2.
- Sakai, N., Sakamoto, Y., Kishimoto, N., Chihara, M., and Karube, I. (1995). *Chlorella* strains from hot springs tolerant to high temperature and high CO₂. *Energy Conversion and Management* 36, 693–696.
- Sanger, F., and Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* 94, 441–448.
- Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison III, C.A., Slocombe, P.M., and Smith, M. (1977). Nucleotide sequence of bacteriophage ϕ X174 DNA. *Nature* 265, 687.
- Saoudi-Helis, L., Dubacq, J.-P., Marty, Y., Samain, J.-F., and Gudin, C. (1994). Influence of growth rate on pigment and lipid composition of the microalga *Isochrysis* aff. *galbana* clone T. iso. *Journal of Applied Phycology* 6, 315–322.
- Schaack, S., Gilbert, C., and Feschotte, C. (2010). Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in Ecology & Evolution* 25, 537–546.
- Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al. (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science* 326, 1112.
- Schulman, A.H. (2015). Genome Size and the Role of Transposable Elements. In *Genetics and Genomics of Brachypodium*, (Springer), pp. 81–106.
- Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., and Schatz, M.C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods* 15, 461–468.
-

-
- Shaish, A., Ben-Amotz, A., and Avron, M. (1991). PRODUCTION AND SELECTION OF HIGH β -CAROTENE MUTANTS OF DUNALIELLA BARDAWIL (CHLOROPHYTA)1. *Journal of Phycology* 27, 652–656.
- Shapiro, J.A. (1969). Mutations caused by the insertion of genetic material into the galactose operon of *Escherichia coli*. *Journal of Molecular Biology* 40, 93–105.
- Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., and Waterston, R.H. (2017). DNA sequencing at 40: past, present and future. *Nature* 550, 345–353.
- Shi, T., Yu, A., Li, M., Ou, X., Xing, L., and Li, M. (2012). Identification of a novel C22- Δ 4-producing docosahexaenoic acid (DHA) specific polyunsaturated fatty acid desaturase gene from *Isochrysis galbana* and its expression in *Saccharomyces cerevisiae*. *Biotechnology Letters* 34, 2265–2274.
- Shoguchi, E., Shinzato, C., Kawashima, T., Gyoja, F., Mungpakdee, S., Koyanagi, R., Takeuchi, T., Hisata, K., Tanaka, M., Fujiwara, M., et al. (2013). Draft Assembly of the *Symbiodinium minutum* Nuclear Genome Reveals Dinoflagellate Gene Structure. *Current Biology* 23, 1399–1408.
- Siguié, P., Gournay, E., and Chandler, M. (2014). Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiology Reviews* 38, 865–891.
- Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics* 8, 272–285.
- Smit, A.F., and Hubley, R. (2010). RepeatModeler Open-1.0.
- Smit, A.F., Hubley, R., and Green, P. (1996). RepeatMasker.
- Söding, J. (2005). Protein homology detection by HMM–HMM comparison. *Bioinformatics* 21, 951–960.
- Song, B., Morse, D., Song, Y., Fu, Y., Lin, X., Wang, W., Cheng, S., Chen, W., Liu, X., and Lin, S. (2017). Comparative Genomics Reveals Two Major Bouts of Gene Retroposition Coinciding with Crucial Periods of *Symbiodinium* Evolution. *Genome Biology and Evolution* 9, 2037–2047.
- Song, B., Chen, S., and Chen, W. (2018). Dinoflagellates, a Unique Lineage for Retrogene Research. *Frontiers in Microbiology* 9.
- Spicer, A., and Molnar, A. (2018). Gene Editing of Microalgae: Scientific Progress and Regulatory Challenges in Europe. *Biology* 7, 21.
- Steinbiss, S., Willhoeft, U., Gremme, G., and Kurtz, S. (2009). Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Research* 37, 7002–7013.
- Sung, W., Ackerman, M.S., Miller, S.F., Doak, T.G., and Lynch, M. (2012). Drift-barrier hypothesis and mutation-rate evolution. *Proceedings of the National Academy of Sciences* 109, 18488–18492.
- T. H. Morgan (1910). Chromosomes and Heredity. *The American Naturalist* 44, 449–496.
- Talbert, L.E., and Chandler, V.L. (1988). Characterization of a highly conserved sequence related to mutator transposable elements in maize. *Molecular Biology and Evolution* 5, 519–529.

Tan, S., Cardoso-Moreira, M., Shi, W., Zhang, D., Huang, J., Mao, Y., Jia, H., Zhang, Y., Chen, C., Shao, Y., et al. (2016). LTR-mediated retroposition as a mechanism of RNA-based duplication in metazoans. *Genome Research* 26, 1663–1675.

Tattini, L., D’Aurizio, R., and Magi, A. (2015). Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Frontiers in Bioengineering and Biotechnology* 3.

Terashima, M., Freeman, E.S., Jinkerson, R.E., and Jonikas, M.C. (2015). A fluorescence-activated cell sorting-based strategy for rapid isolation of high-lipid *Chlamydomonas* mutants. *Plant J* 81, 147–159.

Tesson, S.V., Skjøth, C.A., Šantl-Temkiv, T., and Löndahl, J. (2016). Airborne microalgae: insights, opportunities, and challenges. *Applied and Environmental Microbiology* 82, 1978–1991.

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

The Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69–87.

Thiriet-Rupert, S., Carrier, G., Chénais, B., Trottier, C., Bougaran, G., Cadoret, J.-P., Schoefs, B., and Saint-Jean, B. (2016). Transcription factors in microalgae: genome-wide prediction and comparative analysis. *BMC Genomics* 17.

Thiriet-Rupert, S., Carrier, G., Trottier, C., Eveillard, D., Schoefs, B., Bougaran, G., Cadoret, J.-P., Chénais, B., and Saint-Jean, B. (2018). Identification of transcription factors involved in the phenotype of a domesticated oleaginous microalgae strain of *Tisochrysis lutea*. *Algal Research* 30, 59–72.

Thung, D.T., Slagboom, E.P., Ye, K., Veltman, J.A., and Hehir-Kwa, J.Y. (2014). Mobster: accurate detection of mobile element insertions in next generation sequencing data. 11.

Tillmann, U. Phagotrophy by a plastidic haptophyte, *Prymnesium patelliferum*. 6.

Treangen, T.J., and Salzberg, S.L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* 13, nrg3117.

Tréguer, P., and Pondaven, P. (2000). Silica control of carbon dioxide. 2.

Trentacoste, E.M., Shrestha, R.P., Smith, S.R., Gle, C., Hartmann, A.C., Hildebrand, M., and Gerwick, W.H. (2013). Metabolic engineering of lipid catabolism increases microalgal lipid accumulation without compromising growth. *Proceedings of the National Academy of Sciences* 110, 19748–19753.

Treves, H., Raanan, H., Finkel, O.M., Berkowicz, S.M., Keren, N., Shotland, Y., and Kaplan, A. (2013). A newly isolated *Chlorella* sp. from desert sand crusts exhibits a unique resistance to excess light intensity. *FEMS Microbiology Ecology* 86, 373–380.

Tsuji, Y., and Yoshida, M. (2017). Biology of Haptophytes: Complicated Cellular Processes Driving the Global Carbon Cycle. In *Advances in Botanical Research*, (Elsevier), pp. 219–261.

Turcotte, K., Srinivasan, S., and Bureau, T. (2008). Survey of transposable elements from rice genomic sequences. *The Plant Journal* 25, 169–179.

-
- Tyson, J.R., O'Neil, N.J., Jain, M., Olsen, H.E., Hieter, P., and Snutch, T.P. (2018). MinION-based long-read sequencing and assembly extends the *Caenorhabditis elegans* reference genome. *Genome Research* 28, 266–274.
- Ueki, N., and Nishii, I. (2008). Idaten Is a New Cold-Inducible Transposon of *Volvox carteri* That Can Be Used for Tagging Developmentally Important Genes. *Genetics* 180, 1343–1353.
- Uribe, P., Fuentes, D., Valdés, J., Shmaryahu, A., Zúñiga, A., Holmes, D., and Valenzuela, P.D.T. (2008). Preparation and Analysis of an Expressed Sequence Tag Library from the Toxic Dinoflagellate *Alexandrium catenella*. *Marine Biotechnology* 10, 692–700.
- Van Dolah, F.M., Lidie, K.B., Monroe, E.A., Bhattacharya, D., Campbell, L., Doucette, G.J., and Kamykowski, D. (2009). The Florida red tide dinoflagellate *Karenia brevis*: New insights into cellular and molecular processes underlying bloom dynamics. *Harmful Algae* 8, 562–572.
- Van Wagenen, J., Holdt, S.L., De Francisci, D., Valverde-Pérez, B., Plósz, B.G., and Angelidaki, I. (2014). Microplate-based method for high-throughput screening of microalgae growth potential. *Bioresource Technology* 169, 566–572.
- de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., et al. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science* 348, 1261605–1261605.
- Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The Sequence of the Human Genome. *THE HUMAN GENOME* 291, 51.
- Vieira, C., Fablet, M., Lerat, E., Boulesteix, M., Rebollo, R., Burlet, N., Akkouche, A., Hubert, B., Mortada, H., and Biémont, C. (2012). A comparative analysis of the amounts and dynamics of transposable elements in natural populations of *Drosophila melanogaster* and *Drosophila simulans*. *Journal of Environmental Radioactivity* 113, 83–86.
- Vigani, M., Parisi, C., Rodríguez-Cerezo, E., Barbosa, M.J., Sijtsma, L., Ploeg, M., and Enzing, C. (2015). Food and feed products from micro-algae: Market opportunities and challenges for the EU. *Trends in Food Science & Technology* 42, 81–92.
- Vigeolas, H., Duby, F., Kaymak, E., Niessen, G., Motte, P., Franck, F., and Remacle, C. (2012). Isolation and partial characterization of mutants with elevated lipid content in *Chlorella sorokiniana* and *Scenedesmus obliquus*. *Journal of Biotechnology* 162, 3–12.
- Visser, M.E. (2008). Keeping up with a warming world; assessing the rate of adaptation to climate change. *Proceedings of the Royal Society B: Biological Sciences* 275, 649–659.
- Volk, T., and Hoffert, M.I. (1985). Ocean carbon pumps: Analysis of relative strengths and efficiencies in ocean-driven atmospheric CO₂ changes. *The Carbon Cycle and Atmospheric CO₂: Natural Variations Archean to Present* 32, 99–110.
- Walbot, V. (1992). Reactivation of Mutator transposable elements of maize by ultraviolet light. *MGG Molecular & General Genetics* 234, 353–360.
- Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE* 9, e112963.

-
- Wallace, M.R., Andersen, L.B., Saulino, A.M., Gregory, P.E., Glover, T.W., and Collins, F.S. (1991). A de novo Alu insertion results in neurofibromatosis type 1. *Nature* 353, 864.
- Wang, S.-C., Schnell, R.A., and Lefebvre, P.A. (1998). Isolation and characterization of a new transposable element in *Chlamydomonas reinhardtii*. *Plant Molecular Biology* 38, 681–687.
- Warren, W.D., Atkinson, P.W., and O’Brochta, D.A. (1994). The Hermes transposable element from the house fly, *Musca domestica*, is a short inverted repeat-type element of the hobo, Ac, and Tam3 (hAT) element family. *Genetics Research* 64, 87–97.
- Waterbury, J.B. (1986). Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. *Photosynthetic Picoplankton* 71–120.
- Wenke, T., Dobel, T., Sorensen, T.R., Junghans, H., Weisshaar, B., and Schmidt, T. (2011). Targeted Identification of Short Interspersed Nuclear Element Families Shows Their Widespread Existence and Extreme Heterogeneity in Plant Genomes. *THE PLANT CELL ONLINE* 23, 3117–3128.
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* 8, 973–982.
- Witte, C.-P., Le, Q.H., Bureau, T., and Kumar, A. (2001). Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proceedings of the National Academy of Sciences* 98, 13778–13783.
- Worden, A.Z., Lee, J.-H., Mock, T., Rouzé, P., Simmons, M.P., Aerts, A.L., Allen, A.E., Cuvelier, M.L., Derelle, E., Everett, M.V., et al. (2009). Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes *Micromonas*. *Science* 324, 268.
- Xiao, H., Jiang, N., Schaffner, E., Stockinger, E.J., and van der Knaap, E. (2008). A Retrotransposon-Mediated Gene Duplication Underlies Morphological Variation of Tomato Fruit. *Science* 319, 1527–1530.
- Xue, J., Niu, Y.-F., Huang, T., Yang, W.-D., Liu, J.-S., and Li, H.-Y. (2015). Genetic improvement of the microalga *Phaeodactylum tricornutum* for boosting neutral lipid accumulation. *Metabolic Engineering* 27, 1–9.
- Yamamoto, Y. (2003). Retrotransposon-mediated restoration of *Chlorella* telomeres: accumulation of Zepp retrotransposons at termini of newly formed minichromosomes. *Nucleic Acids Research* 31, 4646–4653.
- Yang, L., and Bennetzen, J.L. (2009). Structure-based discovery and description of plant and animal Helitrons. *Proceedings of the National Academy of Sciences* 106, 12832–12837.
- Yang, J., Pan, Y., Bowler, C., Zhang, L., and Hu, H. (2016). Knockdown of phosphoenolpyruvate carboxykinase increases carbon flux to lipid synthesis in *Phaeodactylum tricornutum*. *Algal Research* 15, 50–58.
- Yang, Y.N., Liu, C.L., Wang, Y.K., and Xue, J.M. (2013). Mutation effects of C2+ ion irradiation on the greasy *Nitzschia* sp. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 751–752, 24–28.
-

-
- Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25, 2865–2871.
- Yoon, B.-J. (2009). Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current Genomics* 10, 402–415.
- Yousuff, C., Ho, E., Hussain K., I., and Hamid, N. (2017). Microfluidic Platform for Cell Isolation and Manipulation Based on Cell Properties. *Micromachines* 8, 15.
- Yu, C., Zhang, J., and Peterson, T. (2011). Genome rearrangements in maize induced by alternative transposition of reversed Ac/Ds termini. *Genetics* genetics-111.
- Yu, Z., Wright, S.I., and Bureau, T.E. (2000). Mutator-like Elements in *Arabidopsis thaliana*: Structure, Diversity and Evolution. 14.
- Zayadan, B.K., Purton, S., Sadvakasova, A.K., Userbaeva, A.A., and Bolatkhan, K. (2013). Isolation, mutagenesis, and optimization of cultivation conditions of microalgal strains for biodiesel production. *Russ J Plant Physiol* 61, 124–130.
- Zayadan, B.K., Purton, S., Sadvakasova, A.K., Userbaeva, A.A., and Bolatkhan, K. (2014). Isolation, mutagenesis, and optimization of cultivation conditions of microalgal strains for biodiesel production. *Russian Journal of Plant Physiology* 61, 124–130.
- Zeller, M.A., Hunt, R., Jones, A., and Sharma, S. (2013). Bioplastics and their thermoplastic blends from *Spirulina* and *Chlorella* microalgae. *Journal of Applied Polymer Science* 130, 3263–3275.
- Zerbino, D.R. (2010). Using the Velvet *de novo* Assembler for Short-Read Sequencing Technologies. In *Current Protocols in Bioinformatics*, A.D. Baxevanis, D.B. Davison, R.D.M. Page, G.A. Petsko, L.D. Stein, and G.D. Stormo, eds. (Hoboken, NJ, USA: John Wiley & Sons, Inc.), p.
- Zhang, X., Zhang, X.-F., Li, H.-P., Wang, L.-Y., Zhang, C., Xing, X.-H., and Bao, C.-Y. (2014). Atmospheric and room temperature plasma (ARTP) as a new powerful mutagenesis tool. *Applied Microbiology and Biotechnology* 98, 5387–5396.
- Zimin, A.V., Puiu, D., Luo, M.-C., Zhu, T., Koren, S., Marçais, G., Yorke, J.A., Dvořák, J., and Salzberg, S.L. (2017). Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Research* 27, 787–792.
- Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A.N., and Alva, V. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology* 430, 2237–2243.

Annexes

Annexe 1

Liste des assemblages de génomes de microalgues qui ont été recensés.

Espèce	Souche	NCBI	Chromosomes ou Contigs	Taille (Mpb)	Publication(s)	Technologie(s)
Chlorophytes						
<i>Auxenochlorella protothecoides</i>	710	GCA_000733215.1	Contigs	22,9	Gao et al. 2014	454 GS FLX Titanium et Illumina HiSeq
<i>Auxenochlorella protothecoides</i>	UTEX 2341	GCA_002154875.1	Contigs	32,7	na	Illumina HiSeq
<i>Auxenochlorella pyrenoidosa</i>	FACHB-9	GCA_001430745.1	Contigs	57	Fan et al. 2015	454
<i>Bathycoccus prasinos</i>	RCC 1105	GCA_002220235.1	Chromosomes	15,1	Moreau et al. 2012	Sanger
<i>Bathycoccus</i> sp.	TOSAG39-1	GCA_900128745.1	Contigs	10,1	na	na
<i>Botryococcus braunii</i>	Showa	GCA_002005505.1	Contigs	184,4	na	PacBio
<i>Chlamydomonas applanata</i>	NIES-2202	GCA_001662365.1	Contigs	78,5	na	Illunima Miseq
<i>Chlamydomonas asymmetrica</i>	NIES-2207	GCA_001662385.1	Contigs	141,9	na	Illunima Miseq
<i>Chlamydomonas debaryana</i>	NIES-2212	GCA_001662405.1	Contigs	120,4	na	Illunima Miseq
<i>Chlamydomonas eustigma</i>	NIES-2499	GCA_002335675.1	Contigs	66,6	Hirooka et al. 2017	454 GS FLX+ et Illumina
<i>Chlamydomonas reinhardtii</i>	CC-503 cw92 mt+	GCA_000002595.3	Chromosomes	120,4	Merchant et al. 2007	ABI 3739
<i>Chlamydomonas sphaeroides</i>	NIES-2242	GCA_001662425.1	Contigs	122,2	na	
<i>Chlorella sorokiniana</i>	1228	GCA_002939045.1	Contigs	61,4	na	PacBio
"	1412	GCA_003116155.1	Contigs	57,9	na	PacBio
"	1230	GCA_003130725.1	Contigs	58,5	na	PacBio
"	1602	GCA_002245835.2	Contigs	59,6	Arriola et al. 2018	PacBio; Illumina HiSeq
<i>Chlorella</i> sp.	A99	GCA_003063905.1	Contigs	40,9	/	Illumina MiSeq; Illumina HiSeq
<i>Chlorella</i> sp.	ArM0029B	GCA_002896455.3	Contigs	93	Ahn et al. 2012	Illumina HiSeq
<i>Chlorella variabilis</i>	NC64A	GCA_000147415.1	Contigs	46,2	Blanc et al. 2010	Sanger
<i>Chlorella vulgaris</i>	UTEX 395	GCA_001021125.1	Contigs	37,3	na	Illumina HiSeq
<i>Chromochloris zofingiensis</i>	SAG 211– 14	na	Chromosomes	58	Roth et al. 2017	PacBio; Illumina HiSeq
<i>Coccomyxa</i> sp.	LA000219	GCA_000812005.1	Contigs	48,5	na	Illumina HiSeq
<i>Coccomyxa</i> sp.	SUA001	GCA_001244535.1	Contigs	11,8	na	IonTorrent
<i>Coccomyxa subellipsoidea</i>	C-169	GCA_000258705.1	Contigs	48,8	Blanc et al. 2012	Sanger
<i>Coelastrella</i> sp.	M60	GCA_001630525.1	Contigs	80,2	na	Illumina MiSeq
<i>Coelastrella</i> sp.	UTEX B 3026	GCA_002588565.1	Contigs	151,5	na	Illumina HiSeq
<i>Cymbomonas tetramitiformis</i>	PLY_AMNH	GCA_001247695.1	Contigs	281,3	na	Illumina HiSeq
<i>Dunaliella salina</i>	CCAP 19/18	GCA_002284615.1	Contigs	343,7	Polle et al. 2018	PacBio; Illumina HiSeq
<i>Eudorina</i> sp.	2010-623- F1-E4	GCA_003117195.1	Contigs	184	na	PacBio; Illumina HiSeq
"	2010-623- F1-E2	GCA_003117095.1	Contigs	168,6	na	PacBio; Illumina HiSeq

Espèce	Souche	NCBI	Chromosomes ou Contigs	Taille (Mpb)	Publications	Technologies
Chlorophytes (suite)						
<i>Gonium pectorale</i>	NIES-2863	GCA_001584585.1	Contigs	148,8	Hanschen, et al. 2016	454, Illumina MiSeq, ABI3730
<i>Helicosporidium</i> sp.	ATCC 50920	GCA_000690575.1	Contigs	12,4	na	Illumina HiSeq, Illumina GAIIx
<i>Micractinium conductrix</i>	SAG 241.80	GCA_002245815.2	Contigs	61	Arriola et al. 2018	PacBio
<i>Micromonas commoda</i>	RCC299	GCA_000090985.2	Chromosomes	21,1	Worden AZ, et al. 2009	na
<i>Micromonas pusilla</i>	CCMP1545	GCA_000151265.1	Contigs	22	Worden AZ, et al. 2010	na
<i>Micromonas</i> sp.	ASP10-01a	GCA_001430725.1	Contigs	19,6	Delmont et al. 2015	Illumina HiSeq
<i>Monoraphidium neglectum</i>	SAG 48.87	GCA_000611645.1	Contigs	69,7	Bogen C, et al. 2013	Illumina GAIIx
<i>Monoraphidium</i> sp.	549	GCA_002814315.1	Contigs	74,7	na	Illumina HiSeq
<i>Ostreococcus lucimarinus</i>	CCE9901	GCA_000092065.1	Chromosomes	13,2	Palenik et al. 2007	na
<i>Ostreococcus tauri</i>	RCC 4221	GCA_002158475.1	Chromosomes	14,8	Derelle et al. 2006; Blanc-Mathieu et al. 2014	Sanger
"	RCC 1115	GCA_000214015.2	Chromosomes	12,9	Blanc-Mathieu et al 2018	Illumina
<i>Ostreococcus</i> sp.	RCC809	na	Chromosome	13,3	na	
<i>Parachlorella kessleri</i>	NIES-2152	GCA_001598975.1	Contigs	59,2	Ota et al. 2016	454 GS FLX
<i>Picochlorum</i> sp.	SENEW3	GCA_000876415.1	Contigs	13,4	Foflonker et al. 2015	Illumina MiSeq
<i>Picochlorum</i> sp. 'soloecismus'	DOE 101	GCA_002818215.1	Contigs	15,3	Foflonker et al. 2015	Illumina, 454
<i>Prototheca cutis</i>	JCM 15793	GCA_002897115.1	Contigs	20	Suzuki et al. 2018	Illumina HiSeq
<i>Prototheca stagnorum</i>	JCM 9641	GCA_002794665.1	Contigs	16,9	Suzuki et al. 2018	Illumina HiSeq
<i>Prototheca wickerhamii</i>	HMC1	GCA_003255715.1	Contigs	27,7	na	Illumina
<i>Raphidocelis subcapitata</i>	NIES-35	GCA_003203535.1	Contigs	51,2	Suzuki et al. 2018	Illumina MiSeq
<i>Scenedesmus quadricauda</i>	LWG 002611	GCA_002317545.1	Contigs	65,4	na	Ion Proton
<i>Tetrabaena socialis</i>	NIES-571	GCA_002891735.1	Contigs	135,8	Featherston et al 2018	Illumina HiSeq, Illumina MiSeq
<i>Tetradismus obliquus</i>	na	GCA_900108755.1	Contigs	107,7	na	na
"	DOE0152z	GCA_002149895.1	Contigs	208,2	na	PacBio
<i>Trebouxia gelatinosa</i>	LA000220	GCA_000818905.1	Contigs	61,7	na	Illumina HiSeq
<i>Trebouxia</i> sp.	TZW2008	GCA_002118135.1	Contigs	69,3	na	Illumina HiSeq
uncultured <i>Bathycoccus</i>	uncultured	GCA_000259855.1	Contigs	5,2	Monier A, et al. 2012	Sanger, Roche 454 FLX
<i>Volvox carteri</i> f. <i>nagariensis</i>	Eve	GCA_000143455.1	Contigs	137,7	Prochnik, et al. 2010	Sanger
<i>Yamagishiella unicocca</i>	2012-1026-YU-F2-6	GCA_003116995.1	Contigs	134,2	na	PacBio, Illumina HiSeq
"	2012-1026-YU-F2-1	GCA_003117035.1	Contigs	140,8	na	PacBio, Illumina HiSeq

Espèce	Souche	NCBI	Chromosomes ou Contigs	Taille (Mpb)	Publication(s)	Technologie(s)
Stramenopiles						
<i>Asterionella formosa</i>	BG1	GCA_002256025.1	Contigs	68.4	na	Illumina NextSeq
<i>Aurantiochytrium</i> sp.	FCC1311	GCA_002897355.1	Contigs	38,9	na	Illumina HiSeq
<i>Aurantiochytrium</i> sp.	KH105	GCA_003116975.1	Contigs	76.82	Iwasaka et al. 2018	Illumina GAIIx, Illumina MiSeq, Roche 454 GS FLX
<i>Aurantiochytrium</i> sp.	T66	GCA_001462505.1	Contigs	43.43	Liu B, et al. 2016	454, Illumina HiSeq
<i>Aureococcus anophagefferens</i>	CCMP1984	GCA_000186865.1	Contigs	56,6	Gobler et al. 2011	Sanger
<i>Fistulifera solaris</i>	JPCC DA0580	GCA_002217885.1	Contigs	49,7	Tanaka et al. 2015	Roche 454 GS
<i>Fragilariopsis cylindrus</i>	CCMP1102	GCA_001750085.1	Contigs	80,5	Mock et al. 2017	Sanger, PacBio
<i>Heterococcus</i> sp.	DN1	GCA_000498555.1	Contigs	60.74	na	Illumina GAIIx
<i>Licmophora abbreviata</i>	CCMP470	GCA_900291995.1	Contigs	29,2	na	na
<i>Nannochloropsis gaditana</i>	CCMP527	GCA_001614215.1	Contigs	25,6	na	Illumina GAIIx
"	B-31	GCA_000569095.1	contigs	27,6	Carpinelli et al. 2014	Roche 454, SOLiD
"	CCMP1894	GCA_002838785.1	contigs	30,9	na	PacBio
"	CCMP526	GCA_000240725.1	Contigs	33,9	Radakovits et al. 2012	454 GS FLX, Illumina GA IIx
<i>Nannochloropsis limnetica</i>	CCMP505	GCA_001614225.1	Contigs	33,5	na	Illumina GA IIx
<i>Nannochloropsis oceanica</i>	LAMB0001	GCA_000226695.1	Contigs	27,6	na	Illumina GA IIx
"	OZ-1	GCA_001614235.1	Contigs	28	na	454 GS FLX, Illumina GA IIx
"	IMET1	GCA_001870945.1	Contigs	31,5	Hu et al. 2014	Roche 454
"	CCMP1779	na	Contigs	28.7	Vieler et al. 2012	Roche 454 et Illumina
<i>Nannochloropsis salina</i>	CCMP1776	GCA_001614245.1	Contigs	24,4	na	Illumina GA IIx
<i>Ochromonadaceae</i> sp.	CCMP2298	na	Contigs	61,1	na	Illumina
<i>Pelagophyceae</i> sp.	CCMP2097	na	Contigs	85,8	na	Illumina
<i>Phaeodactylum tricornutum</i>	CCAP 1055/1	GCA_000150955.2	Chromosomes	27,4	Bowler et al. 2008	na
<i>Pseudo-nitzschia multistriata</i>	B856	GCA_900005105.1	Contigs	55,1	na	na
<i>Pseudo-nitzschia multiseriata</i>	CLN-47	na	Contigs	218.73	na	Sanger et Roche 454
<i>Schizochytrium</i> sp.	CCTCC M209059	GCA_000818945.1	Contigs	38,3	Ji et al. 2015	Illumina HiSeq
<i>Thalassiosira oceanica</i>	CCMP1005	GCA_000296195.2	Contigs	92,2	Lommer et al. 2012	454 GS FLX Titanium
<i>Thalassiosira pseudonana</i>	CCMP1335	GCA_000149405.2	Chromosomes	32.4	Armbrust et al. 2004	na

Espèce	Souche	NCBI	Chromosomes ou Contigs	Taille (Mpb)	Publication(s)	Technologie(s)
Alvéolées						
<i>Chromera velia</i>	CCAP 1602/1	GCA_0005 85135.1	Contigs	193,6	Woo et al. 2015	Illumina HiSeq
<i>Prorocentrum minimum</i>	D-127	GCA_0016 52855.1	Contigs	473	Ponmani et al. 2016	454 GS FLX
<i>Symbiodinium kagautii</i> (Clade F)	CCMP2 468	na	Contigs	1180	Lin et al. 2015	Illumina HiSeq
<i>Symbiodinium minutum</i>	Mf 1.05b.0 1	GCA_0005 07305.1	Contigs	609,5	Shoguchi et al. 2013	454 GS FLX, Genome Analyzer Ix
<i>Symbiodinium microadriaticum</i>	CCMP2 467	GCA_0019 39145.1	Contigs	808,2	Aranda et al. 2016	Illumina HiSeq
<i>Symbiodinium</i> sp. (Clade A)	Y106	GCA_0032 97005.1	Contigs	766,7	Shoguchi et al. 2018	Illumina GAIIx, Illumina HiSeq
<i>Symbiodinium</i> sp. (Clade C)	Y103	GCA_0032 97045.1	Contigs	703,7	Shoguchi et al. 2018	Illumina GAIIx, Illumina HiSeq
<i>Vitrella brassicaformis</i>	CCMP3 155	GCA_0011 79505.1	Contigs	72,7	Woo et al. 2015	Illumina HiSeq
Haptophytes						
<i>Chrysochromulina tobin</i>	CCMP2 91	GCA_0012 75005.1	Contigs	59,1	Hovde et al. 2015	Illumina, 454
<i>Chrysochromulina parva</i>	Ohio	GCA_0028 87195.1	Contigs	65,7	na	Illumina NextSeq, Illumina MiSeq
<i>Emiliania huxleyi</i>	CCMP1 516	GCA_0003 72725.1	Contigs	155,9	Read et al. 2013	Sanger
<i>Pavlova</i> sp.	CCMP2 436	na	Contigs	165,4	na	Illumina
<i>Tisochrysis lutea</i>	CCAP 926/14	na	Contigs	82	Carrier et al. 2018; Bertheliet et al. 2018	Illumina HiSeq et PacBio
Rhodophyte						
<i>Cyanidioschyzon merolae</i>	10D	GCA_0000 91205.1	Chromosomes	16,5	Matsuzaki, et al. 2004; Nozaki, et al. 2007	na
<i>Galdieria sulphuraria</i>	074W	GCA_0003 41285.1	Contigs	13,7	Schönknecht et al. 2013	Sanger, 454
"	107.79	GCA_0017 04855.1	Contigs	12,1	Davis et al. 2016	Oxford N. MiNION
<i>Porphyridium purpureum</i>	NIES 2140	GCA_0003 97085.1	Contigs	19,5	Tajima et al. 2014	Illumina GAIIx
Chryptophytes						
<i>Cryptophyceae</i> sp.	CCMP2 293	na	Contigs	534,5	na	Illumina
<i>Guillardia theta</i>	CCMP2 712	GCA_0003 15625.1	Contigs	87,1	Curtis et al. 2012	Sanger, 454
Glaucophytes						
<i>Cyanophora paradoxa</i>	CCMP3 29	na	Contigs	70	Price et al. 2012	Roche 454, Illumina GAIIx, Sanger
Chlorarachniophytes						
<i>Bigeloviella natans</i>	CCMP2 755	GCA_0003 20545.1	Contigs	91,4	Curtis et al. Nature 2012	Sanger, 454

Annexe 2

Tableau récapitulatif des outils utilisés.

Outils	Méthode	Données d'entrées (E) / de sorties (S)	Difficulté d'installation	Avantages	Inconvénients	Avis
PARTIE 1: Annotation des ETs de <i>T. lutea</i>						
Repeat Explorer	Forme des groupes de <i>short reads</i> chevauchantes et les assemble	E : Fasta (<i>paired-end</i> possible)/ S : Fasta	Moyenne, quelques dépendances	Ils sont utiles lorsque seul un assemblage génomique fragmenté est disponible	Ils détectent différentes séquences répétées, il est donc nécessaire de dissocier les ETs. Les séquences d'ETs sont partiellement construites.	Parmi ces trois outils, les plus efficaces sont RepeatExplorer et dnaPipeTEs
dnaPipeTE	Réalise un échantillonnage aléatoire de <i>short reads</i> et les assemble	E : Fastq (<i>paired-end</i> non possible)/ S : Fasta	Facile, peu de dépendances			
RepArk	Recense les k-mer les plus abondantes dans les <i>short reads</i> et les assemble	E : Fastq: (<i>paired-end</i> non possible)/ S : Fasta	Facile, peu de dépendances			
Repeat Scout	Détection de longues k-mer dans un assemblage de génome	E : Fasta/ S : Fasta	Facile, peu de dépendances	Ils permettent de détecter des ETs à l'aide de méthodologies de novo	Ils détectent différentes séquences répétées, il est donc nécessaire de dissocier les ETs. Ils sont moins efficaces avec un assemblage génomique fragmenté ou pour détecter des ETs peu répétés.	Parmi l'ensemble des outils utilisés TEdenovo est l'un des outils les plus efficaces pour détecter des ETs complets. RepeatScout est moins efficace.
TEdenovo (REPET)	Comparaison de l'assemblage génomique sur lui même pour identifier des séquences répétées	E : Fasta/ S : Fasta				
PASTE (REPET)	Classification des séquences détectées : recherche de structures de séquences et comparaison par rapport à des banques de séquences nucléiques, protéiques ou profils HMM	E : Fasta/ S : Fasta et format .Classif	Difficile : installation de bases de données MySQL et de Sun Grid Engine (SGE)			
TEannot (REPET)	Il combine les outils d'annotation RepeatMasker, Censor et BLASTER	E : Fasta/ S : GFF		Il rejoint les annotations partielles d'un même ETs et enlève les annotations de séquences satellites	Taux de faux positif difficile à estimer	Outil performant pour réaliser une annotation d'ETs

LTR harvest	Recherche les terminaisons LTRs pour identifier des éléments autonomes ou non autonomes	E : Fasta/ S : Fasta et GFF	Facile, peu de dépendances	Il détecte la majorité des éléments LTR complets	Il détecte des faux positifs, il est ainsi important de vérifier les séquences détectées	Outil intéressant comme complément pour détecter des LTR complets
Mgescan_nonLTR	Recherche des domaines protéiques de LINE à l'aide de profils HMM	E : Fasta/ S : Fasta	Facile, peu de dépendances	Détecte des ETs complets, aucun faux positif détecté dans les tests effectués	Outil tr strict et passe à coté d'éléments atypiques	Outil intéressant comme complément pour détecter des LINE complets
Mite Hunter	Recherche des motifs de séquences de MITEs	E : Fasta/ S : Fasta	Facile, peu de dépendances	Permet de détecter des MITEs complets, ce qui s'avère une tâche compliquée	Présence de faux positifs	Outil intéressant comme complément pour détecter des MITEs complets
SINEfinder	Recherche des motifs de séquences de SINEs	E : Fasta/ S : Fasta	Facile, peu de dépendances	Permet de détecter des SINEs complets, ce qui s'avère une tâche compliquée	Présence de faux positifs	Outil intéressant comme complément pour détecter des SINEs complets
HelSearch	Recherche des motifs de séquences d'Helitrons	E : Fasta/ S : Fasta	Facile, peu de dépendances	Permet de détecter des structures d'Helitron, ce qui s'avère une tâche compliquée	Il génère de nombreux fichiers de sortie, qui sont difficile à analyser et détecte un grand nombre de faux positifs	Outil intéressant comme complément pour détecter des Helitrons, cependant il est probablement l'outil le moins précis et le moins efficace
Repeat Masker	Recherche des séquences similaires à des séquences nucléiques d'ETs recensés dans une banque de données de séquences	E : Fasta/ S : Fasta	Facile, peu de dépendances	Ils sont utiles pour détecter des ETs présents en peu de copies dans un assemblage génomique. Ils permettent d'utiliser des banques de données personnelles	Ils détectent principalement des morceaux d'ETs et non des séquences complètes	Ils sont les moins pertinents pour détecter des ETs complets et sont à utiliser comme compléments afin d'optimiser la détection de quelques familles peu répétées dans le génome
TE-HMMER	Recherche des séquences similaires à des séquences protéiques d'ETs recensés dans une banque de données de profils HMM	E : Fasta/ S : Fasta	Distribué à partir de la VM de PiRATE			
CD-HIT-est	Regroupe les séquences similaires et ne garde que la séquence la plus longue de chaque groupe	E : Fasta/ S : Fasta	Facile, peu de dépendances	Il permet de diminuer le nombre de séquences redondantes	Un % d'identité faible peut donner lieu au regroupement d'ETs de familles différentes	Outil rapide, flexible et efficace

Non utilisé						
Helitron Scanner	Recherche de séquences complètes d'Helitrons	na	Installation non reussie	na	na	na

Outils	Méthode	Données d'entrées (E) / de sorties (S)	Difficulté d'installation	Avantages	Inconvénients	Avis
PARTIE 2 : identification d'indels d'ETs						
LoRTE	Compare les <i>long reads</i> à un assemblage du génome pour identifier des indels d'ETs	E : Fasta et GFF/ S : sortie format spécifique	Facile, peu de dépendance	Outil dédié à la prédiction d'indels d'ETs, permet de connaître la structure de la séquence détectée	Sortie difficile à exploiter, pas de .vcf, les <i>long reads</i> possèdent pour le moment un taux d'erreur important	Efficace pour détecter des indels d'ETs homozygotes, mais nécessite une forte couverture pour identifier des indels d'ETs hétérozygotes
Mobster	Recherche des alignements discordants de paires de <i>short reads</i> pouvant correspondre à des insertions d'ETs et les compare à une banque de données d'ETs	E : bam et GFF/ S : format spécifique, peut être converti en .vcf	Moyenne, quelques dépendances	Outil dédié à la prédiction d'insertions d'ETs	Ne permet pas de savoir si les insertions détectées ont été générées par des ETs autonomes ou non autonomes	Nécessite un alignement réalisé avec l'outil MOSAIK. La sortie doit être convertie en .vcf
Break dancer	Recherche des alignements discordants de paires de <i>short reads</i> pouvant correspondre à des délétions d'ETs	E : bam/ S : format spécifique, peut être converti en .vcf	Facile, peu de dépendances	Ne nécessite pas d'annotation préalable des ETs. Il permet d'identifier des délétions d'ETs.	Parmi les délétions prédites, il est nécessaire d'identifier celles qui ont été générées par des ETs	Outil rapide et simple à utiliser, attention aux faux positifs
Pindel	Recherche des alignements discordants de paires de <i>short reads</i> (tronquées), pouvant correspondre à des délétions d'ETs	E : bam/ S : format spécifique, peut être converti en .vcf	Facile, peu de dépendance	Ne nécessite pas d'annotation préalable des ETs. Il permet d'identifier des délétions d'ETs.	Parmi les délétions prédites, il est nécessaire d'identifier celles qui ont été générées par des ETs	Outil rapide et simple à utiliser

Annexe 3

Analyse des IS candidats avec NCBI-BLASTx, ORFfinder et Interproscan. Le seuil des E-value pour BLASTn et BLASTx est inférieur à 10^{-2} .

La couleur rose représente les E-value inférieures à 10^{-2} , la couleur verte correspond à une E-value située entre 10^{-12} et 10^{-2} . HP signifie protéine hypothétique.

Tableau provenant de Chedhomme et Badawi, 2018 mémoire de Master 1.

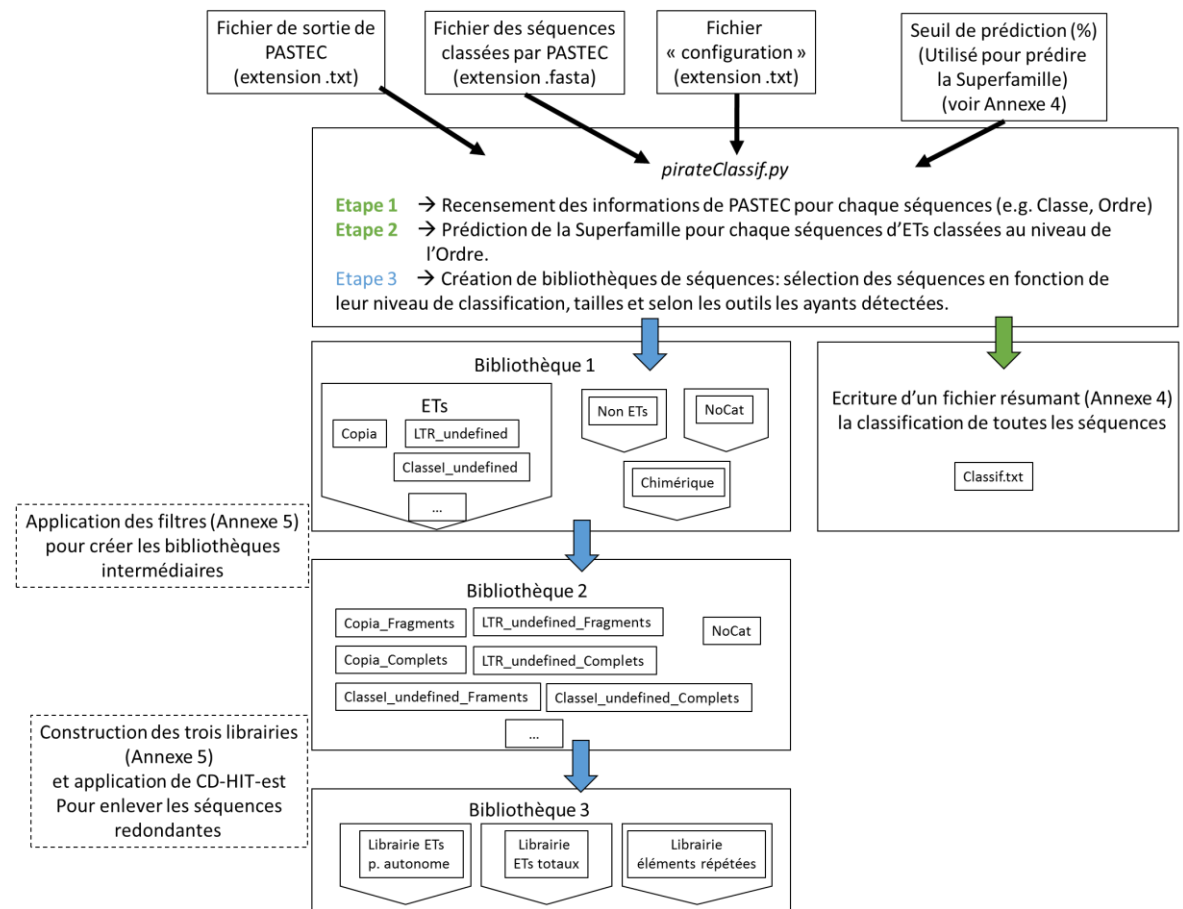
Outils		BLASTx					ORF finder	Interproscan
nom	Taille (pb)	Description	Converture (%)	E-Value	Identité (%)	Localisation (pb)	Taille ORF (pb)	Description
IS-like 1	1919	HP [<i>Emilianiaphuxleyipvirusp202</i>] IS5 family transposase [<i>Leptospirapkinschnei</i>]	15% 16%	1E-15 2E-04	47% 35%	907-1194 898-1218	530	Putative transposase IS4/IS5
IS-like 2	1325	IS5 family transposase [<i>Methanosarcinasp.</i>]	14%	3,00E-03	38%	507-698	272	Putative transposase IS4/IS5
IS-like 3	1553	HP [<i>Emilianiaphuxleyipvirusp202</i>] IS5 family transposase [<i>Caballeroniaphumi</i>]	24% 36%	3E-23 6E-18	45% 36%	411-797 216-785	893	Putative transposase IS4/IS5
IS-like 4	1309	HP [<i>Emilianiaphuxleyipvirusp202</i>] IS5 family transposase [<i>Leptospirapkinschnei</i>]	29% 27%	2E-23 3E-08	45% 38%	473-862 485-847	407	Putative transposase IS4/IS5
IS-like 5	1070	IS5 family transposase [<i>Anaerotruncus</i> sp.]	22%	2,00E-04	34%	10-249	248	Putative transposase IS4/IS5
IS-like 6	1093	HP [<i>Emilianiaphuxleyipvirus202</i>] Transposase [<i>Ca.pyyntrophoarchaeumcaldarius</i>]	27% 20%	9E-11 3E-05	40% 36%	488-784 557-784	320	Putative transposase IS4/IS5
IS-like 7	1016	HP [<i>Emilianiaphuxleyipvirus88</i>]	51%	5,00E-24	34%	490-1014	596	Transposase IS605, OrfB, C-terminal
IS-like 8	1045	Ribosomal protein [<i>Burkholderiappseudomallei</i>] HP [<i>Mooreaproducens</i>]	33% 17%	1E-09 6E-07	47% 48%	11-358 42-209	98	non prédit

Outils		BLASTx					ORF	Outils
nom	Taille (pb)	Organisme	Converture (%)	E-Value	Identité (%)	Localisation (pb)	Taille ORF (pb)	Description
IS-like 9	1040	/	/	/	/	/	545	non prédit
IS-like 10	1099	/	/	/	/	/	167	non prédit
IS-like 11	1496	/	/	/	/	/	294	non prédit
IS-like 12	1300	/	/	/	/	/	455	non prédit
IS-like 13	1254	/	/	/	/	/	285	non prédit
IS-like 14	1039	IS5 family transposase [<i>Methanosarcinasp.</i>]	18%	2,00E-03	38%	474-665	272	Putative transposase IS4/IS5
IS-like 15	1626	HP [<i>Emilianiaphuxleyipvirusp202</i>] IS5 family transposase [<i>Methanosarcinapmazei</i>]	30% 33%	3E-29 3E-10	42% 32%	590-1087 578-114	614	Putative transposase IS4/IS5, IS4-like
IS-like 16	955	HP [<i>Emilianiaphuxleyipvirusp202</i>] Transposase [Ca. <i>Syntrophoarchaenm caldarins</i>]	21% 21%	6E-04 9E-03	41% 37%	453-659 513-713	356	Putative transposase IS4/IS5
IS-like 17	1557	HP [<i>Emilianiaphuxleyipvirusp202</i>] IS5 family transposase [<i>Caballeroniaphumi</i>]	31% 34%	1E-30 4E-13	42% 35%	627-1124 690-1220	761	Putative transposase IS4/IS5, IS4-like
IS-like 18	1156	IS5 family transposase [<i>Methanosarcinasp.</i>]	40%	2,00E-11	29%	415-882	467	Putative transposase IS4/IS5
IS-like 19	1170	/	/	/	/	/	515	non prédit
IS-like 20	1171	Predicted protein [<i>Nematostellapvectensis</i>] RNA pol. sigma factor [<i>Mooreaproducens</i>]	32% 29%	3E-09 3E-08	36% 39%	546-866 600-851	620	non prédit

Annexe 4

Schéma représentatif des processus du programme pirateClassif.

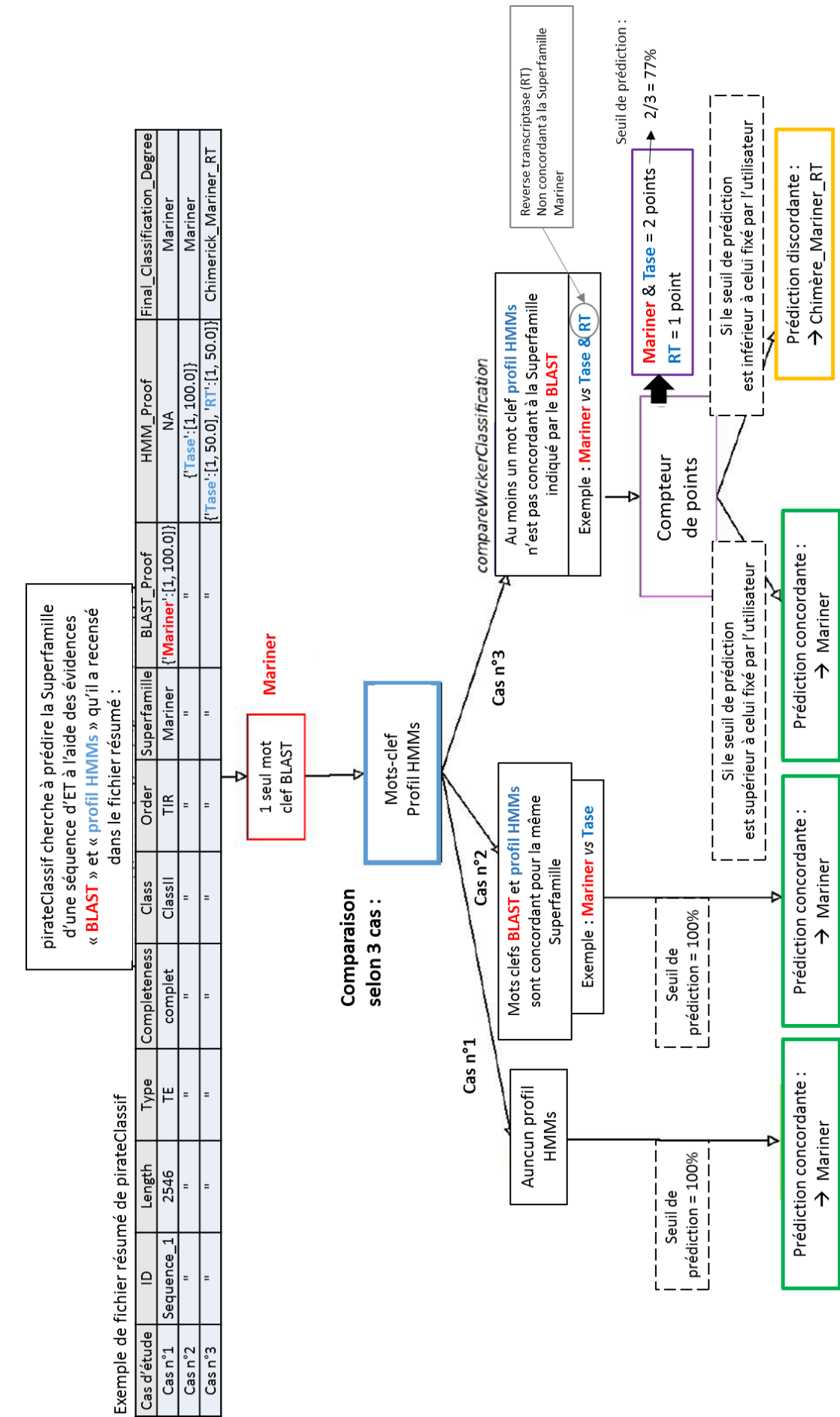
pirateClassif génère trois librairies en « poupées russes » (cf. Chapitre 1). Il crée également un fichier résumé de la classification. Image adaptée de T. Frances et al. mémoire de Master 2.



Annexe 5

Illustration de la méthode de prédiction de Superfamille du programme pirateClassif.

La classification est réalisée par comparaison des évidences « BLAST » et « profil HMMs » identifiées par PASTEC (cf. Chapitre 1). Ce diagramme, illustre les cas de figures possibles lorsqu' une seule évidence « BLAST » est recensée par pirateClassif pour la séquence d'étude. D'autres cas de figures sont possibles : « aucune évidence BLAST » et « plusieurs évidences BLAST ». Ils sont détaillés dans Frances et al. (2018) mémoire de Master 2.



Annexe 6

Aperçu du fichier « configuration.txt » employé par le programme pirateClassif.

Ce fichier est configurable et rassemble les indications choisies par l'utilisateur pour la construction des bibliothèques 2 (jaune) et 3 (vert) (Annexe 4). Plusieurs lignes du fichier, représentées par des « ... » ont été enlevées dans cet aperçu.

Bibliothèque 1 sources	Bibliothèque 2 à construire (filtres = colonnes 3 et 4)	Longueur des séquences à sélectionner	Outils dont les séquences détectées		Séquences à sélectionner pour la construction des trois librairies (Bibliothèque 3)		
			doivent pas être sélectionnés	doivent être uniquement utilisées	ETs potentiellement autonomes	ETs totaux	Éléments répétées
Fichiers	Fichiers FL= full lenght frag= fragment	min:max					
Copia	Copia_FL_sorted	3000:18000	helsearch	na	yes	yes	yes
Copia	Copia_frag_sorted	200:3000	helsearch	na	no	yes	yes
...	...	...	...	...	...	...	...
L1	L1_FL_sorted	2000:18000	ltharvest:helsearch	na	yes	yes	yes
L1	L1_frag_sorted	200:2000	ltharvest:helsearch	na	no	yes	yes
...	...	...	...	...	...	...	...
tRNA	tRNA_sorted	100:1000	ltharvest:helsearch	na	no	yes	yes
...	...	...	...	...	...	...	...
Mariner	Mariner_FL_sorted	1000:10000	ltharvest:helsearch	na	yes	yes	yes
Mariner	Mariner_frag_sorted	200:1000	ltharvest:helsearch	na	no	yes	yes
...	...	...	...	...	...	...	...
Maverick	Maverick_FL_sorted	4000:18000	ltharvest:helsearch	na	yes	yes	yes
Maverick	Maverick_frag_sorted	200:4000	ltharvest:helsearch	na	no	yes	yes
LTR_undefined	LTR_undefined_FL_sorted	3000:18000	helsearch	na	yes	yes	yes
LTR_undefined	LTR_undefined_frag_sorted	200:3000	helsearch	na	no	yes	yes
...	...	...	...	...	...	...	...
TIR_undefined	TIR_undefined_FL_sorted	1000:10000	ltharvest:helsearch	na	yes	yes	yes
TIR_undefined	TIR_undefined_frag_sorted	200:1000	ltharvest:helsearch	na	no	yes	yes
LINE_undefined	LINE_undefined_FL_sorted	2000:18000	ltharvest:helsearch	na	yes	yes	yes
...	...	...	...	...	...	...	...
MITE_undefined	MITE_sorted	100:1500	na	Mite-hunter:Tedenovo:RepeatScout	no	yes	yes
LARD_undefined	LARD_sorted	3000:18000	na	Tedenovo:RepeatScout	no	yes	yes
TRIM_undefined	TRIM_sorted	500:5000	na	Tedenovo:RepeatScout	no	yes	yes
ClassI_undefined	ClassI_undefined_sorted	200:18000	na	Tedenovo:RepeatScout	no	yes	yes
ClassII_undefined	ClassII_undefined_sorted	200:18000	na	Tedenovo:RepeatScout	no	yes	yes
noCAT	noCAT_sorted	200:18000	na	Tedenovo:RepeatScout	no	no	yes

Annexe 7

Liste d'outils visant la prédiction d'insertions et délétions de séquences à l'aide de données génomiques *short reads* ou *long reads*.

Nom	Auteurs
<i>short reads</i>	
Outils non spécialisés pour les ETs	
SVachra	Hampton et al. 2017
TARDIS	Soylev et al. 2017
TIDDIT	Eisfeldt et al. 2017
SvABA	Wal.a et al. 2017
COSMOS	Yamagata et al. 2016
Seeksv	Liang et al. 2016
Breakseq2	Abyzov et al. 2015
CODEX	Jiang et al. 2015
MetaSV	Mohiyuddin 2015
ONCOCNV	Boeva et al. 2014
MindTheGap	Rizk et al. 2014
DiscoVar	Weisenfeld et al. 2014
Socrates	Schroder et al. 2014
PSCC	Li et al. 2014
Gustaf	Trappe et al. 2014
XHMM	Bansai et al. 2014
LUMPY	Layer et al. 2014
MATE-CLEVER	Marschal.l et al. 2013
Excavator	Magi et al. 2013
SoftSearch	Hart et al. 2013
Tigra	Chen et al. 2013
Spliteread	Karakoc et al. 2012
CLEVER	Marschal.l et al. 2012
cn.MOPS	Klambauer et al. 2012
CNAnorm	Gusnanto et al. 2012
CoNIFER	Krumm et al. 2012
DELLY	Rausch et al. 2012
Cortex	lqbal. et al. 2012
PRISM	Jiang et al. 2012
Magnolya	Nijkamp et al. 2012
CREST	Wang et al. 2011
inGAP-sv	Qi et Zhao 2011
JoinSML	Magi et al. 2011
BIC-seq	Xi et al. 2011
CNVnator	Abyzov et al. 2011
ExomeCNV	Sathirapongsasuti et al. 2011
Genome STRiP	Handsaker et al. 2011
ReadDepth	Miller et al. 2011
SegSeq	Abyzov et al. 2011
Svseq	Zhang et Wu 2011
rSW-seq	Kim et al. 2010
CNAseg	Ivakhno et al. 2010
CNVer	Medvedev et al. 2010
SVDetect	Zeitouni et al. 2010
GASVpro	Sindi et al. 2010
HYDRA	Quinlan et al. 2010
Breckseq	Hugo et al. 2009
BreakDancer	Chen et al. 2009

Nom	Auteurs
<i>short reads (suite)</i>	
Outils non spécialisés pour les ETs (suite)	
MoDIL	Lee et al. 2009
PEMer	Korbel et al. 2009
Pindel	Ye et al. 2009
CNVseq	Xie et Tammi 2009
Outils spécialisés pour les ETs	
McClintock	Nelson et al. 2017
STEAK	Santander et al. 2017
RelocaTE2	Chen et al. 2017
TIGER	Tica et al. 2016
PTEMP	Kang et al. 2016
TEBreak	Carreira et al. 2016
splitreader	Quadrana et al. 2016
PoPoolationTE2	Kofler et al. 2016
T-lex2	Fiston-Lavier et al. 2015
TIDAL	Rahman et al. 2015
Jitterbug	Hénaff et al. 2015
ITIS	Jiang et al. 2015
MELT	Ewing et al. 2015
TEMP	Zhuang et al. 2014
discord-retro	Ballinger et al. 2015
lsmapper	Hawkey et al. 2015
Tangram	Wu et al. 2014
TIF	Nakagome et al. 2014
TE-Tracker	Gilly et al. 2014
TranspoSeq	Helman et al. 2014
pecnv	Cridland et al. 2013
Mobster	Thung et al. 2013
RetroSeq	Keane et al. 2013
RelocaTE	Robb et al. 2013
Alu-detect	David et al. 2013
TEA	Lee et al. 2012
ngs_te_mapper	Linheiro et Bergman, 2012
PoPoolationTE	Kofler et al. 2012
TE-locate	Platzner et al. 2012
Pool-Seq	Kofler et al. 2012
T-lex1	Fiston-Lavier et al. 2011
VariationHunter	Hormozdiari et al. 2011
L1-Seq	Ewing et Kazazian, 2010
ME-Scan	Witherspoon et al, 2010
NovelSeq	Hajirasouliha et al. 2010
<i>long reads</i>	
Outils spécialisés pour les ETs	
LoRTE	Disdero et Filée, 2017
Outils non spécialisés pour les ETs	
SMRT- SV	Huddleston et al. 2017
SNIFFLES	Sedlazeck et al. 2018

Annexe 8

Liste de 144 insertions d'ETs qui ont été prédites par le programme Mobster et qui sont potentiellement spécifiques au génome de la souche domestiquée Tiso-LipidePlus.

ID Tiso_S2M2	Famille de l'ET	Contig	Position insertion	Nombre reads soutenant L'allele de reference	Nombre reads tronquées soutenant l'insertion de l'ET	Fréquence allélique
Mobster483	<i>Ace</i>	131	20173	0	14	1
Mobster623	<i>Ace</i>	170	217539	4	7	0,64
Mobster297	<i>Luffy</i>	92	1125	4	6	0,6
Mobster168	<i>Luffy</i>	52	1156188	10	12	0,55
Mobster190	<i>Ace</i>	57	607184	23	24	0,51
Mobster101	<i>Ace</i>	37	427642	26	26	0,5
Mobster102	<i>Luffy</i>	37	136075	12	11	0,48
Mobster456	<i>Shanks</i>	127	894448	10	9	0,47
Mobster686	<i>Ace</i>	178	109926	13	11	0,46
Mobster342	<i>Ace</i>	97	1956140	31	25	0,45
Mobster628	<i>Luffy</i>	173	1595652	22	18	0,45
Mobster590	<i>Ace</i>	161	40364	33	26	0,44
Mobster630	<i>Ace</i>	173	2311585	14	10	0,42
Mobster485	<i>Ace</i>	134	1092065	13	9	0,41
Mobster495	<i>Ace</i>	137	568721	10	7	0,41
Mobster457	<i>Ace</i>	127	1614558	14	9	0,39
Mobster301	<i>Ace</i>	93	2113000	32	20	0,38
Mobster380	<i>Ace</i>	101	184729	29	18	0,38
Mobster740	<i>Ace</i>	189	830328	20	12	0,38
Mobster22	<i>Luffy</i>	12	47276	13	8	0,38
Mobster240	<i>Ace</i>	71	808821	45	26	0,37
Mobster699	<i>Luffy</i>	179	341961	39	23	0,37
Mobster566	<i>Ace</i>	155	311560	27	16	0,37
Mobster501	<i>Ace</i>	138	4757	25	14	0,36
Mobster119	<i>Ace</i>	42	124429	18	10	0,36
Mobster207	<i>Ace</i>	63	390535	14	8	0,36
Mobster503	<i>Luffy</i>	139	1732642	14	8	0,36
Mobster444	<i>Shanks</i>	116	48233	13	7	0,35
Mobster594	<i>Ace</i>	162	597096	11	6	0,35
Mobster629	<i>Ace</i>	173	1833109	29	14	0,33
Mobster191	<i>Ace</i>	57	24461	26	13	0,33
Mobster44	<i>Ace</i>	16	32483	22	11	0,33
Mobster700	<i>Ace</i>	179	654175	22	11	0,33
Mobster175	<i>Ace</i>	53	278496	16	8	0,33
Mobster162	<i>Ace</i>	47	209355	36	17	0,32
Mobster707	<i>Ace</i>	181	269331	23	11	0,32
Mobster208	<i>Ace</i>	63	711379	17	8	0,32
Mobster633	<i>Ace</i>	173	1155942	17	8	0,32

ID Tiso_S2M2	Famille de l'ET	Contig	Position insertion	Nombre reads soutenant L'allele de reference	Nombre reads tronquées soutenant l'insertion de l'ET	Fréquence allélique
Mobster118	<i>Ace</i>	42	184433	46	21	0,31
Mobster542	<i>Ace</i>	150	2652711	44	20	0,31
Mobster281	<i>Ace</i>	89	694862	29	13	0,31
Mobster616	<i>Shanks</i>	168	474666	27	12	0,31
Mobster386	<i>Ace</i>	103	35850	18	8	0,31
Mobster13	<i>Ace</i>	9	768630	21	9	0,3
Mobster459	<i>Ace</i>	127	14960	14	6	0,3
Mobster524	<i>Ace</i>	144	522052	44	18	0,29
Mobster685	<i>Ace</i>	178	1141139	44	18	0,29
Mobster2	<i>Ace</i>	4	242651	34	14	0,29
Mobster543	<i>Ace</i>	150	2155439	34	14	0,29
Mobster254	<i>Ace</i>	73	434846	30	12	0,29
Mobster272	<i>Ace</i>	86	411209	25	10	0,29
Mobster434	<i>Ace</i>	115	1345124	17	7	0,29
Mobster271	<i>Shanks</i>	86	172943	36	14	0,28
Mobster544	<i>Ace</i>	150	168429	33	13	0,28
Mobster306	<i>Ace</i>	93	1914519	31	12	0,28
Mobster155	<i>Ace</i>	46	818899	26	10	0,28
Mobster716	<i>Luffy</i>	184	1438565	18	7	0,28
Mobster396	<i>Ace</i>	106	25176	49	18	0,27
Mobster343	<i>Ace</i>	97	1811743	46	17	0,27
Mobster540	<i>Ace</i>	149	60500	37	14	0,27
Mobster111	<i>Shanks</i>	41	117444	32	12	0,27
Mobster93	<i>Harbinger7</i>	32	331612	36	13	0,27
Mobster593	<i>Ace</i>	162	603564	19	7	0,27
Mobster117	<i>Ace</i>	42	13970	67	24	0,26
Mobster235	<i>Ace</i>	68	75062	70	24	0,26
Mobster478	<i>Ace</i>	130	32684	58	20	0,26
Mobster494	<i>Gypsy1</i>	136	57338	32	11	0,26
Mobster458	<i>Ace</i>	127	318087	25	9	0,26
Mobster303	<i>Ace</i>	93	1735398	44	15	0,25
Mobster302	<i>Ace</i>	93	1519619	59	19	0,24
Mobster304	<i>Ace</i>	93	2152292	45	14	0,24
Mobster255	<i>Shanks</i>	73	254029	41	13	0,24
Mobster344	<i>Ace</i>	97	1965621	37	12	0,24
Mobster413	<i>Shanks</i>	108	71877	31	10	0,24
Mobster519	<i>Luffy</i>	143	78789	31	10	0,24
Mobster701	<i>Ace</i>	179	918158	25	8	0,24
Mobster381	<i>Ace</i>	101	23616	19	6	0,24
Mobster45	<i>Shanks</i>	16	260041	33	10	0,23
Mobster270	<i>Shanks</i>	84	4984	34	10	0,23

ID Tiso_S2M2	Famille de l'ET	Contig	Position insertion	Nombre reads soutenant L'allele de reference	Nombre reads tronquées soutenant l'insertion de l'ET	Fréquence allélique
Mobster427	<i>Ace</i>	109	60615	33	10	0,23
Mobster634	<i>Ace</i>	173	2189116	23	7	0,23
Mobster53	<i>Ace</i>	23	432094	36	10	0,22
Mobster54	<i>Ace</i>	23	430038	29	8	0,22
Mobster546	<i>Ace</i>	150	2155803	28	8	0,22
Mobster192	<i>Ace</i>	57	607859	25	7	0,22
Mobster26	<i>Ace</i>	13	627650	21	6	0,22
Mobster663	<i>Ace</i>	174	168362	65	17	0,21
Mobster226	<i>Ace</i>	65	208335	57	15	0,21
Mobster305	<i>Ace</i>	93	679078	58	15	0,21
Mobster525	<i>Ace</i>	144	489910	55	15	0,21
Mobster112	<i>Ace</i>	41	688564	50	13	0,21
Mobster10	<i>Ace</i>	8	274287	31	8	0,21
Mobster448	<i>Ace</i>	119	337529	31	8	0,21
Mobster688	<i>Ace</i>	178	1115599	27	7	0,21
Mobster40	<i>Ace</i>	15	1245978	22	6	0,21
Mobster55	<i>Ace</i>	23	514265	22	6	0,21
Mobster72	<i>Ace</i>	26	174428	22	6	0,21
Mobster307	<i>Ace</i>	93	692268	40	10	0,2
Mobster336	<i>Ace</i>	94	92806	24	6	0,2
Mobster71	<i>Ace</i>	26	114301	43	10	0,19
Mobster687	<i>Ace</i>	178	1429844	42	10	0,19
Mobster176	<i>Ace</i>	53	202926	39	9	0,19
Mobster734	<i>Ace</i>	185	113757	30	7	0,19
Mobster341	<i>Ace</i>	96	9067	69	15	0,18
Mobster715	<i>Ace</i>	184	1976655	51	11	0,18
Mobster504	<i>Ace</i>	139	427280	28	6	0,18
Mobster631	<i>Ace</i>	173	1473227	90	18	0,17
Mobster339	<i>Shanks</i>	95	410863	48	10	0,17
Mobster39	<i>Ace</i>	15	1111331	45	9	0,17
Mobster83	<i>Ace</i>	29	17041	40	8	0,17
Mobster717	<i>Ace</i>	184	1891118	39	8	0,17
Mobster460	<i>Ace</i>	127	1853207	30	6	0,17
Mobster609	<i>Gypsy2</i>	165	153230	30	6	0,17
Mobster286	<i>Shanks</i>	90	105335	59	11	0,16
Mobster6	<i>Shanks</i>	5	150142	46	9	0,16
Mobster520	<i>Ace</i>	143	311346	36	7	0,16
Mobster567	<i>Ace</i>	155	986515	31	6	0,16
Mobster610	<i>Ace</i>	165	136512	31	6	0,16
Mobster362	<i>Ace</i>	99	1088544	91	16	0,15
Mobster25	<i>Ace</i>	13	793378	50	9	0,15

ID Tiso_S2M2	Famille de l'ET	Contig	Position insertion	Nombre reads soutenant l'allele de reference	Nombre reads tronquées soutenant l'insertion de l'ET	Fréquence allélique
Mobster741	<i>Ace</i>	189	280939	45	8	0,15
Mobster382	<i>Ace</i>	101	151437	39	7	0,15
Mobster502	<i>Shanks</i>	138	11157	75	12	0,14
Mobster308	<i>Ace</i>	93	1308223	66	11	0,14
Mobster718	<i>Ace</i>	184	712611	70	11	0,14
Mobster199	<i>Ace</i>	59	138917	63	10	0,14
Mobster635	<i>Ace</i>	173	458076	49	8	0,14
Mobster449	<i>Ace</i>	119	400743	44	7	0,14
Mobster569	<i>Ace</i>	155	53128	43	7	0,14
Mobster146	<i>Ace</i>	44	1331060	37	6	0,14
Mobster73	<i>Ace</i>	26	169370	66	10	0,13
Mobster531	<i>Shanks</i>	145	50706	60	9	0,13
Mobster27	<i>Ace</i>	13	793686	45	7	0,13
Mobster636	<i>Ace</i>	173	1509686	46	7	0,13
Mobster383	<i>Ace</i>	101	171007	70	10	0,12
Mobster364	<i>Ace</i>	99	1041316	68	9	0,12
Mobster690	<i>Ace</i>	178	1438294	65	9	0,12
Mobster751	<i>Ace</i>	192	232482	67	9	0,12
Mobster719	<i>Ace</i>	184	2262059	58	8	0,12
Mobster692	<i>Ace</i>	178	273668	53	7	0,12
Mobster547	<i>Ace</i>	150	742138	62	8	0,11
Mobster266	<i>Ace</i>	82	90215	59	7	0,11
Mobster571	<i>Ace</i>	155	135898	58	7	0,11
Mobster611	<i>Ace</i>	165	130483	56	6	0,1

Annexe 9

Liste de 32 insertions d'ETs qui ont été prédites par le programme Mobster et qui sont potentiellement spécifiques au génome de la souche domestiquée Tiso-LipideMoins.

ID Tiso_S2M2	Famille de l'ET	Contig	Position insertion	Nombre reads soutenant l'allèle de référence	Nombre reads tronquées soutenant l'insertion de l'ET	Fréquence allélique
Mobster733	<i>Gypsy2</i>	175	19135	1	8	0,89
Mobster796	<i>Ace</i>	184	560523	28	22	0,44
Mobster375	<i>Ace</i>	95	556431	30	16	0,35
Mobster551	<i>Gypsy1</i>	136	57338	27	8	0,23
Mobster121	<i>Luffy</i>	37	500851	24	7	0,23
Mobster765	<i>Ace</i>	178	1144257	20	6	0,23
Mobster683	<i>Ace</i>	168	536880	32	9	0,22
Mobster317	<i>Ace</i>	88	55078	22	6	0,21
Mobster583	<i>Shanks</i>	142	157577	32	8	0,2
Mobster35	<i>Luffy</i>	13	901982	24	6	0,2
Mobster620	<i>Ace</i>	150	1630177	38	8	0,17
Mobster190	<i>Gypsy2</i>	46	677809	34	7	0,17
Mobster788	<i>Copia6</i>	182	243938	52	10	0,16
Mobster119	<i>Ace</i>	37	383296	47	9	0,16
Mobster602	<i>Gypsy1</i>	147	295999	35	6	0,15
Mobster302	<i>Ace</i>	78	103655	61	10	0,14
Mobster560	<i>Ace</i>	138	4573	43	7	0,14
Mobster808	<i>Ace</i>	188	409319	61	9	0,13
Mobster628	<i>Luffy</i>	154	35683	48	7	0,13
Mobster458	<i>Ace</i>	108	430953	39	6	0,13
Mobster523	<i>Ace</i>	127	1582256	41	6	0,13
Mobster524	<i>Ace</i>	127	1727262	60	8	0,12
Mobster280	<i>Ace</i>	71	271498	51	7	0,12
Mobster18	<i>Luffy</i>	9	714817	43	6	0,12
Mobster22	<i>Ace</i>	10	120933	55	7	0,11
Mobster320	<i>Gypsy1</i>	89	35380	59	7	0,11
Mobster74	<i>Ace</i>	23	268150	81	9	0,1
Mobster90	<i>Ace</i>	26	169639	66	7	0,1
Mobster464	<i>Ace</i>	110	82050	61	7	0,1
Mobster792	<i>Ace</i>	184	204040	63	7	0,1
Mobster113	<i>Ace</i>	37	133091	55	6	0,1
Mobster728	<i>Ace</i>	173	2887140	55	6	0,1

Annexe 10

Les séquences protéiques des ORF des ETs autonomes *Luffy*, *Ace* et *Shanks* ont été comparées à des séquences protéiques et des profils HMMs d'ETs provenant de différentes banques de données.

ORF <i>Luffy</i>	Hit	Accession	% Similarité	% identité	% Probabilité
Censor (local)	Mariner-12_HM	/	35	/	/
Geneious (global)	Mariner-12_HM	/	/	22	/
Pfam	na	na	/	/	/
HHpred	Mariner <i>Mos</i>	4U7B_G	/	10	97,32

ORF <i>Ace</i>	Hit	Accession	% Similarité	% identité	% Probabilité
Censor	hAT-7_ATr	/		/	/
Geneious	hAT-7_ATr	/	/	14	/
Pfam	Dimer_Tnp_hAT et DUF659	PF05699 et PF04937	/	/	/
HHpred	hAT <i>Hermes</i>	2BW3_A	/	16	100

ORF1 <i>Shanks</i>	Hit	Accession	% Similarité	% identité	% Probabilité
Censor	hATw-1_BF	/	29	na	na
Geneious	hATw-1_BF	/	/	15	na
Pfam	na	na	/	na	na
HHpred	hAT <i>Hermes</i>	2BW3_A	/	14	93,52
ORF2 <i>Shanks</i>	Hit	Accession	% Similarité	% identité	% Probabilité
Censor	hATw-1_BF	na	30	na	/
Geneious	hATw-1_BF	na	/	15	/
Pfam	na	na	/	/	/
HHpred	na	na	/	na	na

Annexe 11

Les ORF des ETs autonomes de *Luffy*, *Ace* et *Shanks* ont été comparés avec tBLASTn aux assemblages des génomes respectifs des Haptophytes *Emiliana huxleyi*, *Pavlova* sp. et *Chrysochromulina tobin*.

<i>Contigs</i>	Position début	Position fin	Identité	E-value
ORF <i>Luffy</i>				
Pavlovaes_scaffold_134	21808	20921	89/316 (28%)	2,00E-47
Pavlovaes_scaffold_93	174731	173604	130/398 (32%)	5,00E-44
Pavlovaes_scaffold_293	196573	195446	126/398 (31%)	6,00E-42
Pavlovaes_scaffold_33	133553	132426	126/398 (31%)	2,00E-40
Pavlovaes_scaffold_537	84718	83792	102/331 (30%)	2,00E-39
Pavlovaes_scaffold_601	29411	28281	125/399 (31%)	5,00E-39
Pavlovaes_scaffold_2	571480	570356	128/397 (32%)	9,00E-39
Pavlovaes_scaffold_109	91071	92195	125/396 (31%)	2,00E-38
Pavlovaes_scaffold_296	9581	8655	102/333 (30%)	6,00E-38
Pavlovaes_scaffold_49	436698	435571	124/398 (31%)	8,00E-37
Pavlovaes_scaffold_4	977689	976571	120/394 (30%)	2,00E-36
Pavlovaes_scaffold_130	15135	14209	93/330 (28%)	1,00E-35
Pavlovaes_scaffold_854	24775	25671	96/320 (30%)	1,00E-33
Pavlovaes_scaffold_51	325537	325229	50/104 (48%)	2,00E-30
Pavlovaes_scaffold_292	84707	84432	47/93 (50%)	4,00E-30
Pavlovaes_scaffold_323	147529	147915	52/131 (39%)	5,00E-29
Pavlovaes_scaffold_98	128962	130044	108/382 (28%)	2,00E-21
Pavlovaes_scaffold_335	33412	33053	48/120 (40%),	3,00E-15
Emiliana_scaffold_345	71444	71992	62/218 (28%)	2,00E-10
ORF <i>Ace</i>				
Chrysochromulina_tobin_Contig7589	6536	7465	90/338 (26%)	5E-22
Emiliana_scaffold_108	41544	41029	51/177 (28%)	0,000002

<i>Contigs</i>	Position début	Position fin	Identité	E-value
ORF1 <i>Shanks</i> (796 pb)				
Emiliana_scaffold_7	396462	397343	139/303 (45%)	2,00E-72
Emiliana_scaffold_9	88785	89087	39/104 (37%)	4,00E-16
Emiliana_scaffold_31	371841	370960	140/303 (46%)	1,00E-72
Emiliana_scaffold_36	753270	754022	75/260 (28%)	6,00E-13
Emiliana_scaffold_46	9650	8769	140/303 (46%)	2,00E-72
Emiliana_scaffold_67	512466	513347	140/303 (46%)	1,00E-72
Emiliana_scaffold_71	579079	579952	140/303 (46%)	1,00E-72
Emiliana_scaffold_73	527460	528333	140/303 (46%)	1,00E-72
Emiliana_scaffold_80	168844	167963	139/303 (45%)	9,00E-73
Emiliana_scaffold_117	165071	165952	140/303 (46%)	2,00E-72
Chrysochromulina_tobin_Contig6714	4877	4011	87/306 (28%)	4,00E-19
ORF2 <i>Shanks</i> (580 pb)				
Emiliana_scaffold_7	397451	398767	188/441 (42%)	7,00E-77
Emiliana_scaffold_9	89850	90959	162/383 (42%)	8,00E-72
Emiliana_scaffold_31	370852	369536	188/441 (42%)	3,00E-77
Emiliana_scaffold_36	754310	755230	124/313 (39%)	7,00E-43
Emiliana_scaffold_46	8661	7345	188/441 (42%)	3,00E-77
Emiliana_scaffold_67	513964	514761	105/267 (39%)	1,00E-68
Emiliana_scaffold_71	580068	581384	188/441 (42%)	3,00E-77
Emiliana_scaffold_73	528449	529765	188/441 (42%)	3,00E-77
Emiliana_scaffold_80	167855	166539	187/441 (42%)	1,00E-76
Emiliana_scaffold_117	166060	167376	188/441 (42%)	3,00E-77
Emiliana_scaffold_530	11907	13223	187/441 (42%)	2,00E-76
Emiliana_scaffold_1057	2557	1892	85/227 (37%)	8,00E-20
Chrysochromulina_tobin_Contig6714	3920	3126	86/272 (31%)	3,00E-21

Titre : Étude de l'activité des éléments mobiles chez *Tisochrysis lutea* dans un contexte d'amélioration des espèces de microalgues.

Mots clés : bio-informatique, évolution, génome, annotation, éléments transposables

Résumé : Les éléments transposables (ETs) sont des acteurs majeurs de la dynamique des génomes. Ils regroupent l'ensemble des séquences d'ADN capables de se déplacer et de se propager dans les génomes. L'objectif de cette thèse a été d'estimer la contribution des ETs dans la dynamique du génome de la microalgue *Tisochrysis lutea* (Haptophyte). Dans ce but, le séquençage et l'assemblage de novo du génome de cette espèce ont été réalisés. Afin de détecter, classer et annoter les ETs présents dans le génome de *T. lutea*, un pipeline bio-informatique nommé PiRATE (Pipeline to Retrieve and Annotate Transposable Elements) a été développé. Suite à cette annotation, l'activité de deux familles d'ETs candidates baptisées *Luffy* et *Ace* a été mise en évidence à l'aide de données transcriptomiques et protéomiques.

L'étude de la dynamique des ETs a ensuite été conduite chez deux souches domestiquées de cette espèce. La comparaison de leur génome respectif à celui de la souche d'origine a permis la prédiction d'insertions et délétions associées aux ETs. Les deux familles d'ETs identifiées précédemment ainsi qu'une troisième, appelée *Shanks*, participent de façon prépondérante à ces variations génétiques et semblent ainsi fortement impliquées dans la dynamique du génome de *T. lutea*.

Ce travail apporte les premières connaissances de l'activité des ETs chez les Haptophytes attestant ainsi de l'importance de ces éléments dans la dynamique des génomes de microalgues.

Title : Activity of mobile elements in *Tisochrysis lutea* in a context of microalgae domestication.

Keywords : bioinformatics, evolution, genome, annotation, transposable elements

Abstract : Transposable elements (TEs) are known as key drivers of the genome dynamics. They are defined as DNA sequences able to move and spread in genomes. The objective of this thesis was to estimate the contribution of TEs in the genome dynamics of the microalgae *Tisochrysis lutea* (Haptophyte). For this purpose, a new genome assembly of this species was performed. In order to detect, classify and annotate TEs, a new pipeline named PiRATE (Pipeline to Retrieve and Annotate Transposable Elements) was developed. Following this annotation, the activity of two candidate TE families named *Luffy* and *Ace* was revealed using transcriptomic and proteomic data.

The study of the TE dynamics was then conducted in two domesticated strains of *T. lutea*. Comparative analysis between the domesticated and the originated strains genomes highlighted insertions and deletions predictions associated to TEs. The two families previously identified, together with a third one that was called *Shanks*, mainly participate in these genetic variations and seem to be strongly involved in the genome dynamics of *T. lutea*.

This work provides the first knowledges regarding the activity of TEs in Haptophyte and demonstrate their importance in the genome dynamics of microalgae.