Impact des duplications génomiques sur l'évolution des eucaryotes

France Denoeud

Document de synthèse en vue de l'obtention d'une
Habilitation à Diriger des Recherches

Spécialité : Biologie des organismes

Par **France Denoeud**
CEA Institut François Jacob, Genoscope, Laboratoire de Bioinformatique pour la Génomique et la Biodiversité, UMR8030 Génomique Métabolique

Impact des duplications génomiques sur l'évolution des eucaryotes

Présentée et soutenue le 25 mars 2022
au Genoscope, Evry

Composition du Jury :

Laurent Duret Directeur de recherche CNRS UMR 5558	**rapporteur**
Christophe Plomion Directeur de recherche INRAE UMR 1202	**rapporteur**
Hugues Roest Crollius Directeur de recherche CNRS UMR8197	**rapporteur**
Susana Coelho Directrice de Recherche, Max Planck Institute	**examinatrice**
Mathieu Rousseau-Gueutin CR INRAE UMR 1349 IGEPP	**examinateur**
Olivier Lespinet Professeur Université Paris-Saclay	**examinateur**
Marie-Hélène Mucchielli Giorgi Professeur Université d'Evry Val d'Essonne	**examinatrice**

Table des matières

Introduction ... 3
Partie I. Les duplications génomiques ... 5
1/ Les duplications complètes de génomes ... 5
2/ Les duplications segmentales ... 9
3/ Devenir des gènes dupliqués ... 10
4/ Evolution des familles multigéniques ... 13
Partie II. Retour sur mon parcours scientifique ... 16
1/ Thèse à l'Institut de Génétique et Microbiologie, Orsay (2000-2003) ... 16
2/ Postdoctorat à l'IMIM/GRIB, Barcelone (2004-2006) ... 17
3/ Expérience de chercheur « junior » au Genoscope (2006-2009) ... 18
4/ Prise de responsabilités dans trois grands projets (2010-2014) ... 22
a/ Analyse du génome du bananier, *Musa acuminata*, paléopolyploïde ... 22
b/ Analyse du génome du colza, *Brassica napus*, polyploïde récent ... 23
c/ Analyse du génome du caféier, *Cophea canofora*, révélant des duplications en tandem dans les gènes de biosynthèse de la caféine ... 26
5/ Élaboration de projets de recherche personnels ... 28
a/ Projet *Polysuccess* : comment un polyploïde devient une nouvelle espèce ... 29
b/ *PhyloAlps* et *PhyloNorway* : « genome skimming » sur la flore alpine et arctique ... 31
c/ Phaeoexplorer : analyse des génomes de 40 espèces d'algues brunes ... 40
d/ *Tara Pacific* : mise en évidence de duplications en tandem massives dans deux génomes de coraux ... 43
6/ Discussion ... 51
Partie III. Futur projet de recherche ... 53
Conclusion ... 60
Références ... 62
ANNEXES ... 69
Liste des publications/communications ... 69
Liste des encadrés ou co-encadrés d'étudiants ... 74
Liste des collaborations / contrats ... 77

Introduction

En préambule, je souhaite dire que la rédaction de ce document d'HDR a été beaucoup plus délicate pour moi que je ne l'aurais cru de prime abord, car j'étais très inspirée pour parler de science, mais beaucoup moins pour parler de moi. Il s'est agi d'un exercice difficile mais très enrichissant, qui m'a permis de prendre du recul sur mes activités passées et de me projeter sur ma carrière future, et je suis heureuse de pouvoir vous présenter mes travaux. Je souhaite tout d'abord exposer ma motivation pour soutenir l'HDR. Je suis arrivée à un tournant de ma carrière où je souhaite prendre davantage de responsabilités en tant que PI pour des projets, et avoir la légitimité pour le faire. En outre, l'encadrement d'étudiants en thèse est à mes yeux une mission très importante que se doit de remplir tout chercheur accompli. Selon moi, la direction de thèses est certes une opportunité de faire avancer des axes de recherche mais surtout un devoir de formation envers la future génération de chercheurs. C'est pour ces raisons que j'ai choisi de candidater au diplôme d'HDR. J'ai déposé ma demande d'autorisation de soutenance d'HDR à l'Université d'Evry début 2019. Cette demande a été acceptée, et je remercie l'Université d'avoir considéré positivement mon profil un peu atypique. En effet, en tant que chercheuse au Genoscope, j'ai eu la chance d'être impliquée dans de très gros projets de séquençage, qui ont débouché sur des articles dans des revues très prestigieuses où j'étais souvent premier ou co-premier auteur. C'est un point très positif. Le revers de la médaille, cependant, est que mon travail personnel s'est trouvé souvent « noyé dans la masse » de l'ensemble du travail d'énormes consortiums. Du point de vue thématique, j'ai eu la chance de pouvoir souvent creuser ma propre question en parallèle des questions amenées par les collaborateurs. Dans les rares cas où cette liberté ne m'a pas été offerte, j'ai veillé à choisir judicieusement les projets dans lesquels je me suis engagée par la suite. Cependant, il est certain que la nécessité d'apporter des réponses aux questions scientifiques inhérentes au génome séquencé m'ont laissé moins de temps et de liberté pour creuser mon propre sillon, à des échelles plus vastes transcendant les projets de séquençage du Genoscope. Au-delà de ces contraintes thématiques, j'ai surtout été confrontée à des contraintes temporelles. En effet, les données produites par un consortium d'analyse n'ont vocation à être libérées que lorsque l'article du consortium est soumis, ce qui a généré chez moi une certaine réticence à encadrer des thésards. J'ai une grande expérience des délais toujours beaucoup plus longs qu'escomptés lorsqu'il s'agit de rédiger des articles de consortium. J'ai été témoin des déconvenues d'étudiants en thèse ou de postdoctorants dont

le calendrier n'était pas du tout en adéquation avec celui d'un grand consortium et c'est un risque que je refusais de faire courir à mon étudiant(e). C'est pourquoi je n'ai pas encore d'article en dernier auteur avec un de mes étudiants en premier auteur. J'arrive maintenant à une période beaucoup plus favorable, car je suis l'un des partenaires principaux d'un projet qui arrive à la phase d'analyse (projet « Pheaoexplorer ») : je suis la coordinatrice de l'une des thématiques d'analyse, j'aurai donc une grande liberté d'orienter mes recherches, et je propose un sujet pour encadrer un thésard sur ce sujet. Par ailleurs, j'ai prévu de publier un article décrivant un outil développé actuellement par mon stagiaire de dernière année d'école d'ingénieur, sur la prédiction de tailles de génomes de plantes, avant la fin de son stage (dans quelques mois). Les conditions auraient certainement été plus favorables pour moi si j'avais soutenu mon HDR un an plus tard, mais j'ai choisi de relever le défi de vous convaincre dès aujourd'hui de mes capacités à diriger des recherches.

Lors de ma carrière scientifique dédiée à l'exploration de génomes, l'une des premières choses qui m'a « sauté aux yeux » est la présence de duplications génomiques. Bien-sûr, les éléments répétés les plus « visibles », qui influent grandement sur la taille des génomes, sont les éléments transposables. Nous nous intéresserons ici plutôt aux duplications de gènes (ou de génomes entiers) et à leurs conséquences fonctionnelles. La présence de gènes dupliqués peut être massive, si l'on étudie un polyploïde récent (par exemple le colza (Chalhoub, Denoeud, *et al*., 2014)) ou un organisme dont le génome a accumulé un très grand nombre de duplications en tandem (par exemple les coraux (Noel, Denoeud, *et al*., en préparation)). Ces duplications peuvent aussi être plus discrètes mais de grande importance pour la compréhension de l'évolution des génomes, comme les mécanismes de création d'introns (par exemple chez le tunicier *Oikopleura dioica* (Denoeud *et al*., 2009)) ou l'évolution de voies métaboliques d'intérêt (par exemple la biosynthèse de la caféine (Denoeud *et al*., 2012)). Ces différents exemples correspondent à des travaux que j'ai menés depuis le début de ma carrière au Genoscope. Il m'a donc semblé pertinent de choisir le thème des duplications de gènes comme fil directeur de ce document, d'autant plus qu'il rejoint même le sujet de ma thèse qui portait sur les minisatellites polymorphes : d'une certaine façon, je peux ainsi dire que « la boucle est bouclée ». Dans ce manuscrit, je vais tout d'abord définir les duplications génomiques et dresser un état des connaissances actuelles sur l'évolution des gènes dupliqués et des familles multigéniques (partie I). Je reviendrai ensuite chronologiquement sur mon parcours de recherche et d'encadrement et détaillerai mes travaux (partie II). Enfin, je présenterai mon futur projet de recherche (partie III).

Partie I. Les duplications génomiques

Dans cette partie, je fais une revue bibliographique de l'état de la connaissance concernant les duplications génomiques. On peut en distinguer deux types : les duplications complètes de génomes (WGD pour « whole genome duplications ») et les duplications segmentales qui ne concernent que des portions du génome.

1/ Les duplications complètes de génomes

De nombreuses espèces eucaryotes ont subi une ou plusieurs duplications complètes (WGD) de leur génome au cours de leur évolution. Ce phénomène est fréquent chez les plantes (Blanc, 2004 ; Tuskan, 2006 ; Jaillon, 2007 ; Schmutz, 2010 ; Jiao, 2012 ; Jiao, 2014 ; Murat, 2017), où 30 à 80 % des espèces sont polyploïdes (Masterson, 1994 ; Ren, 2018, **Figure 1**). Ces duplications complètes de génomes sont corrélées avec l'explosion du nombre d'espèces chez les angiospermes (De Bodt, 2005 ; Freeling, 2006 ; Leitch, 2008). Elles sont survenues de façon concomitante dans plusieurs lignées à des périodes correspondant à des variations climatiques importantes. Ces évènements de duplication génomique ont vraisemblablement joué un rôle prépondérant dans l'évolution des plantes à fleurs, permettant des innovations morphologiques et physiologiques clefs, ouvrant la voie à la colonisation de nouveaux environnements (Soltis, 2016). En effet, il est probable que les WGDs ont pu faciliter l'apparition de nouvelles espèces de plantes résistantes aux nouvelles conditions, en particulier en provoquant l'amplification de familles de gènes impliquées dans la résistance au stress (Wu, 2020).

Des duplications complètes de génomes ont également été observées chez les levures (Wolfe, 1997 ; Kellis, 2004), les ciliés (Aury, 2006), les arachnides (Schwager, 2017) et les vertébrés où il a été montré que deux duplications complètes de génomes sont survenues il y a 450 millions d'années (événement nommé « 2R ») (Dehal, 2005). D'autres duplications sont ensuite survenues dans la lignée des poissons téléostéens (Glasauer, 2014), ainsi que dans la famille des salmonidés (Berthelot, 2014).

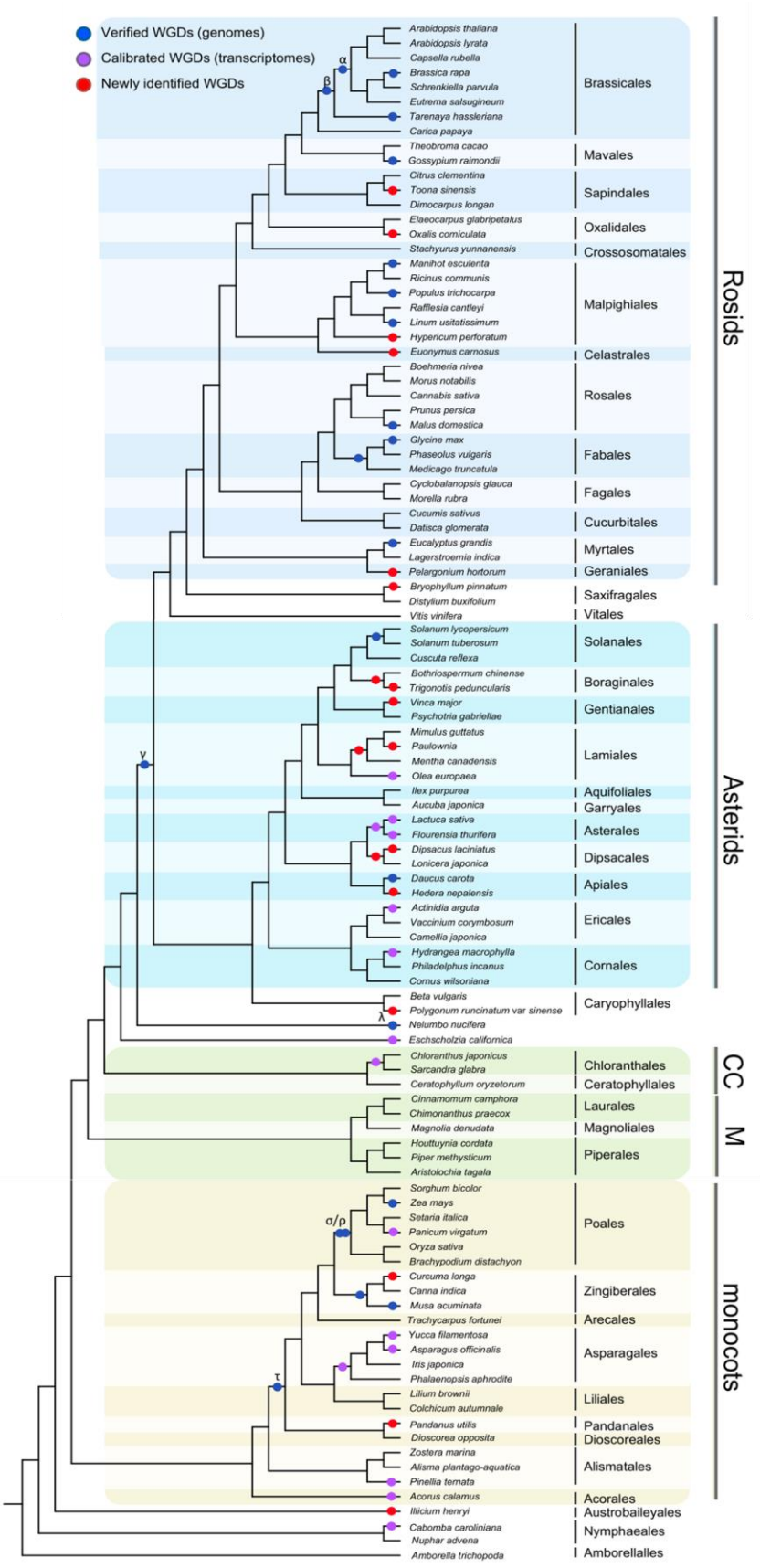

Figure 1 (d'après Ren, 2018 : Figure 3) Phylogénie de 105 angiospermes avec les WGDs détectées (ronds bleus, violets et rouges)

Il existe deux mécanismes permettant la duplication complète d'un génome (**Figure 2**). Le premier, l'autoploïdie, correspond à la non-disjonction des chromosomes dans la lignée germinale au cours de la méiose et crée des gamètes diploïdes. La fusion de deux gamètes 2n crée un zygote 4n. Le second, l'allopolyploïdie, courante chez les plantes, correspond à l'hybridation interspécifique entre deux espèces et résulte en deux ensembles de chromosomes. Chez l'hybride F1, le nombre de chromosomes va doubler et créer une nouvelle espèce si l'hybride est fertile. Par exemple, les plantes du genre *Brassica* ont évolué par allopolyploïdie : je présenterai plus loin dans ce manuscrit mes travaux sur le génome du colza.

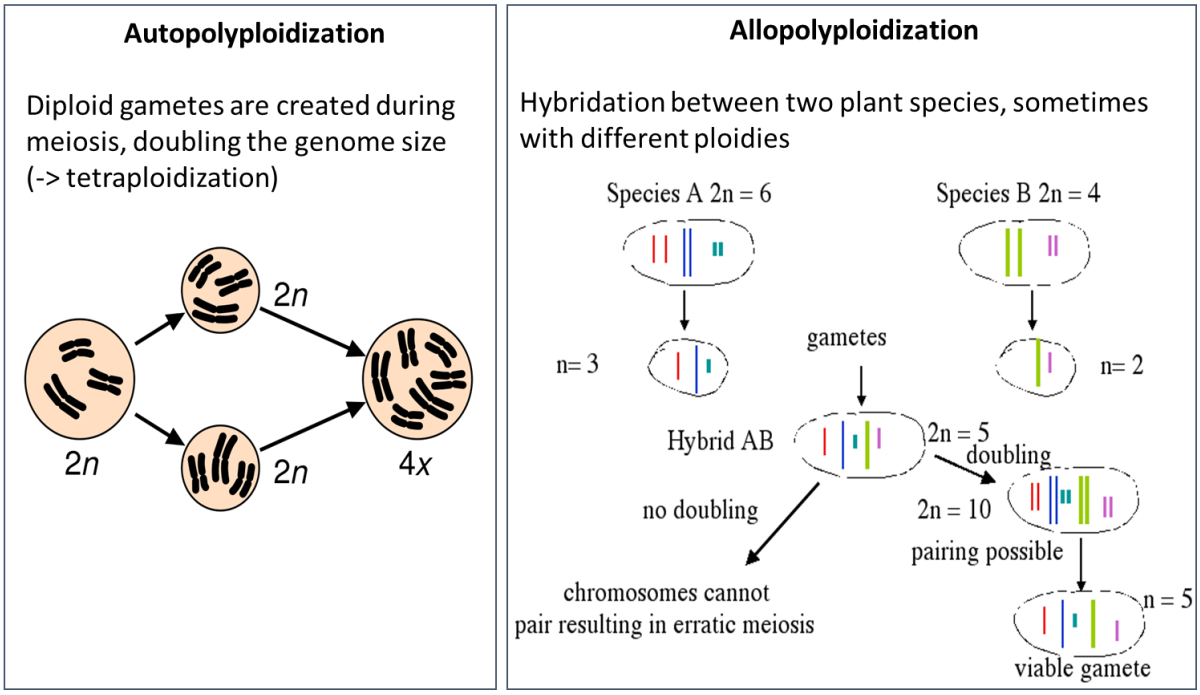

Figure 2 : mécanismes d'autopolyploïdie et d'allopolyploïdie

Même si les polyploïdes sont très communs dans la nature, les individus résultants sont confrontés à des obstacles comme l'augmentation des taux d'erreurs de ségrégation chromosomique et la compétition avec leurs progéniteurs diploïdes (Comai, 2005 ; Otto, 2007). Suite à un évènement d'allopolyploïdisation, deux génomes différents, en termes de contenu génique et d'éléments régulateurs, sont associés dans le même noyau, ce qui peut être à l'origine d'un important choc génomique et transcriptomique. L'expression des copies dupliquées peut ainsi évoluer suivant différents scénarios (Buggs, 2011 ; Edger, 2017), dépendant de mutations dans les sites de liaison des facteurs de transcription (Li, 2005 ; Casneuf, 2006). Enfin, des régulations post-translationnelles peuvent également survenir :

l'évolution rapide de motifs de phosphorylation peut permettre une régulation des gènes dupliqués (Amoutzias, 2010).

La plupart des duplicats sont rapidement perdus, avec une importance des effets de dosage visant à équilibrer la stoechiométrie des complexes (je reviendrai plus en détail sur la notion de « gene balance hypothesis » dans la section « Evolution des familles de gènes »). La perte de certaines copies de gènes pour le retour à un organisme diploïde est aussi nommée « gene fractionation » (Langham, 2004) (**Figure 3**).

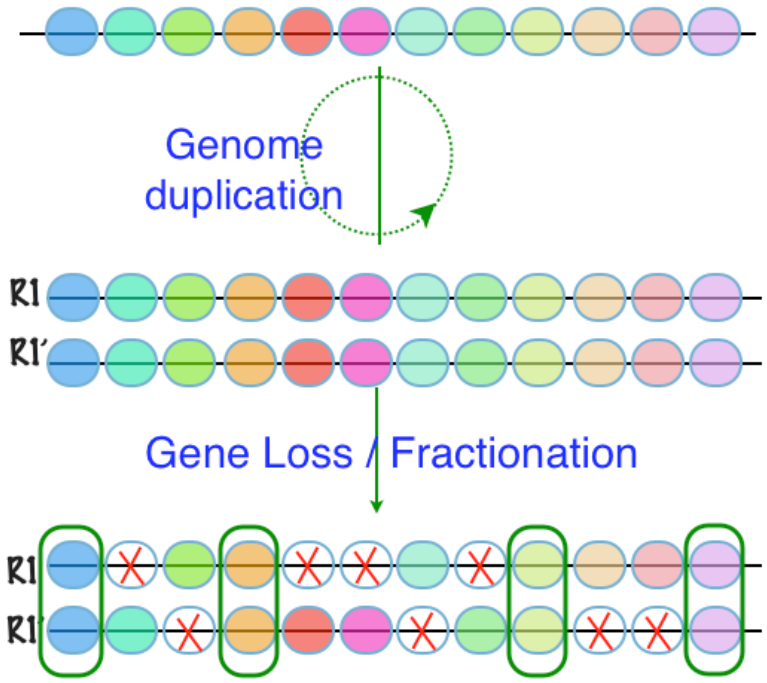

Figure 3 : illustration du phénomène de gene fractionation (perte de gènes après un événement de WGD), d'après CoGePedia (http://genomevolution.org/wiki/index.php)

Pour les raisons évoquées plus haut, il a été suggéré que la polyploïdie est en général une voie évolutive sans issue et que les espèces polyploïdes qui réussissent à s'établir sont plutôt l'exception (Arrigo and Barker, 2012 ; Mayrose, 2011). Cependant, la polyploïdie peut aussi avoir des avantages : en particulier, la redondance des gènes peut protéger les individus polyploïdes des effets délétères des mutations et leur conférer un potentiel adaptatif supérieur (Comai, 2005 ; Ohno, 2013). En outre, il s'agit d'une source de spéciation car les descendants, qui ont des nombres de chromosomes différents de leurs parents, ne peuvent plus se croiser avec des organismes non-polyploïdes. Cette diversification des espèces nécessite toutefois un certain laps de temps (quelques millions d'années après la WGD) : on parle de « Whole Genome Duplication Radiation lag-time model » (Schranz et al. 2012, Tank

et al. 2015). Il est probable que les polyploïdes qui ont réussi à s'établir ont échappé à la compétition avec leurs espèces progénitrices en migrant vers des niches écologiques différentes à la suite d'une forte sélection environnementale (Otto and Whitton, 2000 ; Visger, 2016), comme cela est proposé pour l'apparition de nouvelles espèces d'angiospermes par WGD suite à des périodes de grandes variations climatiques (Vanneste, 2014 ; Ren, 2018 ; Wu, 2020). L'intervention humaine sur des espèces cultivées (via des hybridations et des étapes de sélection) a pu également permettre à des polyploïdes de s'établir (Ballington, 2008) ou de se maintenir (Dubcovsky, 2007).

2/ Les duplications segmentales

Parmi les duplications segmentales, certaines surviennent localement (répétitions directes ou inverses), d'autres à distance. Les principaux mécanismes menant à la duplication d'un ou de quelques gènes sont les suivants (Zhang, 2003). Le premier mécanisme est le crossing-over inégal (« unequal crossing over ») : lors de la méiose, les chromosomes homologues s'apparient sur les régions semblables, et peuvent éventuellement s'échanger. Ainsi, une partie de l'ADN de l'un des deux chromosomes homologues est transférée sur l'autre. On a ainsi une perte de gènes sur un chromosome et une duplication de gènes sur l'autre si cette portion de chromosome porte un ou plusieurs gènes. On obtient alors une disposition des gènes dupliqués les uns à la suite des autres, en tandem (**Figure 4A**). Un mécanisme de glissement lors de la réplication (« replication slippage ») peut également conduire à des répétitions en tandem : il s'agit d'une erreur survenant lors de la réplication, lorsque l'ADN polymérase se dissocie de l'ADN en cours de copie et qu'elle se réassocie en alignant le brin répliqué sur une séquence homologue plus loin sur l'ADN, entrainant une variation du nombre de copies dans une répétition en tandem (Gadgil, 2017). Ce phénomène est plutôt observé pour de courtes séquences (couramment des microsatellites, répétitions en tandem avec une unité inférieure à une dizaine de paires de bases). Il peut également survenir lors de la réparation de cassures double-brin (Pâques, 1998). Un autre mécanisme est la rétrotransposition (**Figure 4B**). En effet, la rétrotransposase peut parfois agir sur des ARN messagers cellulaires : la transcription inverse génère alors de l'ADN à partir d'un transcrit, qui est inséré aléatoirement dans le génome. Ces rétrogènes sont facilement reconnaissables car ils n'ont pas d'introns, en tout cas dans un premier temps. Ils sont aussi généralement flanqués de répétitions directes.

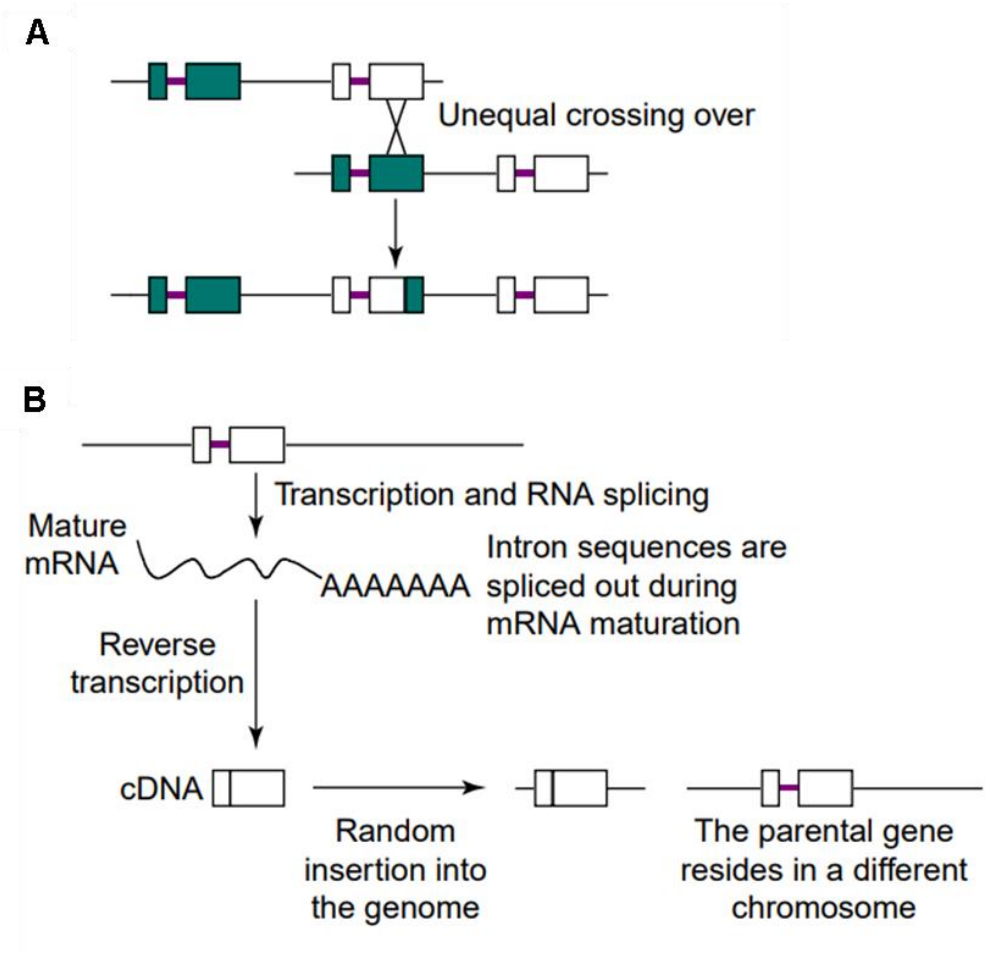

Figure 4 : Deux modèles de duplication de gènes, d'après Zhang (2003). **A.** Crossing over inégal. **B.** Transcription inverse.

3/ Devenir des gènes dupliqués

À la suite d'un évènement de duplication, une proportion non négligeable des gènes dupliqués est conservée. Cette conservation suggère l'existence d'un mécanisme de sélection naturelle qui compenserait la production de pseudogènes (pseudogénisation). En effet, à moins que la présence d'une quantité supplémentaire de protéines soit avantageuse, deux gènes avec des fonctions identiques ont peu de chances d'être maintenus dans le génome. L'un des deux gènes dupliqués, étant redondant avec l'autre copie, peut perdre sa fonction et/ou ne plus être exprimé, Il se transforme ainsi en pseudogène, et au cours de l'évolution finira par être délété ou tellement dégénéré qu'il ne sera plus reconnaissable (Lynch, 2000 ; Zhang, 2003).

Outre la pseudogénisation, on distingue 3 devenirs possibles pour les gènes dupliqués **(Figure 5).** Le premier, nommé subfonctionnalisation, s'applique aux gènes qui

avaient plusieurs « fonctions » ancestrales (ou expressions dans des tissus/conditions différents). Dans ce cas, l'un des gènes dupliqués peut perdre l'une des fonctions, et l'autre gène dupliqué peut perdre la fonction complémentaire. Ainsi, l'organisme est capable d'assurer les deux fonctions, qui sont alors exercées par deux protéines différentes, ou exprimées dans des tissus/conditions différents (Force, 1999 ; Zhang, 2003).

Le second devenir possible, nommé néofonctionnalisation, est à l'origine de l'apparition de nouvelles fonctions. Comme la fonction sélectionnée sur le gène originel est codée par deux copies de gènes, la pression de sélection peut se relâcher sur l'une de ces copies. Il peut ainsi y avoir apparition d'innovations évolutives, et d'une nouvelle fonction. Il s'agit souvent d'une nouvelle fonction liée à la fonction initiale, comme c'est le cas par exemple pour les protéines « opsines » sensibles à la couleur rouge ou verte chez les humanoïdes, qui ont été générées suite à une duplication génique (Yokoyama, 1989 ; Zhang, 2003). Les enzymes N-méthyl transférases, amplifiées en tandem puis diversifiées indépendamment dans les génomes du caféier, du théier et du cacaotier, pour créer les voies de biosynthèse de la caféine, de la théine et de la théobromine, respectivement, ont également évolué par un mécanisme de ce type : je présenterai ce résultat dans la suite de ce document (Denoeud, 2014). Ces deux évolutions, subfonctionnalisation et néofonctionnalisation, qui permettent la divergence de deux copies de gènes, sont vraisemblablement causées par une relaxation de la pression de sélection, mais aussi par une sélection positive des mutations (Zhang, 2003).

Enfin, la dernière voie est celle d'une sélection en faveur d'une (forte) augmentation du nombre de copies de gènes. Il y a deux grands cas de figure où une telle amplification peut procurer un avantage sélectif. Le premier concerne des gènes dont la fonction nécessite d'énormes quantités d'un produit invariable. Il s'agit de gènes en général très exprimés, comme les protéines ribosomales et les histones, et dont les copies restent très fortement conservées (Piontkivska, 2002 ; Rooney, 2002 ; Rooney, 2004 ; Dharia, 2015). Cette conservation est maintenue par une très forte pression de sélection (sélection purifiante) empêchant des variations de la séquence protéique, j'y reviendrai dans la prochaine section (**Figure 5B**). Le second cas de figure concerne les gènes dont la fonction requiert un grand degré de diversité : c'est le cas pour les très grandes familles de récepteurs olfactifs (Hughes, 2018, Niimura, 2003) ou de gènes du système immunitaire et de résistance aux maladies dans toutes les branches du vivant (Nei, 2005 (review) ; Ota, 1994 ; Nei, 1997 ; Hao, 2004 ; Hamada, 2013 ; Plomion, 2018 ; Andersen, 2020) (**Figure 5C**). La sélection opère alors dans le sens d'une

diversification du répertoire (sélection positive de nouvelles structures protéiques), mais la fonction générale reste similaire, on ne peut donc pas parler ici de néofonctionnalisation. Je présenterai par la suite mes travaux sur l'identification d'amplifications en tandem massives pour les gènes de récepteurs du système immunitaire inné dans les génomes de coraux.

Figure 5 : Devenir des gènes dupliqués. **A.** Trois scénarios classiquement proposés (Source: Smedlib, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons). **B** et **C** : Cas des gènes dont l'amplification présente un avantage évolutif, selon deux scénarios : sélection contre la variation (B) et sélection contre l'homogénéisation (C).

4/ Evolution des familles multigéniques

Les familles multigéniques peuvent provenir de duplications complètes de génomes ou de duplications segmentales, et probablement des deux à la fois. La perte des gènes dupliqués pour un retour à l'état monoploïde peut être soumise à des contraintes. En particulier, le modèle nommé « Gene balance hypothesis » prédit que les protéines qui font partie de complexes multiprotéiques ou de réseaux de régulation nécessitent de conserver toutes le même nombre de copies afin que l'équilibre stœchiométrique de leurs interactions ne soit pas altéré (Papp, 2003 ; Birchler, 2005 ; Freeling, 2006 ; Aury, 2006). Il est suggéré que les pressions de sélection liées à l'effet de dosage sont différentes entre les deux cas de figure (Freeling, 2006 ; Freeling, 2009). Selon cette hypothèse, suite à un événement de tétraploïdie, les gènes en complexes (ou « connectés ») sont maintenus en deux copies, car la duplication concerne l'ensemble des gènes du complexe (tous passent du simple au double sans faire changer leurs proportions relatives). A l'inverse, dans le cas de duplication segmentale ne concernant qu'un seul membre du complexe, le maintien de la répétition serait plutôt contre-sélectionné car il nuirait à la stœchiométrie du complexe. Cependant, il peut être difficile, après des millions d'années d'évolution et de remaniements chromosomiques, de distinguer les paralogues provenant de ces différents mécanismes. Cela est d'autant plus compliqué par le fait que les séquences des gènes dans les familles multigéniques sont parfois très homogènes, pouvant rendre l'identification des relations d'orthologie et de paralogie difficiles.

L'observation que les paralogues dérivant de duplications anciennes sont parfois plus conservés entre eux (au sein d'une espèce) qu'avec leurs orthologues (dans une espèce partageant un ancêtre commun déjà dupliqué) (Brown, 1972 ; Nei, 2005) **(Figure 6)**, a conduit à proposer un modèle d'évolution entre différentes copies de gènes dupliquées concertée (Ingram, 1961 ; Hood, 1975 ; Zimmer, 1980 ; review : Eirín-López, 2012). Cette homogénéisation se produirait par un mécanisme de conversion génique ou bien de crossing-over inégal, qui ferait pour sa part varier le nombre de copies. Ce phénomène pourrait se produire d'autant plus facilement que les gènes seraient proches sur le génome, donc les gènes répétés en tandem y seraient particulièrement sujets. On pourrait alors supposer que la translocation de gènes dans des régions plus distantes du génome permettrait d'échapper à cette homogénéisation et donc de fournir des possibilités d'évolution vers de nouvelles fonctions, pour les copies ainsi « libérées » de la conversion génique. Cependant, d'autres phénomènes peuvent expliquer la présence de paralogues très conservés (**Figure 6B,C**).

Effectivement, les analyses menées par Nei et ses collègues, sur de nombreuses familles de gènes dans diverses espèces ont montré qu'en fait, dans les familles dont les membres sont très conservés, le mécanisme en cause est plutôt la pression de sélection purifiante (« purifying selection », **Figure 5B**, **Figure 6C**), qui maintient, parfois drastiquement, la séquence des protéines (Hughes, 1989, Nei, 2000 **;** Nei, 2005 ; Eirín-López, 2012).

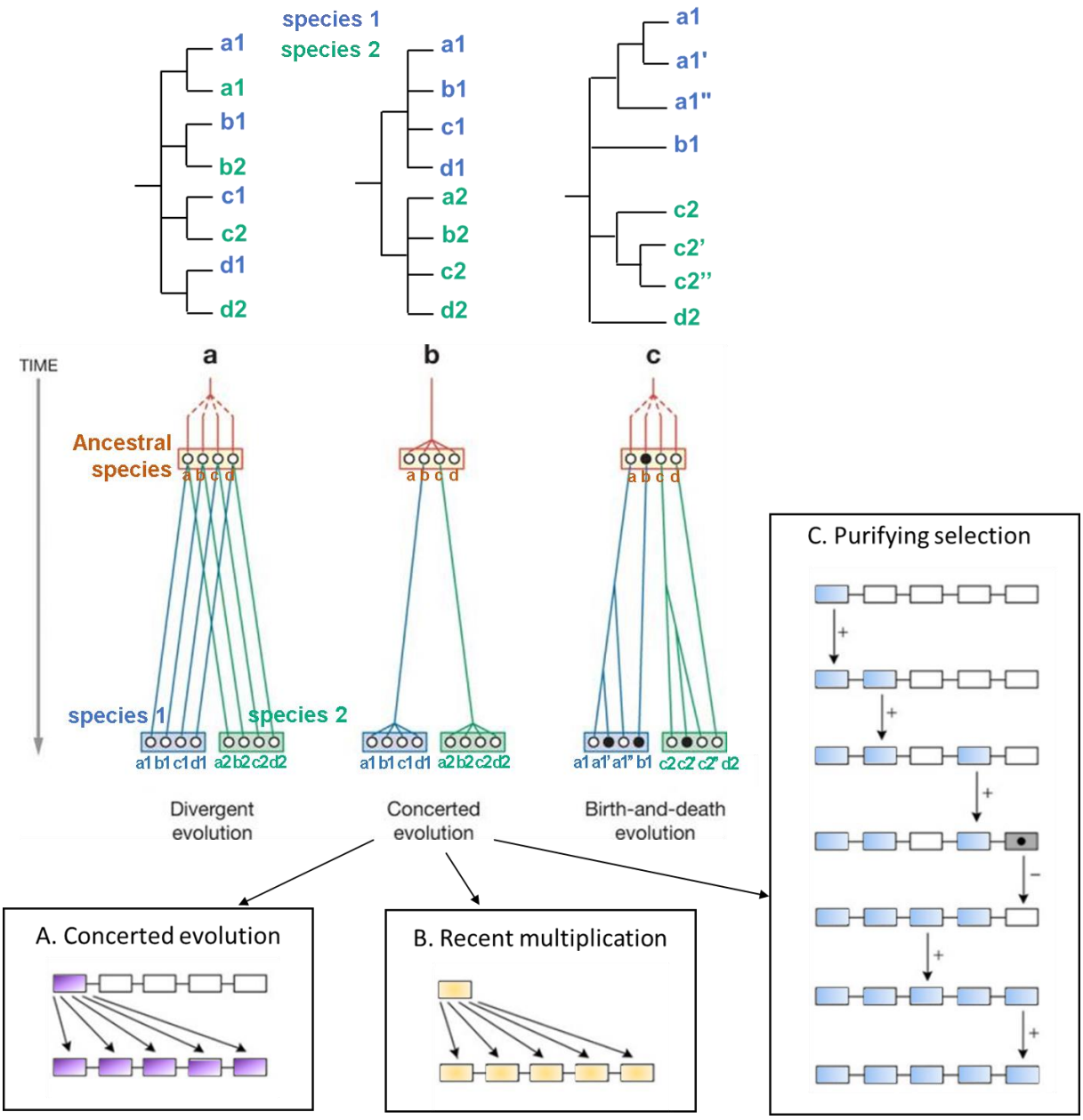

Figure 6 : Différents mécanismes évolutifs conduisant à différentes relations d'homologie entre gènes paralogues et orthologues (d'après Nei, 2005). Le cas b qui pourrait correspondre à l'évolution concertée (b.A) peut aussi s'expliquer par des duplications récentes (b.B) ou une pression de sélection purifiante (b.C).

On peut citer comme exemple marquant les histones (Piontkivska, 2002 ; Rooney, 2002), qui sont extrêmement conservées au niveau protéique (100% d'identité pour l'histone H4 de plantes), mais dont la troisième position du codon varie fortement d'une copie à l'autre (mutations synonymes). Je présenterai dans la section « Disussion et perspectives » les analyses que j'ai menées sur les histones de plantes dans le cadre d'un projet de séquençage de plantes alpines.

Le mécanisme d'évolution des familles multigéniques actuellement le plus reconnu est celui proposé par l'équipe de Nei en 1997 et nommé « Birth and Death evolution » (Eirín-López, 2012). Dans ce modèle, des gènes sont gagnés et perdus au cours de l'évolution à des taux variables selon les espèces, les périodes et les gènes. Ces taux dépendent probablement des avantages évolutifs conférés par les gènes et des nécessités d'adaptation des espèces aux contraintes environnementales. Des gènes sont gagnés par duplication (selon les modèles décrits en section 2) et d'autres gènes sont perdus par pseudogénisation puis délétions. Ce processus conduit à des familles multigéniques de taille variable selon les taxons, constituées d'un mélange de pseudogènes et de gènes fonctionnels montrant divers degrés de similitude entre eux. Afin de comprendre les forces évolutives gouvernant ce processus de gain et perte de gènes, il est nécessaire de pouvoir compter de façon fiable le nombre de copies des gènes dans les familles multigéniques. Je montrerai dans la suite de ce manuscrit comment des assemblages longues lectures (de bonne continuité) et une annotation tenant compte des répétitions en tandem de gènes, ont permis de mettre en évidence une telle dynamique de gains et pertes de gènes dupliqués en tandem chez les coraux.

Partie II. Retour sur mon parcours scientifique

1/ Thèse à l'Institut de Génétique et Microbiologie, Orsay (2000-2003)

Lors de ma thèse (et précédemment, pendant mon stage de maîtrise en informatique appliquée et mon stage de DEA) dans le laboratoire de Gilles Vergnaud à l'Institut de Génétique et Microbiologie d'Orsay, j'ai d'abord développé une base de données de répétitions en tandem dans les génomes séquencés, que j'ai ensuite exploitée afin d'identifier des minisatellites polymorphes (Denoeud et al, 2004). Il est important de souligner que dès mon stage de DEA, Gilles Vergnaud m'a chargée de rédiger une review pour Genome Research sur les minisatellites hypermutables, ce que j'ai été capable de mener à bien de façon autonome (Vergnaud & Denoeud, 2000). Pendant ma thèse, j'avais une double compétence en bioinformatique et biologie moléculaire : j'ai partagé mon temps entre l'ordinateur et la paillasse, où j'ai utilisé des techniques de biologie moléculaire (PCR, électrophorèse, Southern Blot) pour tester le polymorphisme des minisatellites identifiés, et mettre en évidence des minisatellites hypermutables humains. J'ai également développé des méthodes statistiques pour identifier des critères prédictifs des séquences des minisatellites sur leur polymorphisme en collaboration avec Gary Benson (Mount Sinaï School of Medicine, New York) (Denoeud et al, 2003).

Cette thèse ayant été effectuée sur le campus de l'université d'Orsay, j'ai été amenée à participer à la création et à l'encadrement de travaux pratiques de biologie moléculaire sur l'identification de souches bactériennes : j'ai été en charge de la partie analyse bioinformatique (utilisation des profils de génotypage de minisatellites de *Bacillus anthracis* pour identifier des souches bactériennes, calcul de matrices de distances et génération d'arbres phylogénétiques). J'ai aussi encadré deux stagiaires de licence, que j'ai formés à la recherche de minisatellites dans les bases de données et au génotypage par PCR.

Enfin, j'ai eu l'occasion de rédiger plusieurs articles scientifiques en toute autonomie, notamment dès mon stage de DEA (Vergnaud & Denoeud, 2000), ce qui m'a permis d'acquérir des compétences rédactionnelles très utiles pour la suite de ma carrière.

2/ Postdoctorat à l'IMIM/GRIB, Barcelone (2004-2006)

Lors de mon post-doctorat (à l'IMIM, Barcelone dans le laboratoire de Roderic Guigó), dans le cadre du projet ENCODE (ENCyclopedia Of DNA Elements, NIH), j'ai étudié en particulier le paysage transcriptionnel dans les régions ciblées par le projet pilote ENCODE, représentant 1% du génome humain. Le projet ENCODE était mené par un consortium international de grande envergure et mon implication dans ce projet a été l'occasion de nombreuses collaborations internationales (entre-autres avec les laboratoires de Tom Gingeras, chez Affymetrix, Alexandre Reymond à l'Université de Lausanne, Jennifer Harrow au Sanger Center, Ewan Birney à l'EBI) (Harrow J*, Denoeud F* et al, 2006). Etant la responsable de l'analyse dans le groupe de Roderic Guigó, j'ai participé à de nombreux workshops et « data fairs » ce qui m'a familiarisée avec la gestion de projet au sein d'un grand consortium (contraintes variées pour atteindre les dates limites de production et traitement de données, organisation et mise en œuvre de l'analyse de ces données, organisation et mise en œuvre de la rédaction des articles…). J'ai également encadré un ingénieur chargé du support technique sur ce projet.

Mon expérience post-doctorale m'a permis d'acquérir des connaissances variées sur les analyses génomiques et post-génomiques à grande échelle, tant dans mon domaine de prédilection (structure et expression des gènes) que dans d'autres domaines tels que la structure de la chromatine, les sites de fixation de facteurs de transcription (Chip-Seq), l'identification de séquences conservées dans la phylogénie, la détection de régions sous sélection positive.

Mon implication dans le groupe d'analyse « ENCODE Genes and Transcripts » a permis de mettre en évidence de nombreux transcrits avec des exons étendant des gènes en 5', les reliant parfois à un gène amont (Denoeud et al, 2007). J'ai été en charge des analyses bioinformatiques visant à traiter les résultats des expériences de 5'RACE, afin d'identifier les exons 5' candidats. Nous avons ainsi pu valider par RT-PCR la présence de transcrits chimères entre deux gènes voisins (**Figure 7**). Leur fonction reste toutefois à élucider.

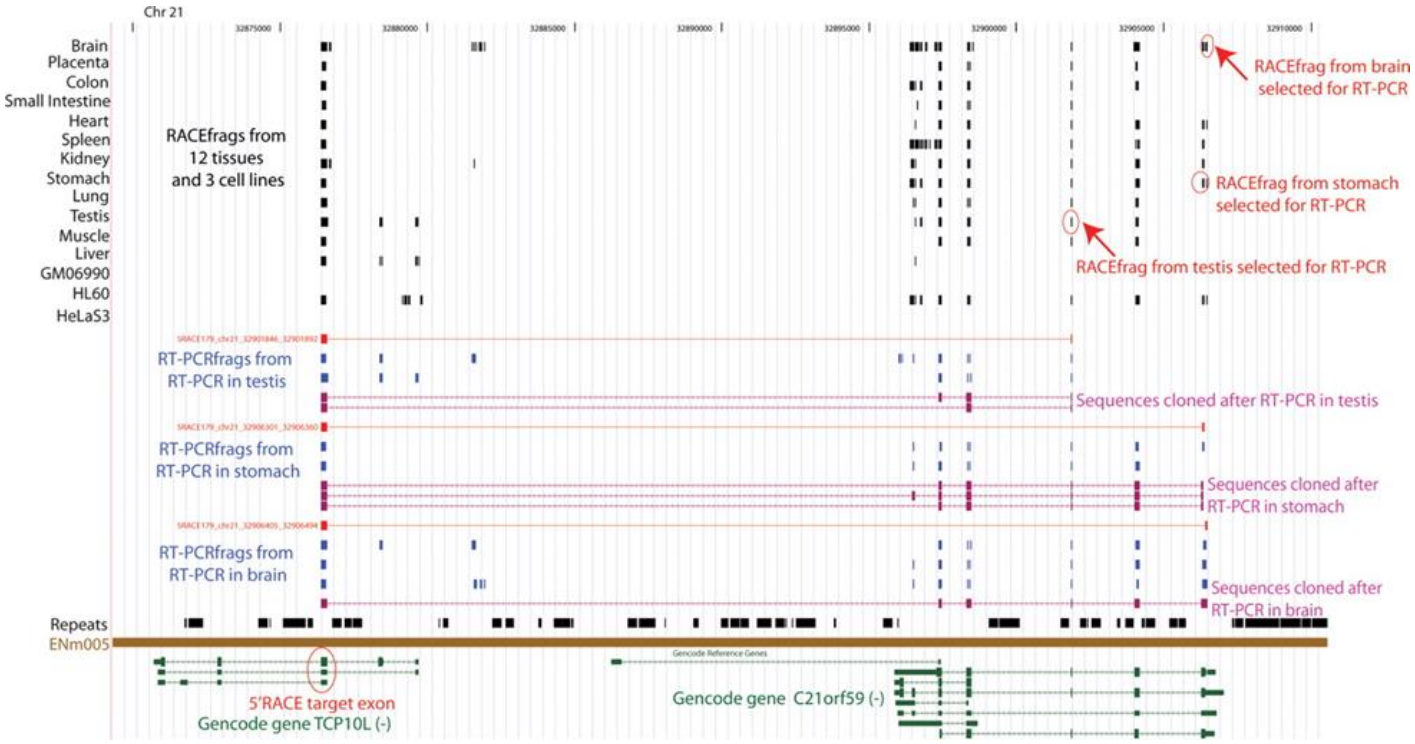

Figure 7 : Example of a transcription-induced chimera between C21orf59 and TCP10L. The results of a 5' RACE/tiling array analysis of the HSA21 *TCP10L*gene are presented. The GENCODE-annotated transcripts are shown (green, at the *bottom*), the index exon where the primer used for the 5' RACE maps is indicated. RACEfrags-positive regions obtained upon hybridization of the tiling array by the RACE reactions performed in 12 human tissues and three cell lines are shown (black boxes, *upper*part). Red boxes joined by thin red lines depict connectivity between index exons and RACEfrags selected to be independently verified by RT-PCR. The corresponding RACEfrags are highlighted in the *upper* part of the panel. The hybridization of these RT-PCR reactions to the same tiling arrays allowed us to identify RT-PCRfrags (blue boxes, see text for details). The cloning and sequencing of the RT-PCR reactions amplimers revealed the exon composition and chimeric nature of transcripts containing the targeted RACEfrags (purple transcripts). **(Denoeud et al, 2007)**

(NB : je n'ai pas traduit cette légende car elle contient trop de jargon impossible à traduire en Français)

3/ Expérience de chercheur « junior » au Genoscope (2006-2009)

Lorsque je suis entrée au Genoscope, ce groupement d'intérêt public n'était pas encore intégré au CEA (Commissariat à l'Energie Atomique). J'y ai fait mes débuts dans l'équipe de bioinformatique, dirigée à l'époque par François Artiguenave. Par la suite, après l'intégration du Genoscope au CEA en 2008, cette équipe est devenue le LABIS (Laboratoire d'Analyse Bioinformatique des Séquences), dirigée par Patrick Wincker. Ce laboratoire a par la suite été renommé LAGE –laboratoire d'analyses génomiques des eucaryotes- ce qui illustre sa

thématique de recherche portant sur la génomique comparative des eucaryotes visant à l'élucidation des relations entre la structure génomique et les traits fonctionnels.

Dès mon arrivée au Genoscope, j'ai été en charge de l'analyse de divers génomes eucaryotes, et du développement des méthodologies bioinformatiques nécessaires. J'ai été recrutée pour mener l'étude du génome du tunicier *Oikopleura dioica* en collaboration avec l'équipe de Daniel Chourrout au SARS, Bergen, Norvège. J'ai tout d'abord élaboré un pipeline permettant d'identifier les régions alléliques dans cet assemblage hétérozygote. J'ai ensuite développé un outil pour comparer les positions d'introns entre différentes espèces, afin d'identifier les introns ancestraux et récemment acquis. Il s'avère qu'*Oikopleura* a un taux extrêmement élevé de turnover d'introns, c'est-à-dire que la fréquence d'introns perdus et gagnés (à de nouvelles positions dans les gènes) est très élevée. Suite à des recherches bibliographiques poussées sur l'évolution des introns, j'ai pu mettre en évidence des mécanismes originaux de gain d'introns, par insertion d'éléments transposables et « reverse splicing » (ou « intron transposition »), mécanisme consistant à ré-insérer un intron épissé dans une molécule d'ARN messager, qui est par la suite reverse-transcrite puis intégrée dans le génome par recombinaison homologue. Ce phénomène, décrit de façon théorique, n'avait jamais été observé auparavant (**Figure 8B-C**, Denoeud et al, 2010 - voir en particulier la section du « supplementary material » portant sur les introns: https://www.science.org/doi/suppl/10.1126/science.1194167/suppl_file/denoeud.som.revision.1.pdf, section 10 p16-29-).

Oikopleura dioica a par ailleurs une proportion très élevée d'introns non canoniques (**Figure 8A**), qui ont des caractéristiques particulières en terme de taille, d'âge d'apparition, et de phase d'insertion. J'ai émis l'hypothèse que le spliceosome majeur (U1/U2) d'*Oikopleura dioica* est permissif, et épisse donc les introns non canoniques, ce qui permet de tolérer des insertions massives d'introns dans cette espèce soumise à un cycle de vie très court et une évolution très rapide. Pour publier ces travaux, j'ai mené l'ensemble des analyses sur l'évolution des introns de façon autonome, ce qui a nécessité des recherches bibliographiques poussées et m'a permis d'acquérir une grande expertise sur ls mécanismes d'épissage et d'évolution des introns. J'avais rédigé un article complet sur cette thématique, que je pensais soumettre en tant que compagnon, mais qui a finalement été intégré dans l'article principal décrivant le génome.

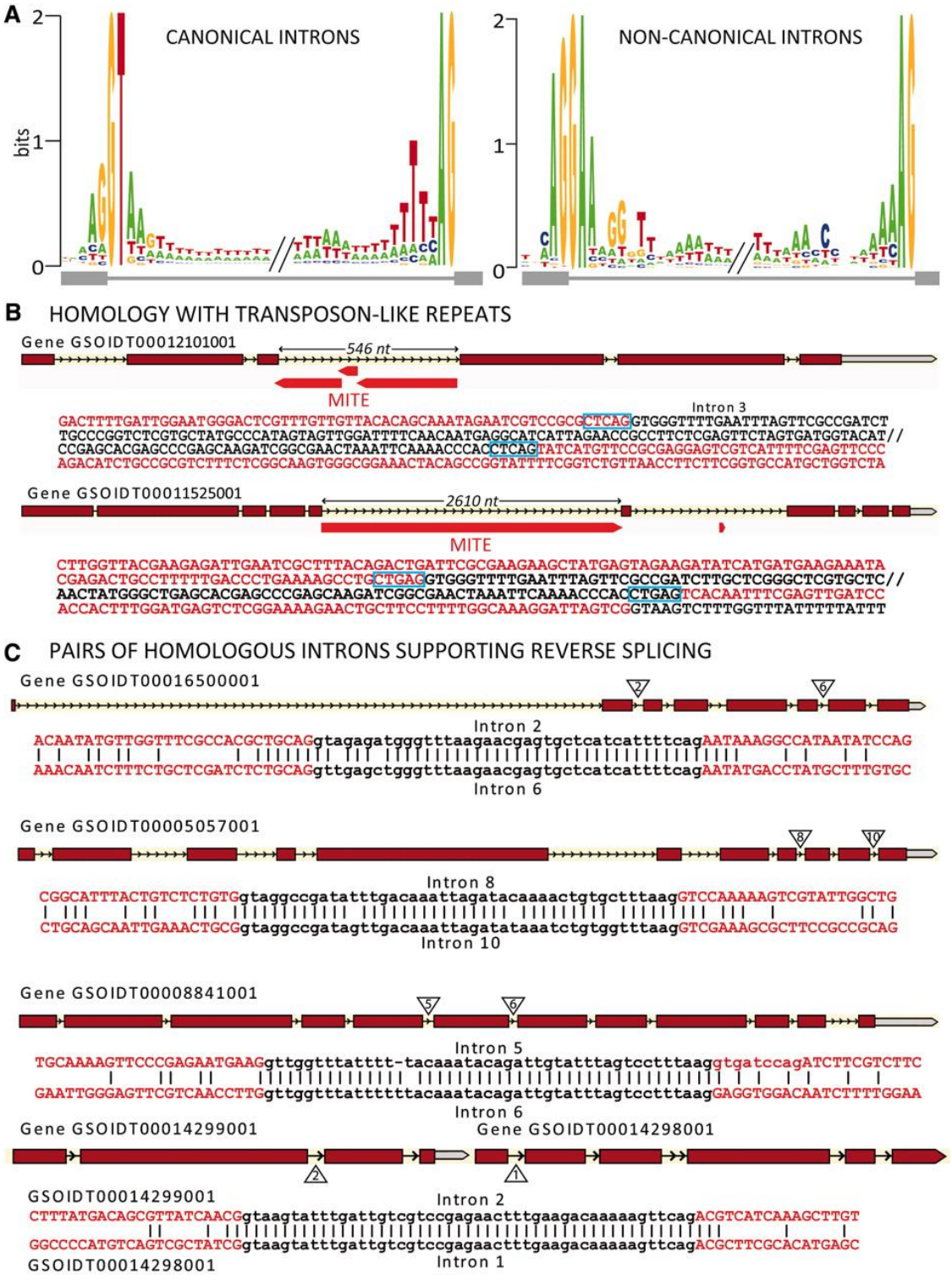

Figure 8. **Structure des introns et scenarios de gains d'introns chez O. dioica.** (A) Logos des principales catégories d'introns. (B) Gain par insertion de transposon : les site d'insertion dupliqués encadrés en bleu permettent à des insertions de type MITE (miniature inverted repeat transposable element) d'être épissées exactement. (C) Gain par reverse splicing (épissage inverse) : quatre paires d'introns homologues (en noir) et leurs environnements exoniques (en rouge) **(Denoeud et al, 2010).**

J'ai également développé la première version d'une nouvelle méthode d'annotation des génomes à partir de lectures RNA-Seq, sans génome de référence, qui a depuis évolué pour devenir l'outil (« Gmove ») utilisé pour l'annotation des génomes eucaryotes au Genoscope (**Figure 9**, Denoeud et al, 2008).

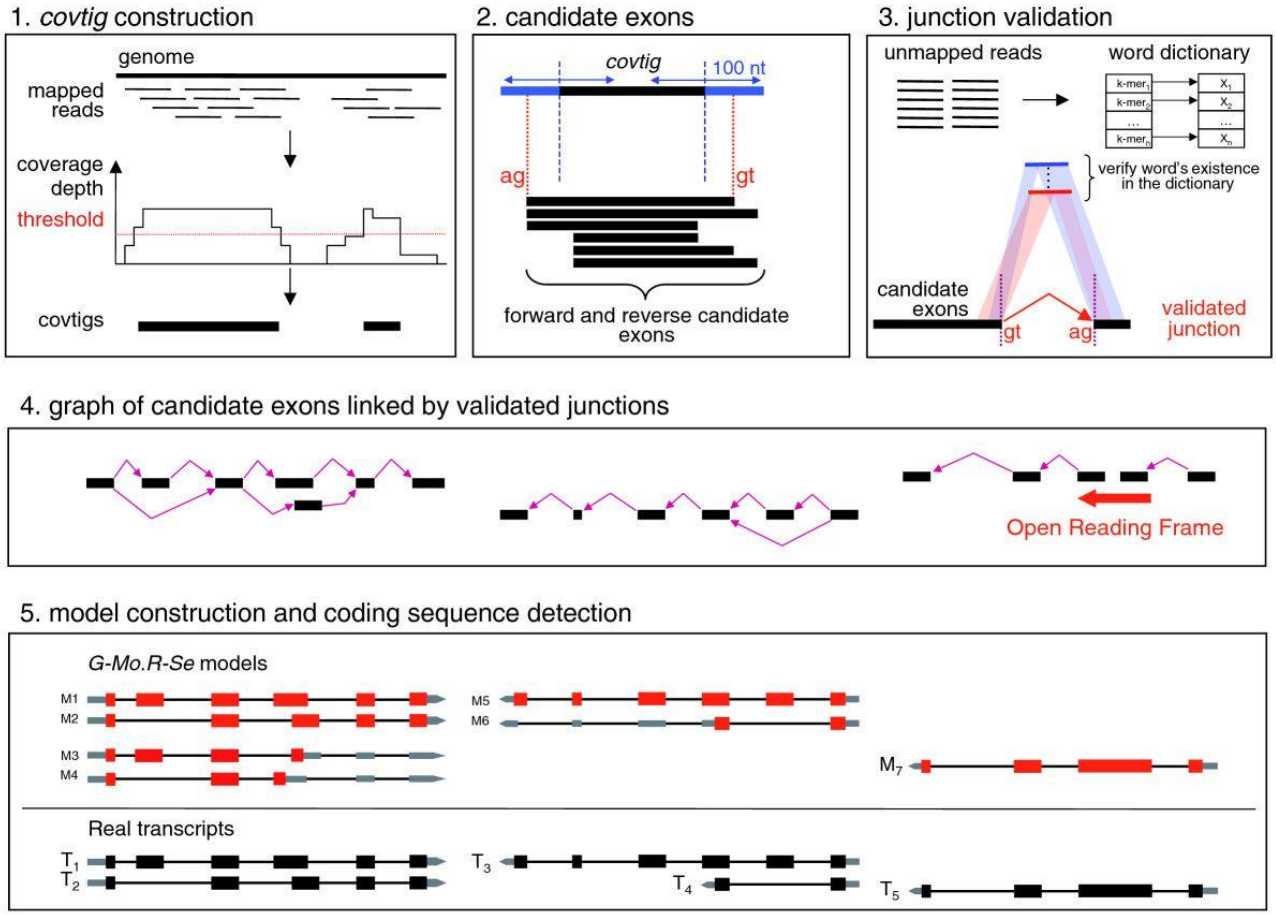

Figure 9. Méthode *G-Mo.R-Se* pour construire des modèles de gènes à partir de lectures RNA-Seq. Les cinq cadres noirs montrent les 5 étapes de l'approche. L'étape 1 (construction des covtigs) construit des covtigs (coverage contigs) à partir des positions où les lectures sont alignées avec une profondeur supérieur à un certain seuil. L'étape 2 (exons candidats) est la définition d'une liste d'exons candidats orientés dérivés de chaque covtig. Les sites d'épissage sont recherchés 100 nucléotides en amont et en aval de chaque covtig, ce qui permet l'orientation de l'exon candidat sur le brin sens ou antisens. L'étape 3 (validation des jonctions) consiste en la validation des jonctions entre des exons candidats en utilisant le dictionnaire de mots construits à partir des lectures non alignées. Pendant l'étape 4 (construction du graphe des exons candidats liés par des jonctions validées), un graphe est créé où les nœuds sont les exons candidats (en noir) et les arrêtes orientées sont les jonctions validées (flèches violettes). Dans l'étape 5 (construction des modèles et détection des séquences codantes), le graphe est parcouru et tous les chemins possibles sont extraits. Chaque chemin représente un transcrit prédit, et une CDS est recherchée pour chaque transcrit **(Denoeud et al, 2008).**

J'ai par la suite participé à l'analyse des génomes de divers eucaryotes séquencés au Genoscope : *Blastocystis* (parasite intestinal humain) (Denoeud et al, 2011), *Tuber melanosporum* (truffe noire du Périgord) (Martin et al, 2010), *Chondrus crispus (*algue rouge) (Collen et al, 2013), *Phytomonas* (trypanosomes de plantes) (Porcel et al, 2014). Pour ces génomes, j'ai expertisé les annotations automatiques et j'ai effectué diverses analyses comparatives. En particulier, dans le génome de *Blastocystis* j'ai mis en évidence des événements de transfert horizontal de gènes, provenant de bactéries mais également d'eucaryotes. J'ai également mené des analyses de synténie, ainsi que des analyses d'orthologie et de paralogie (par exemple entre les deux souches séquencées *Phytomonas* mais également entre les génomes étudiés et d'autres espèces déjà séquencées). J'ai rédigé les paragraphes décrivant ces analyses dans les articles décrivant ces génomes. Je me suis ensuite spécialisée dans l'analyse de génomes de plantes.

4/ Prise de responsabilités dans trois grands projets (2010-2014)

J'ai en effet été en charge de l'analyse de trois génomes de plantes au sein de grands consortiums internationaux : bananier, colza, caféier. Pour ces trois grands projets, j'étais la représentante du Genoscope pour le pilotage des analyses (auxquelles j'ai très largement pris part), ainsi que dans les comités de rédaction des articles scientifiques, qui ont mené à la publication de trois articles dans des grandes revues, dont j'étais le 1er auteur ou co-auteur.

a/ Analyse du génome du bananier, *Musa acuminata*, paléopolyploïde

Tout d'abord, le séquençage du génome du bananier (D'hont*, Denoeud* et al, 2012), première plante de la classe des monocotylédones séquencée en dehors de la famille des Poaceae (céréales) a servi de référence pour l'étude de l'évolution des génomes. Nous avons mis en évidence trois étapes de duplication complète du génome (WGD) spécifiques de la lignée Musa. Certaines copies ont persisté et permis l'émergence de nouvelles fonctions biologiques, comme certains facteurs de transcription impliqués dans la maturation des fruits. J'ai été impliquée en particulier dans la caractérisation des événements de duplication complète du génome par la mise en évidence des régions présentes en trois copies dans le génome du bananier (**Figure 10**).

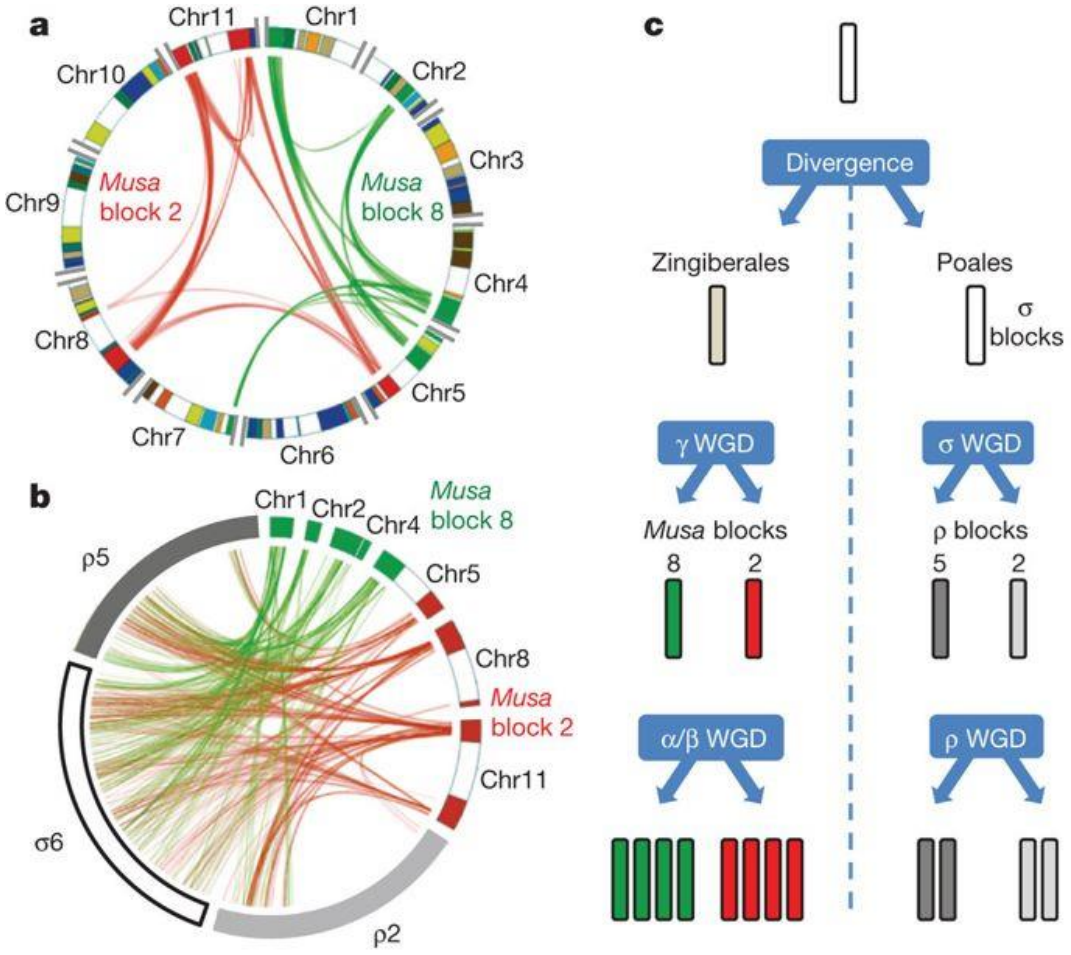

Figure 10. Evènements de duplications complètes de génomes chez *Musa acuminata*. A Relations de paralogie entre les segments de chromosomes correspondant aux blocs ancestraux *Musa* α/β 2 (en rouge) et 8 (en vert). Les 12 blocs *Musa* α/β ancestraux sont représentés par différentes couleurs sur le cercle **b**, Relations d'orthologie entre les blocs ancestraux *Musa* α/β 2 et 8 et les blocs ancestraux de riz ρ2, ρ5 et σ6. Nous n'observons pas de relation 1:1 entre un bloc *Musa* α/β et un bloc ρ, ce qui suggère que les duplications γ et σ sont deux évènements distincts **c**, Représentation de l'événement de WGD déduit **(D'hont et al, 2012).**

b/ Analyse du génome du colza, *Brassica napus*, polyploïde récent

Le colza, *Brassica napus,* est le premier polyploïde récent dont le génome a été séquencé (Chalhoub*, Denoeud* et al, Science 2014). Le colza provient en effet d'un événement d'hybridation entre un chou (*Brassica oleracea*) et une navette (*Brassica rapa*). Cette lignée a été sujette à des cycles successifs de polyploïdisation (duplications complètes de génome),

faisant de son génome un des plus hautement dupliqués chez les plantes à fleurs (72 copies) (**Figure 11**).

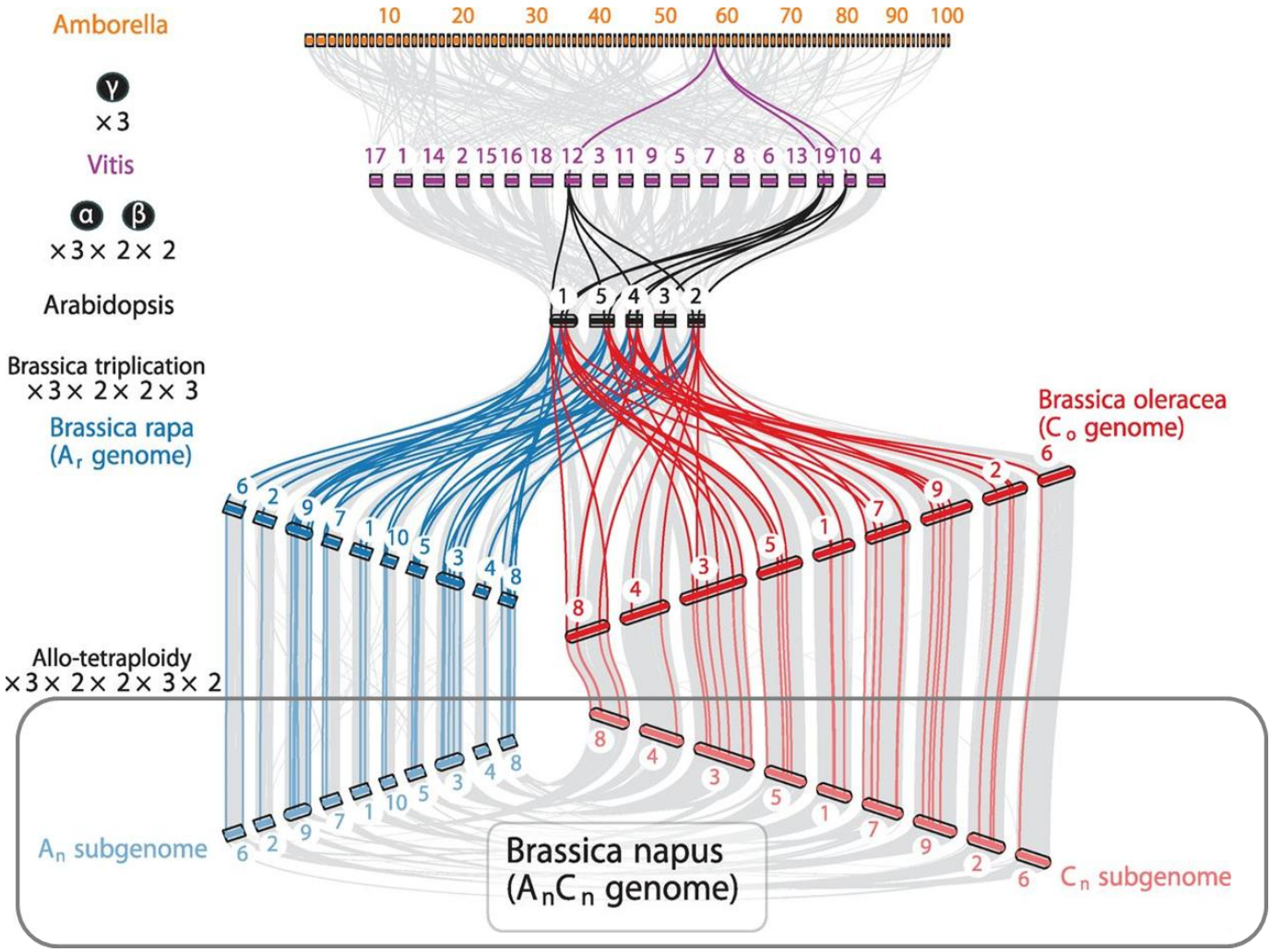

Figure 11. Duplications de génomes récurrentes chez B. napus. Les alignments génomiques entre l'angiosperme basal *Amborella trichopoda*, l'eudicot basal *Vitis vinifera*, l'espèce modèle crucifère *Arabidopsis thaliana*, ainsi que *B. rapa* et *B. oleracea* et *B. napus* sont présentés. On s'attend à ce qu'une région ancestrale d'*Amborella* s'aligne sur jusqu'à 72 régions de *B. napus* (pour la région présentée dans cet exemple 69 ont été détectées). Les régions grisées représentent blocs synthéniques conservés de plus de 10 paires de gènes **(Denoeud et al, 2014).**

La comparaison du génome du colza avec celui de ses deux parents (chou et navette) a permis de montrer que la grande majorité des 101 000 gènes du colza existent en deux copies qui sont toutes deux exprimées et participent donc conjointement à leur fonction. Des échanges de gènes et d'ADN entre les deux sous-génomes parentaux du colza ont été mis en évidence. Lors de ces échanges homéologues, la séquence d'un sous-génome parental est remplacée par celle de l'autre sous-génome. J'ai développé une méthode de détection de ces

échanges, en utilisant la profondeur d'alignement de lectures de différents cultivars de colza sur les génomes parentaux A et C. J'ai ainsi pu montrer que ces échanges se produisent davantage sur les paires de chromosomes les plus synténiques (A01/C01 et A02/C02) et sont également plus fréquents dans le sens C vers A (délétion du génome « C » au profit du génome « A », c'est-à-dire qu'on trouve une profondeur ~double de lectures de colza mappant sur le génome parental A et une profondeur ~nulle sur le génome C). En particulier, j'ai identifié une région d'échange homéologue contenant un gène impliqué dans la biosynthèse des glucosinolates partiellement délété (inactif), copié du génome A au génome C, ce qui peut être mis en relation avec les processus de sélection effectués sur les colzas cultivés, dont on cherche à limiter la teneur en glucosinolates (**Figure 12**).

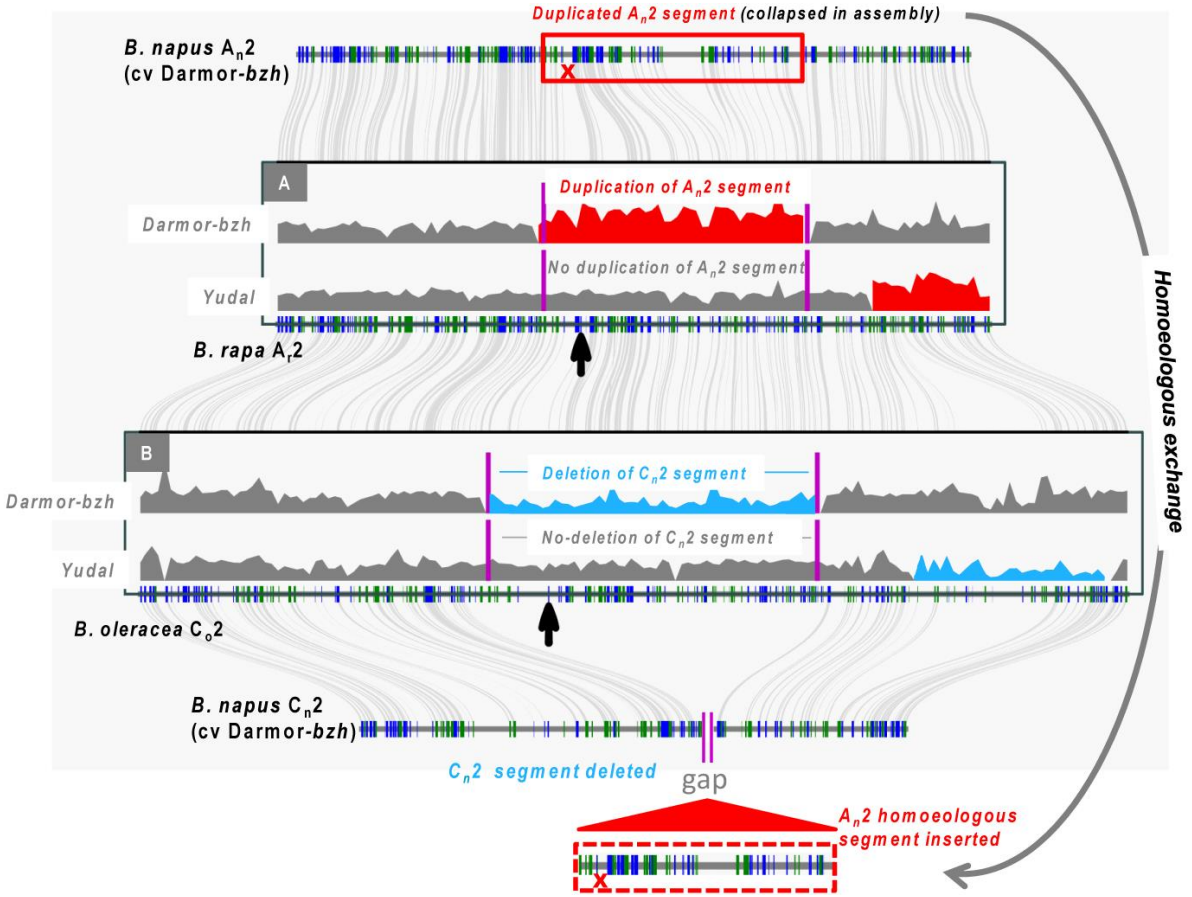

Figure 12 : Echange homéologue dans une région contenant un gène impliqué dans la biosynhtèse des glucosinolates (sa position est représentée par la croix rouge), survenu chez la variété de colza Darmor-bzh mais pas chez la variété Yudal (qui présente un taux de glucosinolates supérieur).

c/ Analyse du génome du caféier, *Cophea canofora*, révélant des duplications en tandem dans les gènes de biosynthèse de la caféine

Enfin, il était important de séquencer le génome du caféier car cette plante constitue la première richesse de nombreux pays tropicaux (Denoeud et al, 2014). L'espèce qui a été séquencée au Genoscope est *Coffea canephora*, alias robusta : il s'agit du génome haploïde parent de *C. arabica*. L'analyse comparée des génomes du café et d'autres plantes (asteridées et rosidées) a révélé que l'ordre des gènes du caféier est le plus conservé au sein des astéridées et très proche de celui de l'espèce ancestrale dont toutes les plantes eudicotylédones ont dérivé. De plus, hormis la triplication commune à tous les eudicots, le caféier n'a pas subi d'autre duplication globale du génome (**Figure 13**).

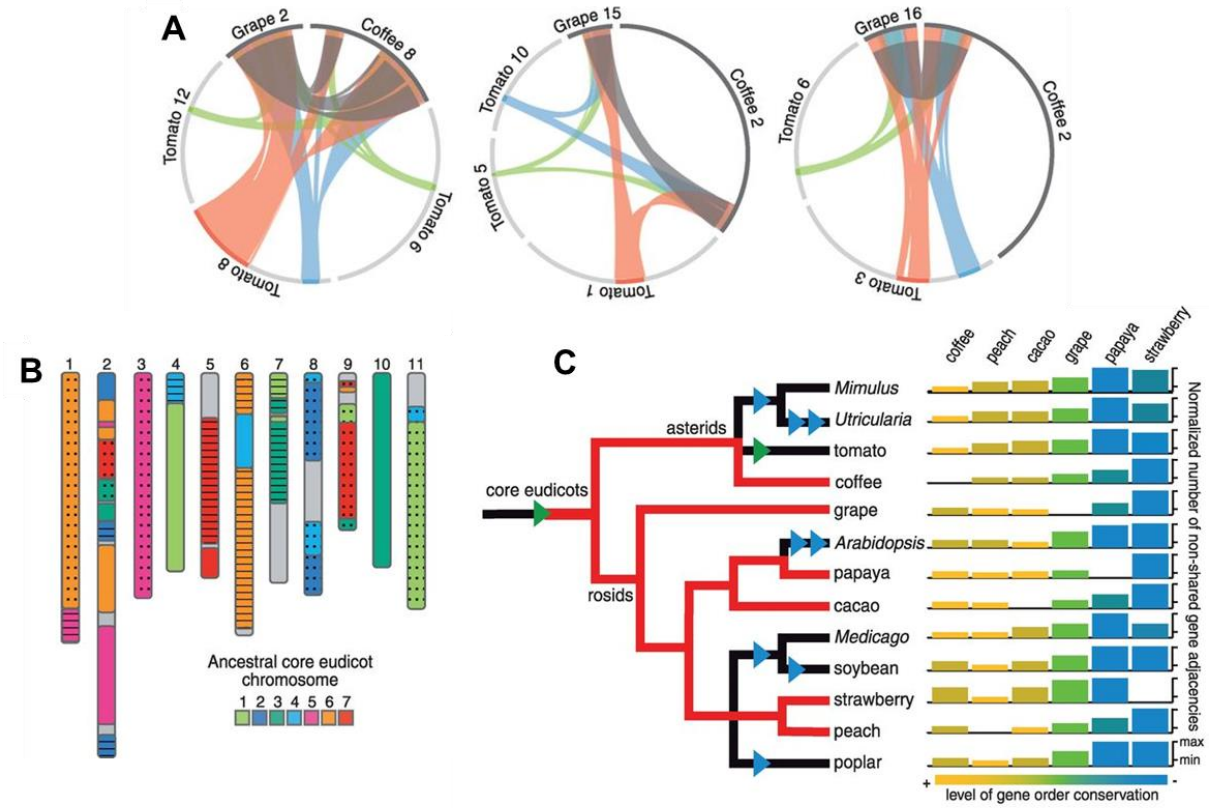

Figure 13 : Structure du génome de *C. canephora* (**A**) Comparaison de trois chromosomes de vigne (descendants de l'eudicot ancestral avant hexaploïdie) alignés sur un seul chromosome de caféier et sur trois régions du génome de la tomate. (**B**) Blocs descendant de sept chromosomes eudicots ancestraux sur les chromosomes de caféier. Pour chaque chromosome ancestral, les trois descendants paralogues partagent la même couleur mais des textures différentes (**C**) Phylogénie et histoire des duplications du génome eudicot ancestral. Les flèches indiquent les évènements de tétraploïdisation (bleu) ou d'hexaploïdisation (vert). Les histogrammes reflètent le degré de divergence de l'ordre des gènes entre paires d'espèces. L'ordre des gènes entre les astéridées est plus conservé avec le caféier qu'avec les autres espèces (et chez les rosidées, c'est le cas pour le pêcher et le cacaoyer) (**Denoeud, 2014**).

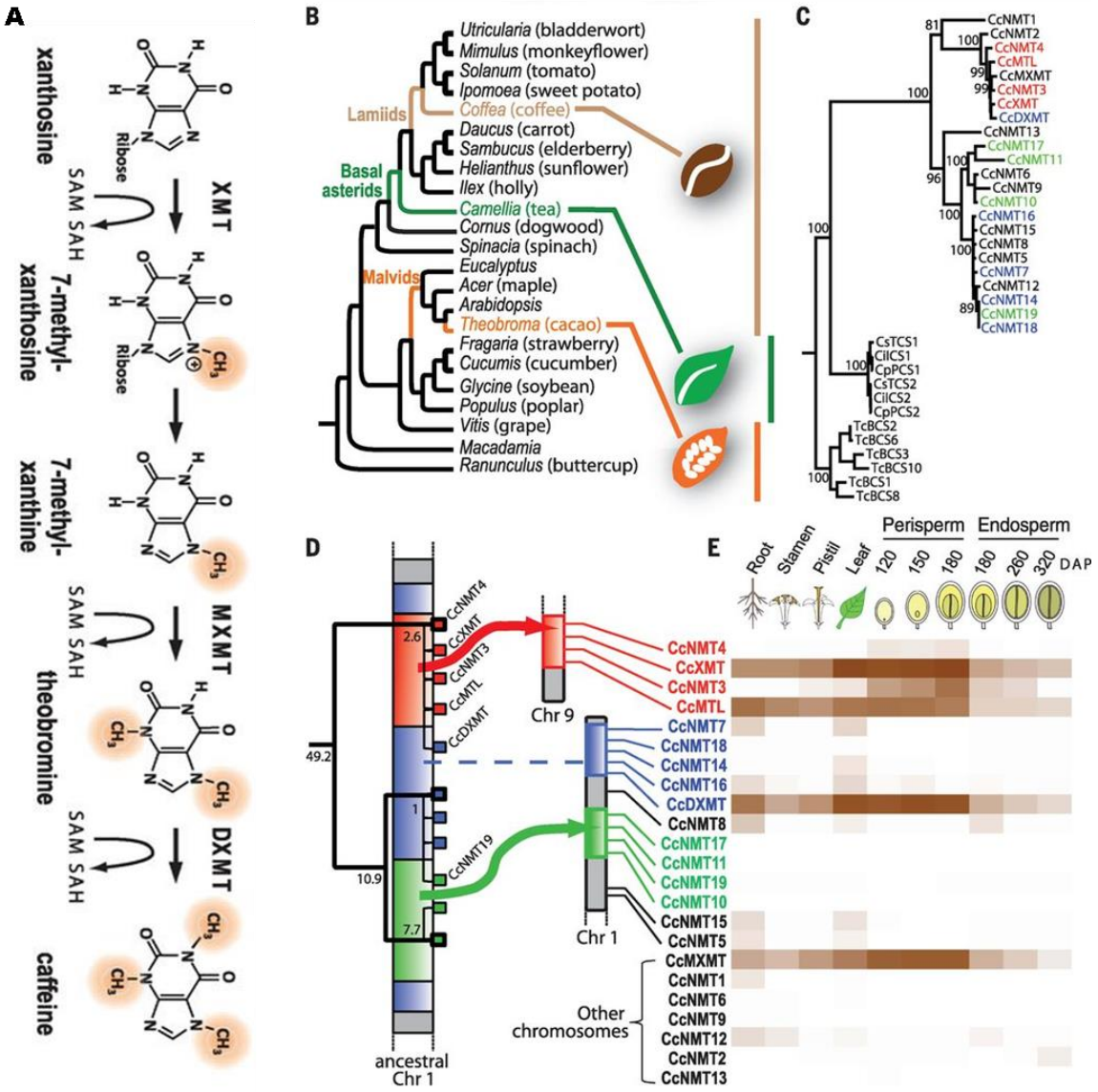

Figure 14: Evolution de la biosynthèse de caféine. (**A**) Voie de biosynthèse de la caféine. Trois étapes de méthylation sont nécessaires pour produire de la caféine à partir de la xanthosine, impliquant trois N-méthyl transférases (NMTs) : XMT, MXMT, et DXMT. (**B**) Arbre phylogénétique représentant la position des trois plantes synthétisant de la caféine(/théine/théobromine) par rapport aux autres eudicots (**C**) Phylogénie (maximum likelihood) des NMTs de caféier, théier et cacaoyer. Les valeurs de support de Bootstrap pour 1000 réplicats sont reportées à côté des branches. Les couleurs des gènes correspondent aux blocs génomiques montrés en D. (**D**) (Gauche) Modèle résumant l'histoire des duplications des NMT de Coffea. Trois blocs de gènes dupliqués en tandem ont évolué sur le chromosome 1 à partir des gènes entourés en gras. Les blocs rouge et vert ont ensuite été transloqués vers le chr9 ou réarrangés plus loin sur le chr 1, respectivement. (Droite) Ordre des gènes sur les chromosomes actuels. Les gènes du bloc rouge ont été transloqués sauf le gène *CcDXMT*, resté sur le chr 1. De la même façon, *CcNMT19* dérive d'un gène du bloc bleu mais a été transloqué avec le bloc vert (**E**) Profils d'expression (RPKM) des NMT de *C. canephora*. Les gènes appartenant au cluster métabolique putatif (bloc rouge) ainsi que *CcDXMT* et *CcMXMT* montrent des profils d'expression similaires, plus élevés dans le périsperme que dans l'endosperme. (**Denoeud, 2014**).

Une analyse comparative avec le génome du cacaoyer montre par ailleurs que la biosynthèse de caféine est due à des enzymes propres à chaque espèce, apparues à divers moments au cours de l'évolution. Les N-méthyl transférases responsables de la synthèse de caféine dans le café sont apparues par duplications en tandem successives, suivies de translocations et de réarrangements (**Figure 14**).

L'expertise que j'ai acquise au cours de cette période sur les duplications de génomes de plantes m'a valu d'être invitée en tant que que Keynote speaker à la conférence RECOMB Computational Genomics, Lyon Villeurbane, 17-19 oct 2013 : "Whole genome duplications in plants".

Il me semble important de souligner ici que certaines de ces expériences, bien que très enrichissantes scientifiquement, m'ont laissé une légère frustration : en effet, j'ai mené avec grand plaisir des analyses passionnantes, mais sur certains de ces projets, j'ai un peu manqué de liberté pour mener mes propres analyses. J'ai pu parfois être traitée comme le « postdoc » digne de confiance qui génère tous les résultats, au service d'un PI. Ce rôle m'a très bien convenu pendant mon postdoc, mais depuis le début de ma carrière au Genoscope, dans le laboratoire de Patrick Wincker, j'avais pris l'habitude de faire mes analyses en grande autonomie. En particulier, j'avais énormément apprécié, dès mon arrivée, d'avoir eu la liberté de creuser la thématique autour des introns d'*Oikopleura dioica*. Je me suis rendu compte, à la suite de ces expériences, qu'il fallait que je développe ma propre thématique, transversale, qui transcenderait l'intérêt d'un seul projet, tout en profitant de la manne de données dont j'ai la chance de disposer en travaillant au Genoscope. Cela s'est avéré plus facile à dire qu'à faire, car j'ai souvent été rattrapée par des projets tous plus intéressants les uns que les autres, que j'ai d'ailleurs pour la plupart choisis moi-même en fonction de mes appétences : difficile dans ces conditions de trouver le temps pour creuser mon sillon personnel. Cette démarche a donc nécessité quelques années de maturation supplémentaires.

5/ Élaboration de projets de recherche personnels

J'ai pris un congé parental entre octobre 2014 et septembre 2015. A mon retour, j'ai commencé à développer une thématique de recherche personnelle et j'ai sélectionné des projets en rapport avec celles-ci. A cette période, j'ai été la responsable bioinformatique pour plusieurs projets France Génomique. France génomique (https://www.france-

genomique.org/) est une infrastructure créée grâce à un financement « Investissements d'Avenir » dans le cadre du projet « Infrastructures nationales en biologie et santé », née de la volonté de renforcer et de placer à la pointe de l'état de l'art les capacités françaises dans le domaine de la génomique à haut débit et de la bioinformatique associée. Elle rassemble plusieurs plateformes de séquençage en France, dont bien-sûr le Genoscope. Je présente ci-après les différents projets choisis et comment ils s'intègrent dans mon projet de recherche, ainsi que mon implication dans chacun d'eux. Pour le premier de ces projets, Polysuccess, j'ai surtout fait de la gestion de projet, avec juste quelques analyses, mais les autres m'ont permis d'imaginer et de mener à bien des projets de recherche personnels.

a/ Projet *Polysuccess* : comment un polyploïde devient une nouvelle espèce

Le premier projet vise à étudier les conséquences à court et moyen terme d'un évènement de polyploïdisation récente en utilisant des colzas synthétiques (projet Polysuccess « how a polyploid becomes a new species, the Brassica model » – porteuse de projet : Anne-Marie Chèvre, INRA de Rennes). Ce projet s'intéresse aux variations fonctionnelles (expression des gènes) et structurales (échanges homéologues, éléments transposables) survenus à court (dans les polyploïdes synthétiques) et long terme (dans les colzas naturels). En plus de la coordination du projet, j'ai été en charge de l'analyse des échanges homéologues dans les colzas synthétiques et naturels. Nous avons montré que ces événements sont plus fréquents chez les colzas synthétiques que chez les colzas naturels cultivés qui présentent seulement des translocations de petite taille (Chalhoub 2014 ; Hurgobin 2018) (**Figure 15**). Il semblerait donc qu'un mécanisme de régulation de la recombinaison homoéologue se soit mis en place rapidement après la fomation du colza (Rousseau-Gueutin et al, en préparation). Bien que ces colzas resynthétisés présentent une instabilité génomique certaine et une faible fertilité, ces échanges entre chromosomes homéologues ne sont pas forcément délétères et au contraire peuvent avoir des effets bénéfiques en contribuant par exemple à l'apparition de nouveaux phénotypes mieux adaptés à l'environnement. Ces effets sont étudiés par nos collaborateurs dans l'équipe de Rennes.

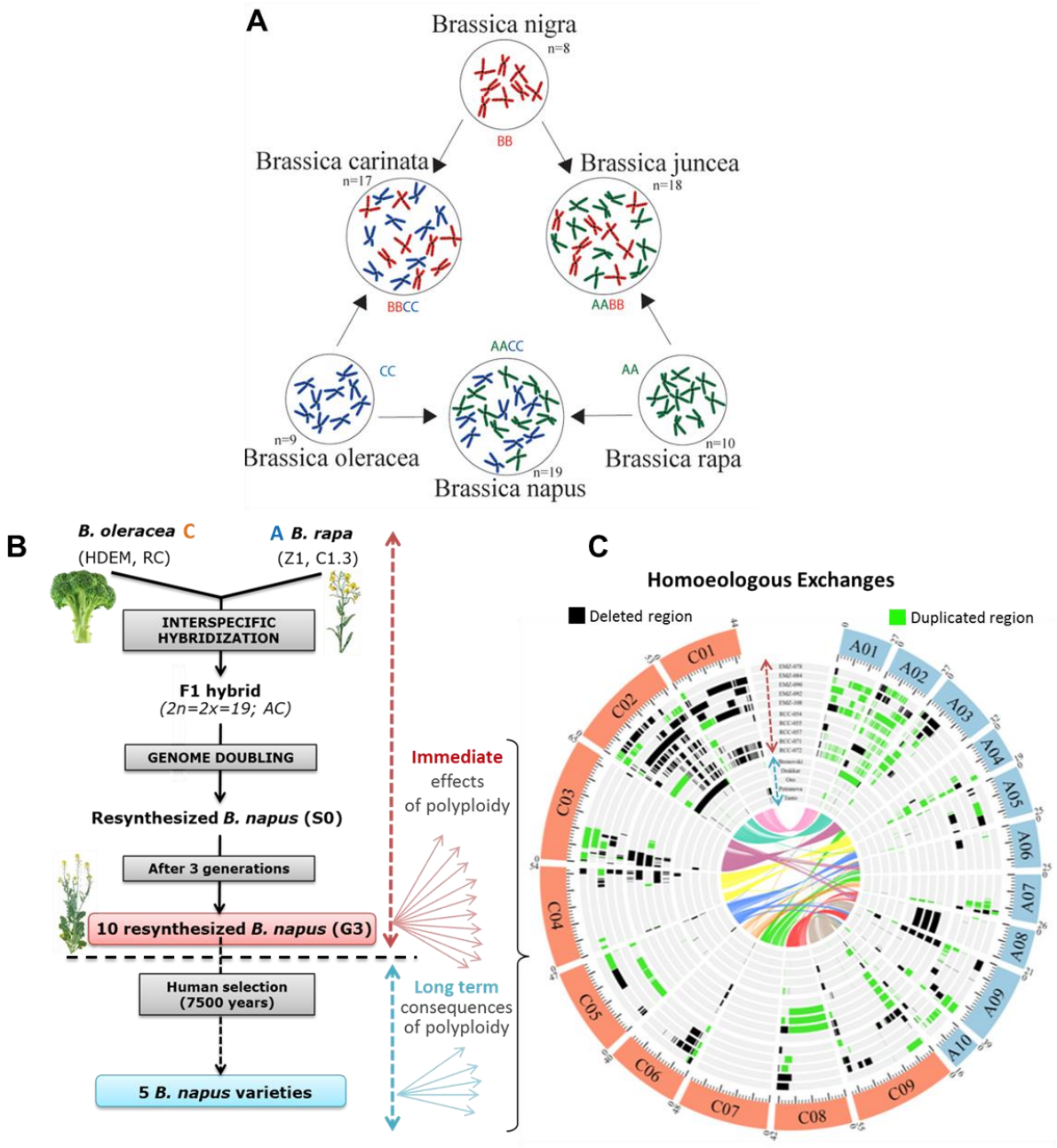

Figure 15 : Evénements d'échanges homéologues dans des colzas synthétiques et naturels. **A.** triangle de U représentant les événements d'hybridations survenant dans le genre Brassica : le colza, Brassica napus (génome AC), provient d'une hybridation entre Brassica rapa (génome A) et Brassica oleracea (génome C). **B.** Schéma représentant les étapes effectuées pour obtenir des colzas synthétiques. **C.** Représentation circos des 19 chromosomes du colza, avec leurs liens de synténie (au centre). Les 10 souches synthétiques sont sur l'extérieur et les 5 souches naturelles sur l'intérieur. Les événements sont très abondants sur les paires de chromosomes A01/C01 et A02/C02 (qui sont les plus synténiques), ainsi que chez les colzas synthétiques par rapport aux colza naturels.

En parallèle de ces travaux portant sur la dynamique évolutive structurale de colzas synthétiques, j'ai aussi participé à la génération et à la publication d'assemblages à l'échelle chromosomique pour les espèces diploïdes parentales des colzas resynthétisés (*Brassica rapa* Z1 (moutarde jaune) et *Brassica oleracea* HDEM (brocoli) (Belser, 2018). Cette même technique a aussi été utilisée pour améliorer l'assemblage du génome de référence du colza cv. Darmor (Rousseau-Gueutin, 2020).

b/ *PhyloAlps* et *PhyloNorway* : « genome skimming » sur la flore alpine et arctique

Ce projet a été pour moi l'occasion d'imaginer des méthodes originales pour « faire parler » des données fragmentaires, qui pourront être appliquées à d'autres jeux données génomiques ou métagénomiques. Il a aussi été l'occasion pour moi d'encadrer plusieurs stagiaires de niveau M2. Le projet PhyloAlps (porteur de projet : Sébastien Lavergne, laboratoire d'écologie alpine de Grenoble) vise à explorer la diversité de la flore alpine (hotspot de biodiversité comprenant 20% de la flore européenne). Dans ce but, plus de 5000 échantillons de plantes alpines ont été prélevés et séquencés à faible couverture (60% des échantillons ont une profondeur de séquençage inférieure à 1X), afin d'assembler les génomes chloroplastiques et d'établir une phylogénie des espèces alpines. Cette approche (séquençage à faible couverture pour obtenir la « crème » de l'information) est nommée "genome skimming" (qui se traduirait en Français par « écrémage génomique ») (Coissac, 2016, Mc Kain, 2018). Les assemblages de génomes d'organelles peuvent être compliqués par leur structure complexe, en particulier dans l'organisation des éléments répétés (Kozik, 2019 ; Alsos, 2020). Nos collaborateurs ont donc développé une méthode nommée ORTHOSKIM permettant de capturer *in silico* les séquences de gènes cibles à partir des données de séquençage de type « genome skimming », qu'ils ont testée sur les données PhyloAlps pour 2 familles de plantes (Primulaceae et Ericacea) (Pouchon, 2022). Une autre étude portant sur la diversification des espèces alpines et utilisant la phylogénie obtenue pour 6 clades de plantes avec les échantillons PhyloAlps est en cours de publication (Smyčka, sous presse).

En parallèle du projet PhyloAlps, notre collaborateur Eric Coissac, du LECA de Grenoble, nous a mis en relation avec une équipe norvégienne de l'Arctic University de Tromsø, dirigée par Inger Greves Alsos, qui menait le même type de projet sur la flore arctique. Le Genoscope a donc également séquencé plus de 2000 spécimens de plantes de la flore arctique, essentiellement issus d'herbiers (Alsos, 2020). Ces données ont été générées principalement afin de disposer de séquences de référence pour identifier les espèces présentes dans des

échantillons d'ADN ancien prélevés dans des sédiments (« sedaDNA »). Deux grandes études utilisant les données PhyloNorway ont été publiées ou soumises. La première concerne l'ensemble de l'arc arctique et remonte jusqu'à il y a 50 000 ans. Elle a mis en évidence une transition de la flore lorsque le climat s'est réchauffé et humidifié, passant d'une steppe-toundra adaptée au froid vers une mosaïque entre des régions boiseuses et d'autres herbeuses, vraisemblablement à l'origine de l'extinction des mammouths (Wang, 2021). La seconde étude s'est focalisée sur la Scandinavie et remonte jusqu'à il y a 16 000 ans. Elle a montré qu'après la dernière grande glaciation, la diversité de plantes connue aujourd'hui en Scandinavie a mis des milliers d'années à s'établir (Alsos et al., soumis).

A partir des données de séquençage obtenues pour les projets PhyloAlps et PhyloNorway, j'ai eu la liberté de mener les analyses qui me semblaient pertinentes, sans aucune directive des collaborateurs : en effet, ceux-ci s'intéressaient aux séquences chloroplastiques, tandis que pour ma part j'ai souhaité mettre à profit l'ensemble des données pour obtenir de l'information, non pas sur le génome chloroplastique, mais sur le génome nucléaire des différentes espèces échantillonnées. Il s'agit donc de projets de recherche totalement personnels. C'est pourquoi c'est sur cette thématique que j'ai choisi de recruter des stagiaires/potentiels futurs thésards. Lors de ces différents stages de M2 que j'ai encadrés, nous avons montré que les données de type « genome skimming » permettent de mesurer l'amplification (duplication génomique) de gènes/fonctions d'intérêt et de prédire la taille des génomes.

En 2018, j'ai encadré sur ce sujet une étudiante de Master 2 (BiB Biologie informatique / Bioinformatique, Université Paris Diderot) : Lina Alferkh. Son stage visait à mettre en évidence des familles de gènes amplifiées (par duplications en tandem par exemple), dans des voies métaboliques spécifiques, et à les mettre en relation avec des stress environnementaux tels que l'altitude extrême. En effet, les génomes de plantes sont fréquemment sujets aux duplications de gènes, mais également aux duplications segmentales ou en tandem (Panchy, 2016). Certaines voies métaboliques ont évolué à partir de duplications de gènes : par exemple, comme je l'ai montré précédemment, les enzymes produisant la caféine chez le caféier proviennent d'amplifications par duplications en tandem suivies de translocations et de réarrangements (Denoeud, 2014 ; **Figure 13**). Les plantes vivant en altitude sont soumises à d'importants rayonnements ultra-violets. Il a été montré que des métabolites secondaires tels que les anthocyanes (Zhou, 2007 ; Zoratti, 2015) et l'acide ascorbique (Roman, 2013) sont

davantage produits dans les plantes vivant à plus haute altitude, ce qui les protège du stress oxydatif. Ces études portent généralement sur la comparaison de la concentration en métabolites entre des plantes d'une même espèce, échantillonnées à différentes altitudes, et peuvent résulter d'un contrôle transcriptionnel ou post-transcriptionnel de la production de ces métabolites secondaires. Lina Alferkh a utilisé les données issues du projet PhyloAlps afin de mettre en évidence des régulations au niveau génomique (amplifications de gènes), et non plus transcriptionnel ou post-transcriptionnel, entre différentes espèces vivant à des altitudes différentes. Pour cela, elle a quantifié les nombres de copies (abondances relatives) de gènes dans trois voies métaboliques d'intérêt (biosynthèse des caroténoïdes, des anthocyanes et de l'acide ascorbique) dans les échantillons PhyloAlps par alignement des lectures traduites en séquences protéiques sur les groupes de gènes orthologues –ou KO (KEGG orthologs (http://www.genome.ad.jp/kegg/)- de ces trois voies Nous avons montré que cette couverture, normalisée en utilisant la profondeur sur des protéines monocopies peut être utilisée comme proxy du nombre de copies de gènes dans les différents KO (**Figure 16**).

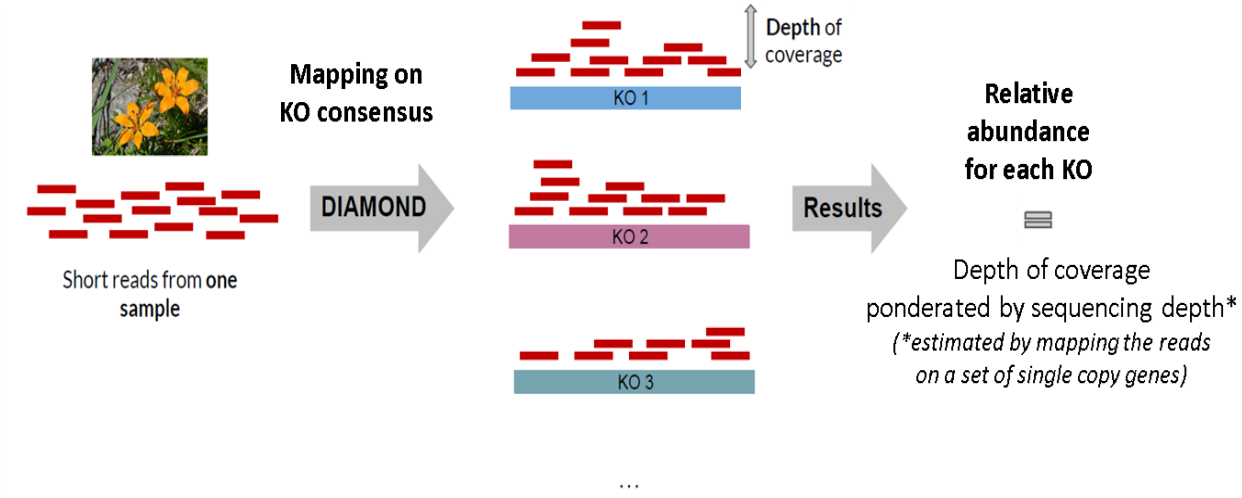

Figure 16 : méthode de quantification de l'abondance relative (c'est-à-dire du nombre de copies génomiques) de gènes à partir de l'alignement de courtes lectures, développée lors du stage de M2 de Lina Alferkh.

Lina s'est ensuite attelée à identifier des corrélations entre l'amplification génomique de ces gènes/voies et le mode de vie des plantes (altitude, rayonnement UV, forme de vie…) provenant de la base de données « Flora indicativa ». Elle a obtenu des résultats intéressants pour certains gènes, mais l'interprétation a été compliquée par le fait que les données sur les traits de vie des espèces provenant de Flora indicativa n'étaient pas assez précises. Mon expérience d'encadrement avec Lina s'est révélée très enrichissante, et m'a amenée à lui proposer de poursuivre avec moi pour une thèse. J'ai déposé en 2018 un sujet à l'école

doctorale SDSV. Lina s'est présentée en parallèle à d'autres concours pour d'autres sujets de thèse et elle a reçu une réponse positive avant la tenue du concours de l'école doctorale SDSV : elle a donc préféré poursuivre une autre thèse, dans le domaine de la prédiction de structures d'ARN, et j'ai eu le plaisir d'écouter sa soutenance il y a quelques semaines. Forte de cette expérience positive d'encadrement, j'ai déposé début 2019 mon dossier de candidature pour l'autorisation de soutenir l'HDR.

L'année suivante (en 2019), j'ai encadré un autre étudiant du Master 2 BiB Denis Diderot, Daniel de Murat, afin d'étendre ces analyses de quantification à l'ensemble des gènes de plantes (tous les Kegg Orthologues ou KO : https://www.kegg.jp). Il a pu mettre en évidence des « signatures métaboliques » pour différentes lignées de plantes. En effet, il a montré que les profils d'abondance des KO permettent de regrouper les plantes selon leur phylogénie (**Figure 17A**). Cette analyse a aussi permis d'identifier des gènes amplifiés dans certaines lignées par rapport à d'autres (par exemple K14488 « SAUR-like auxin-responsive family » est significativement plus abondant chez les monocotylédones que chez les dicotylédones. Ce résultat est cohérent avec le fait que ces deux lignées n'ont pas la même réponse aux herbicides auxiniques (McSteen, 2010) (**Figure 17B**). Hélas, Daniel n'a pas eu envie de poursuivre en thèse par la suite.

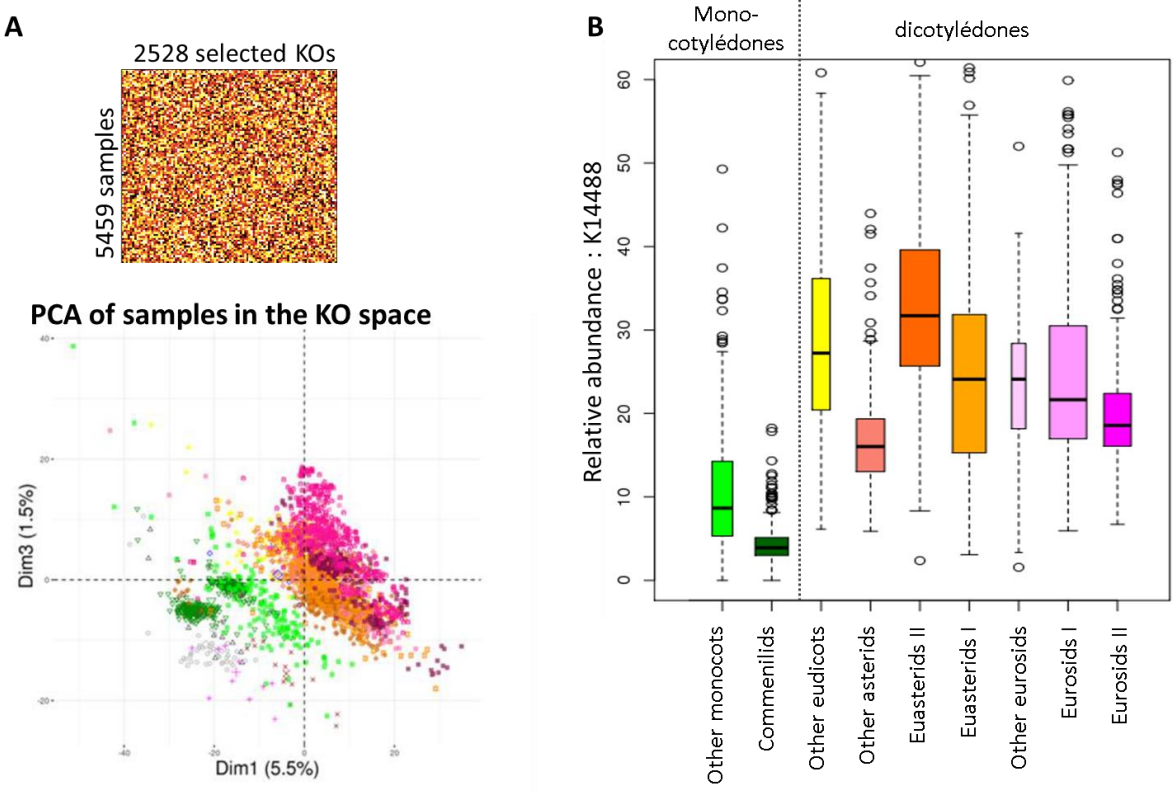

Figure 17 : Résultats obtenus par Daniel De Murat lors de son stage de M2. **A :** ACP obtenue sur la matrice de valeurs d'abondance des KO pour les différents échantillons : les branches phylogénétiques correspondent au même code couleur que sur le panel B. **B.** Exemple d'un KO significativement plus abondant chez les dicotylédones que les monocotylédones.

En 2020, je n'ai pas recruté de stagiaire car j'ai souhaité prendre du recul sur les résultats déjà engrangés (ce qui est plutôt bien tombé compte-tenu du contexte particulier cette année là). J'ai développé un autre axe de recherche sur le projet PhyloAlps, qui concerne les tailles de génomes. En effet, les génomes de plantes ont des tailles très variables (Soltis, 2003 ; Bennet, 2005 ; Bennetzen, 2005) et les contraintes évolutives en cause dans l'équilibre entre les mécanismes d'augmentation et de diminution de taille des génomes sont encore mal comprises. J'ai voulu tirer profit des lectures génomiques obtenues pour un grand nombre d'échantillons (des deux projets PhyloAlps et PhyloNorway) afin de prédire la taille des génomes de ces plantes (inconnue pour beaucoup d'entre-elles).

J'ai mis en évidence, en utilisant les tailles de génomes connues, disponibles dans la base de données de référence au Kew botanical garden (http://data.kew.org/cvalues/), que la profondeur de couverture sur un jeu de protéines monocopies est très bien corrélée à la profondeur de séquençage estimée à partir de la taille du génome monoploïde (1Cx) (**Figure 18A**). J'ai utilisé les protéines monocopies issues de l'initiative « OneKp » (One thousand plant transcriptomes). A partir des transcriptomes de 1090 espèces de plantes (Viridiplantae), les auteurs ont constitué un jeu de 410 gènes présents en une seule copie dans tous les génomes de plantes, qu'ils ont utilisé pour reconstruire une phylogénie (Leebens-Mack, 2019). Des tests préliminaires effectués en utilisant les gènes BUSCO Plants, jeu de gènes monocopies générés dans le but d'estimer la complétion des annotations de génomes (Simão, 2015), ont fourni des résultats presque aussi satisfaisants mais nécessitaient des temps de calcul un peu plus longs. J'ai donc développé une procédure utilisant l'alignement des lectures traduites sur les protéines OneKp pour extrapoler la taille des génomes. Cette procédure est similaire à la procédure d'alignement sur les KO : il s'agit de calculer la profondeur moyenne sur un ensemble de protéines monocopies. C'est d'ailleurs cette valeur utilisée comme proxy de la profondeur de séquençage pour normaliser les abondances relatives des KO (**Figure 16**). Pour chaque grand clade de plantes (correspondant aux couleurs sur la **Figure 18B**), une régression linéaire appliquée sur les espèces dont la taille de génome était connue m'a permis d'extrapoler les tailles des génomes inconnues.

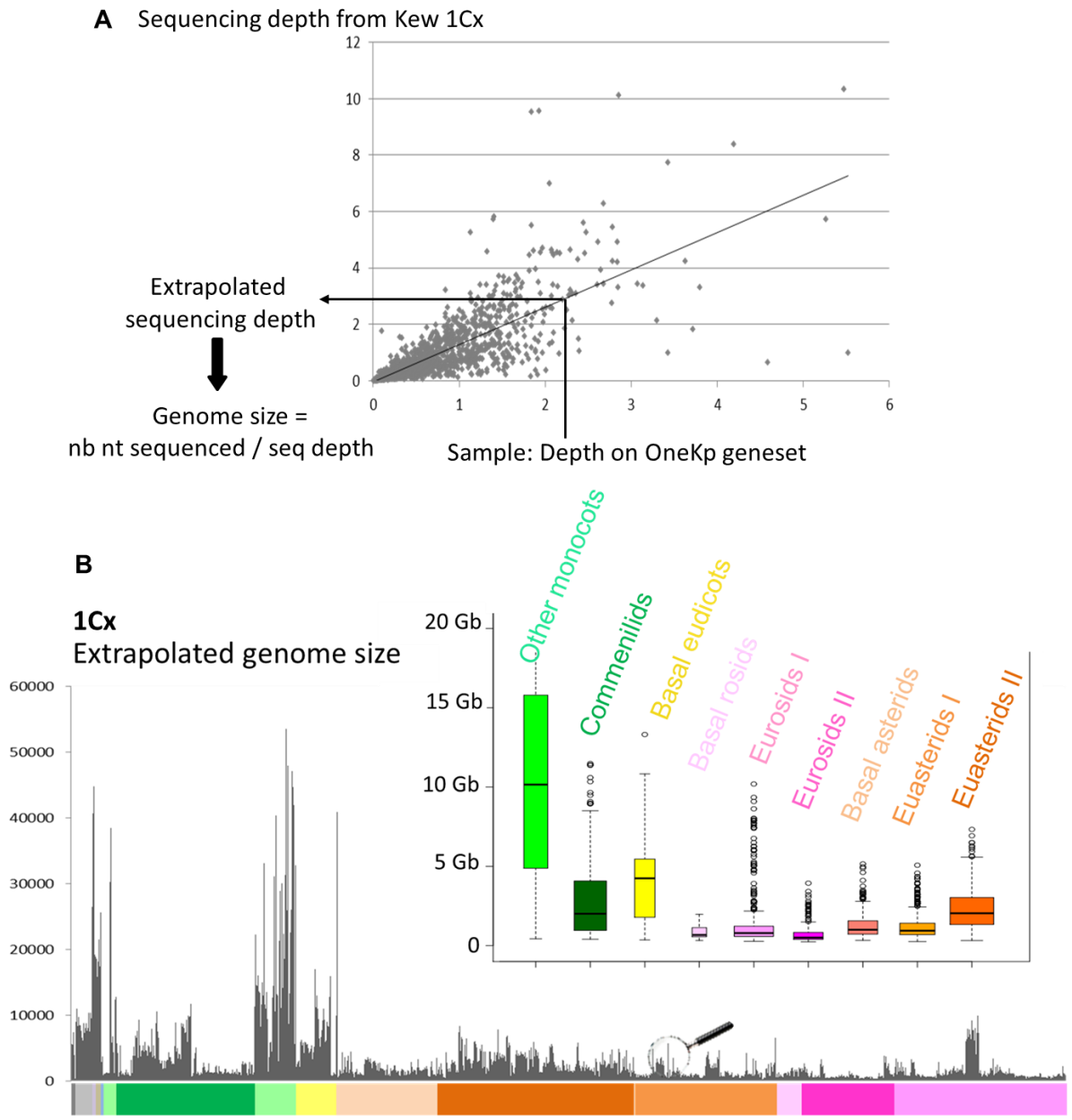

Figure 18 : A. Corrélation entre les profondeurs d'alignements de lectures sur un jeu de protéines monocopies et la profondeur de séquençage calculée à partir de la taille de génome monoploïde (1Cx). **B.** Estimations de tailles de génomes obtenues par régression linéaire dans les échantillons PhyloAlps classés par branche phylogénétique. On y observe comme attendu que les gymnospermes (brun) et les monocots (vert clair) ont de très grands génomes. On distingue aussi un groupe parmi les Eurosids I présentant des grandes tailles de génomes : il s'agit de légumineuses telles que le pois, qui sont connues pour contenir de très grandes quantités d'éléments transposables.

Je prévois par la suite de rechercher des corrélations entre ces tailles de génomes et certaines caractéristiques des plantes (par exemple, il a déjà été montré que la taille des génomes est corrélée à la taille des graines et à leur degré de dispersion (Beaulieu, 2007 ;

Pandit, 2014). Par ailleurs, des résultats très intéressants sont apparus lorsque j'ai recherché des corrélations entre l'amplification de certaines familles de gènes et la taille des génomes : le nombre de copies de gènes d'histones est fortement corrélé à la taille du génome monoploïde. Les histones sont les constituants des nucléosomes, complexes protéiques autour desquels s'enroule la molécule d'ADN. Ce sont des familles multigéniques très abondantes, et extrêmement conservées (Piontkivska, 2002). Les histones de plantes ont plusieurs propriétés très différentes des histones d'animaux. Tout d'abord, chez les plantes, les gènes d'histones sont dispersés tout au long du génome alors qu'ils se répartissent dans plusieurs grands clusters de gènes répétés en tandem chez les animaux. De plus, chez les métazoaires, les ARNm d'histones ne contiennent pas de queue polyA mais une structure « stem loop » qui permet la fixation de la protéine SLBP et la maturation par snRNA U7 (Marzluff, 2008). En revanche, chez les plantes, les ARNm d'histones contiennent des queues polyA, et pas de stem loop, et la protéine SLBP est absente, tout comme le snRNA U7. Les transcrits d'histones de plantes ressemblent donc davantage à des ARN messagers classiques. Une régulation au niveau transcriptionnel mais aussi post transcriptionnel (lié à la phase S de synthèse d'ADN), est observée chez les métazoaires, mais seule la régulation transcriptionnelle semble s'opérer chez les plantes, même si les mécanismes qui permettent aux protéines d'histones de s'accumuler pendant la phase S chez les plantes ne sont pas encore connus à ce jour. La très forte corrélation entre le nombre de copies d'histones et la taille du génome monoploïde observée chez les plantes suggère que dans ces espèces, la quantité d'histones pourrait être régulée par le nombre de copies génomiques. On pourrait imaginer un mécanisme un peu simpliste, où au fur et à mesure de la réplication, les gènes d'histones libérés seraient transcrits puis traduits afin d'être disponibles pour compacter l'ADN nouvellement synthétisé.

En outre, les histones sont une famille multigénique intéressante du point de vue de leur évolution : la pression de sélection (purifiante) y est telle que leur séquence protéique est extrêmement conservée (100% d'identité dans l'ensemble des Embryophytes pour l'histone H4) (Draizen, 2016). En revanche, les séquences nucléotidiques divergent (aux positions synonymes) et des analyses préliminaires que j'ai menées à partir des lectures provenant d'échantillons PhyloAlps montrent que plusieurs séquences nucléiques d'histone H4 coexistent dans chaque espèce, et confirment que le phénomène de conversion génique n'a pas lieu. Il serait intéressant de développer des approches pour assembler les séquences nucléiques des histones dans les jeux de données à faible couverture dont nous disposons afin d'étudier plus en détail les mécanismes évolutifs en jeu dans cette famille multigénique.

Entre mars et septembre 2021, j'ai accueilli une étudiante du Master 2 « AMI2B » de l'université de Paris Saclay, Tolulopé Apanishile. Son projet visait à utiliser cette procédure de prédiction de taille de génomes dans la famille des astéracées. En effet, nous collaborons avec Oriane Hidalgo et Jaume Pellicer (Institut Botànic de Barcelona et Kew royal botanic gardens) qui ont mesuré par cytométrie de flux de nombreuses tailles de génomes dans la famille des astéracées. Ces mesures estiment le contenu d'ADN dans une cellule, c'est à dire la taille du génome diploïde (2C) ou holoploïde (1C), selon le type de cellule étudié (en général des gamètes, 1C). La combinaison de ces mesures avec notre approche, qui estime la taille du génome « monoploïde » (1Cx) (Greilhuber, 2005 ; **Figure 19**) permet ainsi d'estimer le niveau de ploïdie pour certaines espèces où il reste inconnu.

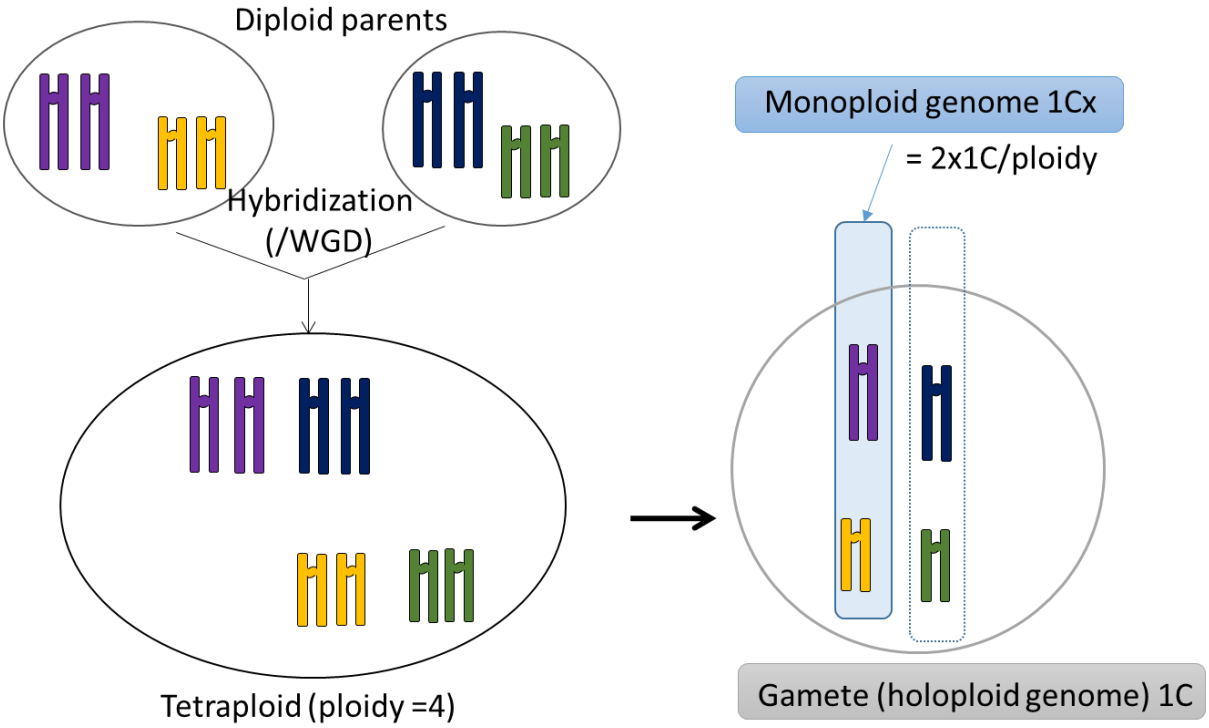

Figure 19 : Définition des valeurs de tailles de génome 1C et 1Cx.

Enfin, Tolulopé a calculé les corrélations entre les tailles de génomes prédites et les abondances relatives des KO dans les échantillons d'Astéracées et elle a pu confirmer une observation que j'avais pu faire sur l'ensemble des espèces échantillonnées dans les projets PhyloAlps et PhyloNorway : le nombre de copies de gènes d'histones est corrélé à la taille du génome monoploïde. Tolulopé a développé pendant ce stage un grand intérêt pour l'analyse de données (en particulier, elle a été très motivée lorsque je lui ai fait tester des approches

d'intelligence artificielle). Elle a donc choisi de se diriger vers le domaine des « data science » et a commencé un nouveau Master dans ce domaine.

Depuis mars 2022, j'accueille un stagiaire en dernière année d'école d'ingénieur (Ecole Nationale Supérieure Agronomique de Toulouse), Pierre Guenzi-Tibéri. Son stage consistera à développer le pipeline de prédiction de tailles de génomes afin qu'il soit utilisable par la communauté des botanistes qui étudient les tailles de génomes et les niveaux de ploïdie. Nous avons prévu de rédiger un article décrivant ce pipeline, dont Pierre sera le premier auteur et moi le dernier. Il appliquera ensuite cet outil à l'ensemble des échantillons PhyloAlps : les résultats seront publiés dans l'article décrivant l'ensemble du jeu de données ou bien dans un article compagnon. Par ailleurs, cette méthode de quantification du nombre de copies des gènes par l'alignement de courtes lectures peut être appliquée à d'autres problématiques. Je montrerai en section d) comment je l'ai utilisée pour comparer des génomes de coraux. De la même façon, dans le projet Phaeoexplorer (décrit en section c), je propose d'utiliser cette méthode pour quantifier les nombres de copies dans les familles multigéniques chez les algues brunes. Cette approche, reposant sur les courtes lectures, a l'avantage d'être indépendante des méthodes/qualités d'assemblage et d'annotation, qui peuvent être hétérogènes selon les génomes. Si Pierre souhaite poursuivre en thèse sur le projet Phaeoexplorer, il s'agira d'une bonne transition avec son sujet de stage.

Comme je l'ai déjà évoqué en préambule de ce document, même si j'ai eu une pleine liberté thématique pour utiliser les données de genome skimming issues des projets PhyloAlps et PhyloNorway, j'ai quand-même manqué de liberté au niveau temporel : en effet, la publication de mes résultats sur les tailles de génomes et leur corrélation avec le nombre d'histones est en attente de la soumission des données PhyloAlps depuis plusieurs années. J'ai pu faire le même type d'observation avec des données publiques, mais la portée de la découverte sera bien plus importante en utilisant les milliers d'espèces PhyloAlps, c'est pourquoi j'ai préféré attendre avant de publier mes résultats. La rédaction de l'article décrivant la phylogénie des plantes alpines commence maintenant, et les données seront rendues publique dans l'année. Je suis donc en train de préparer l'article à soumettre très prochainement.

c/ Phaeoexplorer : analyse des génomes de 40 espèces d'algues brunes

Depuis 2018, j'occupe une position de co-PI avec Mark Cock (CNRS Station Biologique de Roscoff) sur projet de séquençage d'une quarantaine d'espèces d'algues brunes (projet Phaeoexplorer). Dans ce projet, et je gère donc le projet de façon globale (par exemple, je prends des décisions concernant la mise en place de deadlines, le gel des données, et l'inclusion de membres dans le consortium et le partage des données à l'extérieur). J'ai également été la porteuse d'un projet ANR, j'y reviens plus loin. Je suis également en charge du groupe d'analyse « general genome features », je jouerai donc un rôle essentiel dans l'orientation des analyses et dans la rédaction de l'article décrivant ces génomes. Ce projet me permettra d'aborder deux des thématiques de recherche qui me sont chères : l'évolution des introns et l'amplification de familles de gènes et ses conséquences fonctionnelles et évolutives.

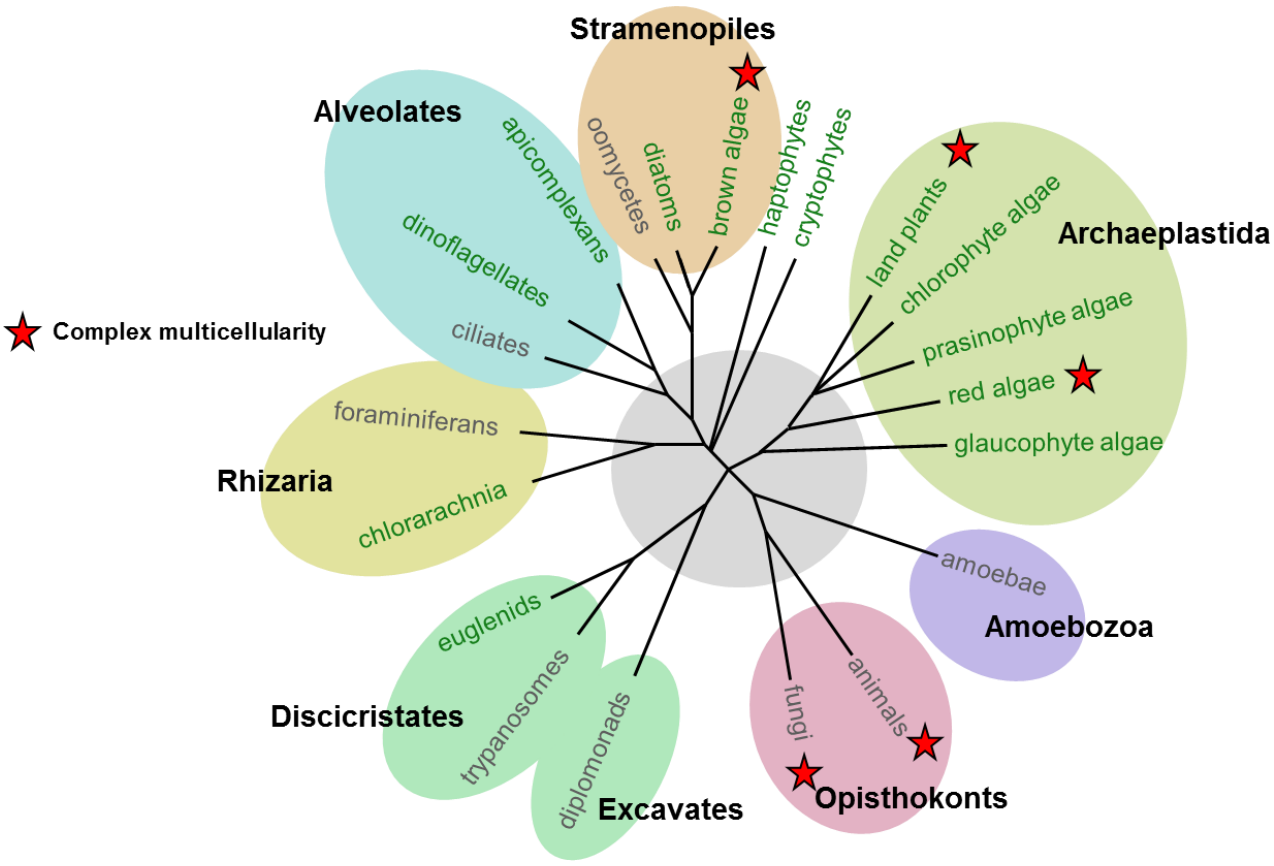

Figure 20 : Apparition de la multicellularité dans les différentes branches du vivant. La multicellularité (étoile rouge) n'est apparue (indépendamment) que dans 5 branches du vivant, son apparition la plus récente étant survenue dans la lignée des algues brunes.

Les algues brunes (Phaeophyceae) sont des algues multicellulaires photosynthétiques (Cock, 2011) qui font partie du groupe des straménopiles (qui inclut aussi les diatomées et les

oomycetes) et ont évolué indépendamment des autres lignées eucaryotes majeures depuis plus d'un milliard d'années (Cock, 2010). Cette histoire évolutive a été associée à l'émergence de nombreuses caractéristiques remarquables. En particulier, les algues brunes ont évolué vers la multicellularité complexe indépendamment des plantes et des animaux, et forment donc la troisième lignée multicellulaire la plus complexe du monde vivant (**Figure 20**), ce qui en fait un groupe clé pour la recherche sur la compréhension de cette transition évolutive majeure (Cock, 2010).

Les algues brunes montrent aussi une capacité d'adaptation remarquable aux environnements hostiles des côtes, où elles jouent des rôles écologiques essentiels, en tant qu'abri et source d'alimentation pour les espèces de l'écosystème côtier (Brodie, 2017). Ces adaptations incluent des parois cellulaires robustes mais très flexibles (contenant des polysaccharides avec un large spectre d'applications industrielles et biomédicales). Elles montrent également des adaptations physiologiques à leur environnement abiotique variable (zone intertidale, immergée et émergée selon la marée) et ont développé des mécanismes de défense essentiels pour ces organismes sédentaires, multicellulaires. En outre, la grande variété de cycles de vie (Cock, 2014) et de détermination sexuelle (Coelho, 2019) en font un modèle d'étude en biologie de la reproduction. Enfin, dans le monde entier, les algues brunes sont devenues une source de nourriture humaine importante, avec l'amplification de l'aquaculture (Buschmann, 2017), certaines fermes marines étant même visibles depuis l'espace.

Malgré tous ces intérêts, le groupe des algues brunes est encore peu étudié, avec à ce jour seulement 6 espèces ayant leur génome séquencé (Cock, 2010 ; Ye, 2015), dont certaines appartiennent au même ordre taxonomique. Afin de combler ce vide de connaissance, le projet France Génomique « Phaeoexplorer » (https://phaeoexplorer.sb-roscoff.fr) a été lancé en 2016, et implique un consortium international regroupant 30 laboratoires dans 11 pays. Ce projet vise à obtenir les séquences génomiques et les annotations de 67 souches d'algues brunes correspondant à 40 espèces, représentant tous les ordres majeurs, ainsi que 4 espèces sœurs. Même s'il est plus modeste, ce projet, s'inscrit dans la lignée des initiatives de génomique comparative à grande échelle qui ont vu le jour récemment. Ces projets ont pour but de séquencer le génome d'un grand nombre d'espèces, en particulier des espèces menacées d'extinction. On peut notamment citer l'initiative européenne ERGA « European Gene Atlas » (https://www.erga-biodiversity.eu) et le projet britannique « Darwin Tree of Life » (https://www.darwintreeoflife.org).

Les assemblages génomiques et les annotations sont en train d'être finalisés au Genoscope et la phase d'analyse va débuter. Je suis en charge de la coordination du groupe de travail « general genome features » qui regroupe plusieurs équipes du consortium Phaeoexplorer. Afin de mener ces analyses, je prévois d'accueillir un étudiant en thèse (projet proposé au concours de l'école doctorale pour un financement à partir d'octobre 2022). Il s'agira d'effectuer l'analyse comparative des génomes d'algues brunes et des espèces sœurs, par des approches bioinformatiques, dans le but d'identifier des mécanismes évolutifs d'intérêt dans cette lignée. Les analyses mises en œuvre comporteront l'étude de la synténie, des gains/pertes de gènes et des amplifications de familles de gènes. Un intérêt particulier sera porté aux gènes communs à l'ensemble des algues brunes (« core » genes) et aux gènes spécifiques de différentes espèces ou groupes d'espèces. De plus, des analyses seront aussi menées sur des familles de gènes d'intérêt, identifiées par notre approche à l'échelle des génomes entiers ou par des approches plus ciblées de reconstruction de réseaux métaboliques (par exemple les gènes impliqués dans la synthèse des parois cellulaires). La recherche de transferts horizontaux sera également effectuée, et mise en relation avec les résultats du groupe d'analyse qui s'intéresse, au sein du consortium Phaeoexplorer, aux bactéries vivant en association avec les algues (principalement dans leurs parois). Enfin, je prévois d'étudier la structure et l'évolution des introns dans ces espèces. Certaines contiennent peut-être des introns « exotiques » comme c'est le cas d'autres organismes marins que j'ai eu l'occasion d'étudier : le tunicier *Oikopleura dioica* ou les dinoflagellés, *Amoebophrya*. Je comparerai également les structures des gènes pour quantifier et décrire les phénomènes de gains et pertes d'introns dans cette lignée. Afin de faciliter ces analyses, j'ai été la porteuse d'un projet ANR soumis à l'AAP2020, nommé « Phaeodiscovery », qui visait à poursuivre l'effort de séquençage entrepris lors du projet Phaeoexplorer par un effort d'analyse. Ce projet n'a malheureusement pas été retenu.

L'un des défis de cette analyse de plusieurs dizaines de génomes réside dans l'hétérogénéité de la qualité des données pour ces différents génomes : certains sont à l'état de « drafts » (séquencés uniquement avec des courtes lectures, compte-tenu de la grande difficulté d'extraire de l'ADN de haut poids moléculaire des algues brunes) et d'autres sont de bonne qualité (séquencés avec des longues lectures nanopore). La thèse que je propose d'encadrer sera l'occasion d'inventer des approches méthodologiques nouvelles afin de comptabiliser le nombre de copies de gènes dans tous les génomes, indépendamment de leur qualité d'assemblage ou d'annotation. En particulier, l'approche que j'ai développée pour le

projet PhyloAlps, qui permet une quantification de nombre de copies de gènes, par alignement de lectures courtes, pourra être employée. Une analyse fine de la structure des gènes (en particulier, gains/pertes d'introns) sera effectuée parmi toutes les algues brunes à notre disposition, mais également au sein d'un groupe d'espèces plus restreint : le groupe des Ectocarpales, dans le but d'identifier des phénomènes de microévolution et de rechercher des signatures éventuelles de spéciation. Ce projet de recherche permettra de mieux comprendre l'évolution des différents génomes de la lignée des algues brunes et de lier leurs caractéristiques génomiques à des innovations biologiques majeures.

d/ *Tara Pacific* : mise en évidence de duplications en tandem massives dans deux génomes de coraux

Début 2020, le laboratoire « Bioinformatique pour la Génomique et la Biodiversité » (LBGB), dirigé par Jean-Marc Aury a été intégré à l'UMR « Génomique Métabolique », et Jean-Marc Aury et Patrick Wincker m'ont proposé de rejoindre cette équipe pour y renforcer la thématique d'analyse de génomes. J'ai donc rejoint ce laboratoire en septembre 2020, ce qui m'a occasionné un regain de motivation et de productivité. En effet, au LAGE, j'étais plutôt isolée, d'une part par ma thématique car j'étais la seule à ne pas travailler sur les métagénomes marins (provenant des projets TARA), et surtout d'autre part parce que le laboratoire est dédié exclusivement à l'analyse, et que le développements bioinformatiques s'opèrent principalement dans le laboratoire de Jean-Marc Aury, qui est constitué en grande partie d'ingénieurs. Ayant l'habitude de développer moi-même les outils (de façon sommaire, je ne me revendique pas comme une développeuse) dont j'ai besoin pour mes analyses, j'ai trouvé plus naturel de rejoindre cette équipe. C'est grâce à cela que je peux cette année encadrer un stagiaire d'école d'ingénieur sur un projet de développement, car l'un des ingénieurs de l'équipe, Benjamin Istace, que je remercie, encadre la partie « développement » du stage. L'arrivée dans ce nouveau laboratoire me donne donc davantage de possibilités pour publier mes articles personnels, portant sur des développements méthodologiques. A mon arrivée au LBGB, j'ai aussi eu la chance de me joindre à Benjamin Noel et Jean-Marc Aury pour analyser deux génomes de coraux. J'ai éprouvé un grand plaisir à mener ces analyses, car le travail en équipe m'avait manqué. Par ailleurs, j'ai pu à mettre à profit mon expertise concernant les duplications de gènes, cette fois-ci en étudiant des familles de gènes dupliqués en tandem. J'ai aussi pu utiliser mon approche de quantification des familles par alignement de lectures courtes pour valider l'annotation des génomes de coraux.

Dans le cadre du projet France Génomique Tara Pacific, j'ai pris part à l'analyse des génomes de deux espèces de coraux : *Porites lobata* et *Pocillopora meandrina*. Le projet Tara Pacific a consisté en une campagne d'échantillonnage dans des récifs répartis dans tout l'océan Pacifique entre 2016 et 2018 (Planes, 2019). Trois espèces (deux coraux ou sclératiniens : *Porites lobata* et *Pocillopora meandrina* et un corail de feu, *Millepora platyphylla*, cnidaire éloigné des sclératiniens dans la phylogénie) ont été ciblées. Sur chaque site, pour chaque espèce, des échantillons de colonies de corail et d'eau environnante, ont été prélevés. Le corail est un holobionte (Knowlton, 2003 ; Pogoreutz, 2020) : une espèce de corail vit en symbiose avec une microalgue (dinoflagellé *Symbiodinium*), qui lui fournit des nutriments essentiels via la photosynthèse. D'autres organismes peuvent être associés aux colonies de coraux, en particulier des bactéries et des virus. Afin d'analyser les échantillons métagénomiques prélevés lors de l'expédition Tara Pacific, le séquençage de génomes de référence pour *P. lobata* (appartenant à la branche des « complexes ») et *P. meandrina* (appartenant à la branche des robustes) était nécessaire. Ces génomes ont été séquencés au Genoscope avec des longues lectures (Oxford nanopore), ce qui a permis d'obtenir des assemblages de très bonne continuité (N50 de 2.15 Mb et 4.7 Mb respectivement pour *P.lobata* et *P.meandrina*). De plus, la méthode d'annotation employée a permis de constater qu'une grande proportion (environ un tiers) des gènes sont dupliqués en tandem dans les génomes de *P.lobata* et *P.meandrina* (**Figure 21**).

Un phénomène d'une telle ampleur n'avait pas été décrit pour les génomes de coraux précédemment publiés, même si des duplications en tandem avaient déjà été observées (Shinzato, 2011 ; Voolstra, 2017 ; Cunning, 2018 ; Ying, 2018 ; Ying, 2019 ; Buitrago-López, 2020). Nous avons cherché à savoir si cette observation pouvait s'expliquer par des biais dans les processus d'assemblage ou d'annotation. Afin d'éliminer de tels biais, j'ai utilisé l'approche d'alignement de courtes lectures décrite pour le projet PhyloAlps (**Figure16**). A partir des orthougroups (OG) calculés pour 22 espèces de cnidaires, j'ai construit des séquences consensus, sur lesquels les lectures Illumina provenant des différentes espèces ont été alignées. Le nombre de copies de chaque OG a été estimé en utilisant la profondeur de couverture pondérée par la profondeur obtenue sur 705 OG présents en une seule copie dans les génomes de coraux. J'ai constaté que l'estimation du nombre de copies était similaire parmi tous les *Porites* et tous les *Pocillopora*, tandis que le nombre de gènes annotés est supérieur pour *P. lobata* ainsi que pour *P. meandrina*. Il semble donc que les assemblages provenant de séquençages courtes lectures manquent des copies dans les familles

multigéniques par rapport aux assemblages longues lectures que nous avons générés (**Figure 21C**).

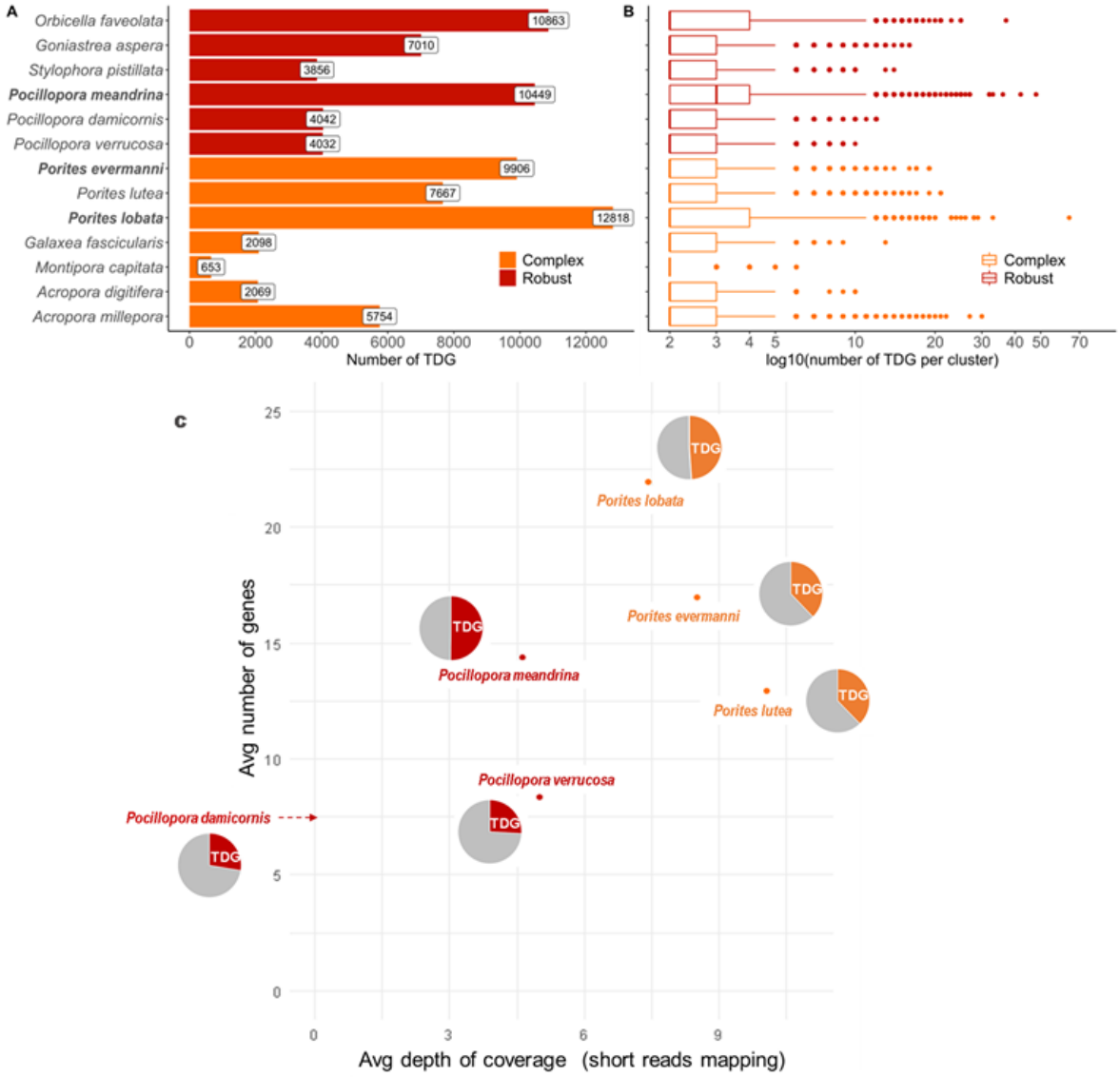

Figure 21 : Quantification des gènes dupliqués en tandem dans les génomes de coraux. A. Nombre de TDG détectés pour chaque espèce **B.** Distribution du nombre de gènes par cluster TDG. **C.** Pour 499 familles de gènes (orthogroupes (OG) avec >=10 gènes dans *P. meandrina* ou *P. lobata*), le nombre de gènes dans les espèces de *Pocillopora* et *Porites* est comparé à l'estimation du nombre de copies du gène à partir de l'alignement de lectures courtes (profondeur normalisée d'alignement de lectures sur le consensus de l'OG). Les camemberts représentent la proportion de gènes TDG dans chaque espèce. Pour *Pocillopora damicornis*, nous n'avons pas pu estimer le nombre de copies à partir des lectures car nous n'avons pas trouvé de jeu de lectures illumina à télécharger **(Noel*,Denoeud* et al, en préparation)**.

J'ai identifié les familles de gènes (OG) amplifiés chez les coraux par rapport aux anémones de mer, et j'ai observé que ces amplifications correspondent en très vaste majorité à des duplications en tandem (**Figure 22A**). Les fonctions des gènes amplifiés chez les coraux

correspondent principalement à des récepteurs extracellulaires, probablement impliqués dans l'immunité innée et/ou dans la relation avec le symbionte (Hamada, 2013)**.** Même si les fonctions amplifiées sont similaires, les orthogroupes amplifiés individuellement peuvent différer d'une espèce à l'autre : il semble que nous assistons donc ici à un phénomène d'évolution convergente, où l'amplification/diversification du répertoire de récepteurs extracellulaires a été sélectionnée, mais pas la présence d'un type de récepteur en particulier. Cette observation est réminiscente de ce que nous avons observé pour les N-méthyltransférases de caféier, théier et cacaoyer. On constate aussi des différences entre *Porites lobata* et *Pocillopora meandrina* : la plupart des orthogroupes amplifiés sont plus abondants chez *Porites lobata* que chez *Pocillopora meandrina* (**Figure 22B**). Une exception notable réside dans les gènes contenant le domaine EGF_CA (« Epidermal Growth Factor, CA binding »). Il s'agit de protéines liant le calcium, trouvées dans la matrice extracellulaire et potentiellement impliquées dans les processus de calcification (Wang, Zoccola *et al.* 2021). On peut émettre l'hypothèse que cette abondance est liée à la structure ramifiée (branchée) des colonies de *Pocillopora*, par opposition aux colonies massives de *Porites*. Par ailleurs, les espèces du genre *Porites* contiennent davantage de récepteurs liés à l'immunité innée que les autres espèces de coraux (**Figure 22C**). Il est possible que ce vaste répertoire soit impliqué dans la grande longévité et la résilience de espèces du genre *Porites* (Carili, 2012) comme cela a été proposé pour les gènes de résistance chez le chêne (Plomion, 2018).

Afin de mieux comprendre l'histoire évolutive des différentes familles de gènes amplifiés, j'ai calculé les taux de mutations synonymes (Ks) entre les paires de gènes dupliqués en tandem dans les orthogroupes. Les valeurs obtenues sont très variables entre les OG, et entre les TDG dans un même OG. Il apparaît donc que les duplications ont eu lieu à diverses périodes : il s'agit d'un processus dynamique, qui est probablement encore en cours.

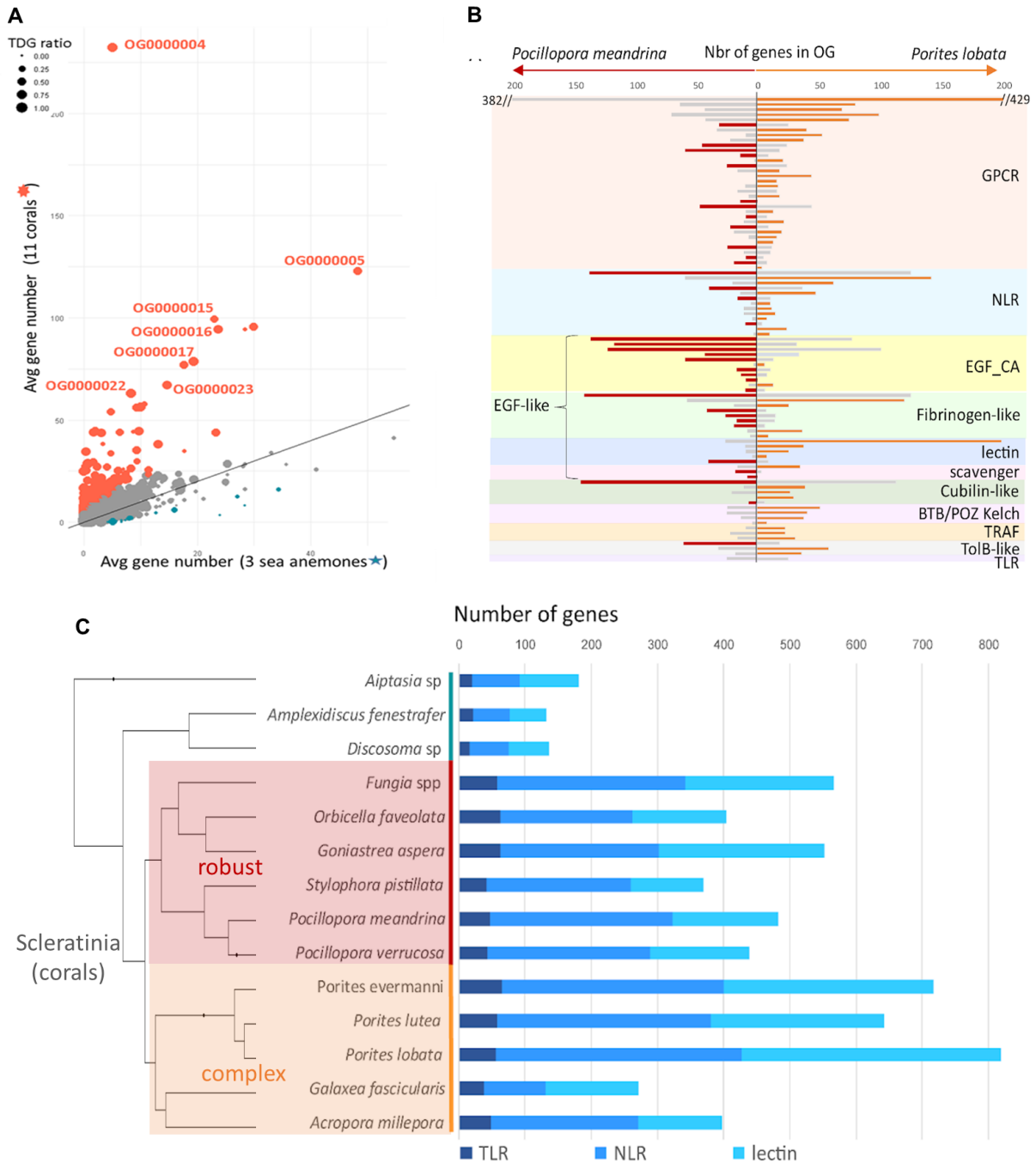

Figure 22 : Familles de gènes (OG) amplifiées chez les coraux. A. Nombre de copies chez les coraux par rapport aux anémones de mer. Les OG colorés en orange et en bleu ont significativement plus et moins de copies (respectivement) dans les coraux comparés aux anémones de mer (test binomial, P-value ajustée < 0.001). La taille des points correspond au ratio de TDG dans chaque OG pour 11 génomes de coraux. **B.** Bar-plot représentant le nombre de gènes dans les familles (OG) amplifiées dans les génomes de *Porites lobata* (à droite) et *Pocillopora meandrina* (à gauche). Les OG sont regroupés par catégories fonctionnelles. La barre correspondant à l'espèce contenant le plus de gènes est colorée en rouge lorsqu'il s'agit de *P. meandrina*, en orange lorsqu'il s'agit de *P. lobata.* **C.** Bar-plot cumulatif représentant le nombre de gènes récepteurs du système immunitaire identifiés à partir des domaines détectés par InterProScan dans 14 cnidaires, pour trois catégories de récepteurs (TLR=Toll-like receptors, NLR=NOD-like receptors) **(Noel*,Denoeud* et al, en préparation)**.

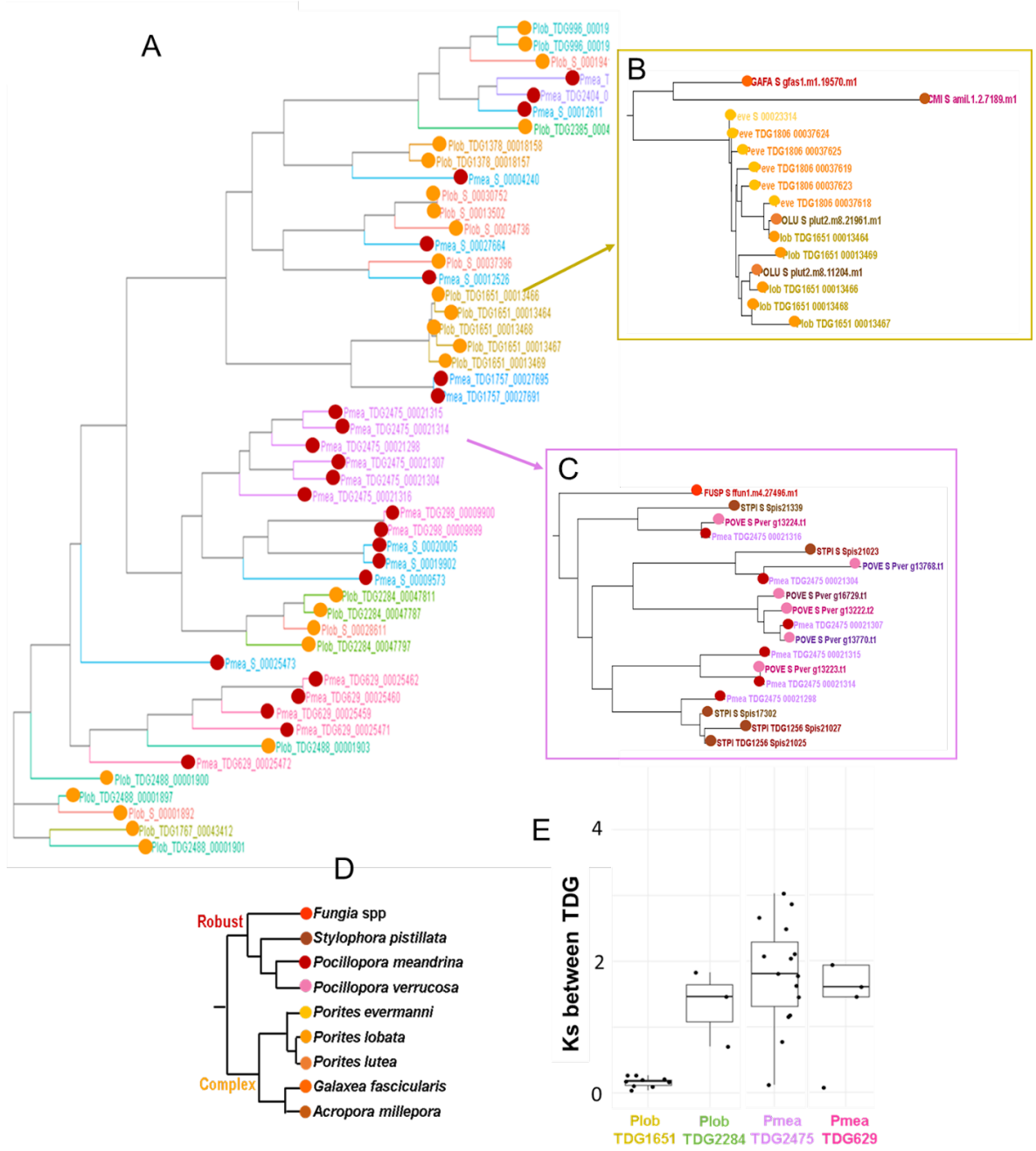

Figure 23. A. Arbre phylogénétique (Approximately maximum-likelihood) obtenu avec FastTree après l'alignement des protéines de l'orthogroupe OG0000106 (TIR domain-containing) dans *Porites lobata* (ronds orange) et *Pocillopora meandrina* (ronds rouges). Les couleurs correspondent aux clusters de gènes répétés en tandem (les singletons sont en rouge) **B,C:** Arbres obtenus par l'alignement des protéines de 15 espèces de coraux pour deux clusters TDG. Les couleurs des ronds correspondent à celles de l'arbre des espèces en **D. E:** Distribution des valeurs de Ks entre les paires de gènes appartenant aux différents clusters TDG, de *P. lobata* et *P. meandrina* **(Noel*,Denoeud* et al, en préparation)**.

La **Figure 23** présente un exemple de famille de gènes amplifiée par duplications en tandem chez les coraux (Toll-like receptor, contenant un domaine TIR, impliqué dans

l'immunité innée (Poole, 2014)). On constate que certaines duplications semblent précéder la divergence entre espèces robustes et complexes, et d'autres sont spécifiques de certaines espèces ou groupes d'espèces, comme les *Porites* (**Figure 23B**) ou les *Pocilloporidae* (**Figure 23C**). Comme attendu, les duplications plus anciennes (partagées par davantage d'espèces) ont des taux de mutations synonymes (Ks) supérieurs, par rapport aux paires de gènes dupliqués récemment au sein d'une seule espèce ou groupe d'espèce (**Figure 23E**).

Le fait que le processus soit dynamique (gains et pertes actuellement en cours) et que le nombre de copies soit variable d'une espèce à l'autre, avec différentes familles amplifiées dans différentes espèces, est en accord avec le modèle d'évolution des familles de gènes nommé "birth and death evolution" (Nei, 2005). La naissance de nouvelles copies dupliquées en tandem surviendrait par crossing-over inégal (**Figure 4A**). En effet, différentes observations montrent que les duplications ont lieu au niveau génomique : on observe plusieurs cas de duplication de plusieurs gènes adjacents ensemble, et la conservation entre les séquences dupliquées est visible également dans les introns, même si l'homologie est perdue plus vite (**Figure 24**).

Bien que l'on observe que le taux de conservation entre les copies dupliquées en tandem est plus faible pour les paires plus distantes sur le génome que pour les paires adjacentes, il est notable que certains gènes TDG très anciens (c'est-à-dire avec une grande divergence entre les copies) sont restés co-localisés sur le génome. Ce n'est pas le cas chez les plantes, par exemple le caféier où les NMT initialement dupliquées en tandem ont été transloquées sur d'autres chromosomes (Denoeud, 2014). Cette observation inhabituelle peut être mise en relation avec la forte conservation de la synténie entre les génomes de coraux. On peut en effet supposer que ces espèces subissent peu de remaniements chromosomiques. Ce faible taux de remaniements explique vraisemblablement pourquoi tant de gènes sont détectés comme dupliqués en tandem dans les coraux (à condition d'avoir des assemblages bien continus), alors que le signal est généralement perdu plus rapidement dans les autres génomes. Cela suggère que les duplications en tandem peuvent être sous-estimées dans les génomes soumis à de forts réarrangements.

Figure 24: Vue du génome browser représentant un cluster de 3 gènes dupliqués en tandem appartenant à l'OG0000023 (lectin-like domain). **B.** Comparaison des structures des gènes annotés : les exons sont en bleu et les introns en noir (les nombres représentent leurs longueurs). Les exons homologues sont liés par des lignées orangées. **C.** Alignement multiple d'une région contenant 2 exons et un intron, montrant une plus grande homologie dans les exons **(Noel*,Denoeud* et al, en préparation)**.

En outre, la présence de copies très peu conservées à proximité les unes des autres démontre que le phénomène d'évolution concertée, qui a été beaucoup remis en question (review par Eirín-López, 2012), n'a pas cours chez les coraux (ou en tout cas de façon très anecdotique). En effet, seulement une poignée de paires de gènes dupliqués ont une parfaite

conservation dans les introns, les autres ont rapidement accumulé des mutations dans les séquences introniques. Il semble donc que la translocation (conduisant aussi à la perte de synténie) n'est pas nécessaire pour permettre aux gènes d'échapper à l'homogénéisation et donc de diverger pour éventuellement acquérir de nouvelles fonctions. Un article décrivant ces analyses, ainsi que d'autres non mentionnées ici, est sur le point d'être soumis (Noel*, Denoeud*, *et al.* En préparation).

6/ Discussion

J'ai pu montrer au travers de divers exemples présentés dans ce document en quoi les remaniements structuraux des génomes peuvent avoir un impact sur les phénotypes. En particulier, les duplications de gènes, qu'elles proviennent de duplications complètes de génomes ou de duplications segmentales (en tandem) sont un levier d'évolution/diversification du répertoire de protéines et donc d'adaptation des espèces.

Les reliques de ces différents types de duplications ne sont pas toujours évidentes à distinguer car une espèce donnée peut avoir connu plusieurs types d'évènements et il est parfois compliqué de retracer leur histoire évolutive. Si l'on souhaite s'intéresser aux conséquences des duplications complètes de génomes, il est intéressant d'étudier des polyploïdes récents, comme le colza, ou encore mieux, des polyploïdes resynthétisés (colzas synthétiques). De tels modèles permettent d'étudier en particulier l'impact d'un évènement de polyploïdisation sur l'expression des gènes dupliqués, notamment de répondre aux questions « l'expression est-elle partagée entre les deux copies dupliquées de façon équitable ou en spécialisant chacune des copies (subfonctionnalisation) ? » ou encore « l'une des copies s'exprime-t-elle tandis que l'autre s'inactive (en voie de pseudogénisation) ? ». L'observation des évènements d'échanges homéologues entre les deux sous-génomes est également informative pour comprendre les mécanismes qui ont permis le succès des polyploïdes pour s'établir dans la nature. Le fait de disposer de génomes de référence de qualité pour les progéniteurs des colzas synthétiques est un atout pour ce type d'analyses (Belser, 2018) ; Rousseau-Gueutin, 2020).

Si l'on s'intéresse au devenir des gènes dupliqués en tandem sur le long terme, les génomes de coraux constituent un bon modèle, car ils ont la particularité d'avoir conservé l'organisation de leurs gènes dupliqués pendant une longue période. En effet, mon intuition

(toute personnelle, et discutable) est que les génomes de coraux ne sont pas davantage sujets aux duplications en tandem que d'autres génomes, mais que ces duplications y sont davantage visibles car les copies sont restées adjacentes, et que moins de duplicats ont été perdus (en particulier, les gènes du système immunitaire inné qui leur confèrent probablement un net avantage). Il serait intéressant de comprendre pourquoi les génomes de coraux sont moins soumis aux réarrangements que d'autres espèces. L'analyse plus fine de la synténie entre les génomes de coraux pourrait fournir des pistes pour répondre à cette question, et bénéficiera de la production d'assemblages de haute qualité et grande continuité pour d'autres génomes de coraux. Nous sommes actuellement en train d'assembler au Genoscope deux génomes de coraux des profondeurs. Je serai impliquée dans l'analyse de ces génomes, qui ont la particularité de ne pas dépendre d'un symbionte photosynthétique pour leur apport en énergie. Il sera intéressant de comparer leur répertoire de gènes immunitaires à ceux des coraux associés à des *Symbiodinium*.

Les quelques exemples cités plus haut montrent bien l'interrelation entre structure et fonction. Les modifications structurales des génomes sont souvent à l'origine de cascades d'événements complexes, aboutissant à la création de nouvelles espèces et à l'augmentation de la biodiversité. Nous avons désormais accès à des jeux de données concernant de nombreuses espèces à l'échelle d'écosystèmes entiers. Même si certaines de ces données sont partielles (courtes lectures, faible profondeur…), je suis convaincue qu'il est possible de les « faire parler ».

Partie III. Futur projet de recherche

Nous sommes entrés dans une ère où le séquençage longues lectures (Nanopore et PacBio) couplé à des technologies dites « long range », comme les carte optiques Bionano Genomics, et les données de Hi-C (chromosomal conformation capture), permet de viser des standards d'assemblage à l'échelle de chromosomes entiers. Ce degré d'exigence devient accessible même pour des projets très ambitieux (comme par exemple le projet ERGA « European Genome Atlas » qui prévoit de générer des génomes de référence de haute qualité pour l'ensemble de la biodiversité européenne). Comme je l'ai illustré par mes travaux, c'est par l'analyse de ces génomes de haute qualité que les observations les plus pertinentes peuvent être faites. Il est passionnant de s'apercevoir que chaque nouveau génome séquencé recèle de nouveaux mystères à élucider et peut nous livrer une information nouvelle sur un mécanisme évolutif encore mal compris. Les génomes et métagénomes déjà étudiés par le passé gagneraient d'ailleurs à être réexaminés régulièrement à l'aune de ces nouvelles découvertes. Les génomes de très haute qualité constituent une ressource précieuse en tant que référence pour l'analyse de jeux de données plus parcellaires, comme c'est le cas pour les génomes d'organismes non cultivables, ou dont l'ADN est difficile à extraire (par exemple les algues brunes) mais aussi pour les données de génomique environnementale (métagénomiques ou de type « genome skimming »). Au Genoscope, j'ai le privilège d'avoir accès à ces deux types de données, qui offrent des opportunités d'analyse très enthousiasmantes. Au cours de ma carrière scientifique, j'ai développé un intérêt central pour l'étude de l'évolution structurelle des génomes et des implications fonctionnelles et évolutives qui en découlent. En effet, j'ai eu la chance d'être aux premières loges pour étudier des génomes très variés, et j'ai constaté que leur structure actuelle recèle des reliques des événements passés. Chaque nouveau génome étudié a été l'occasion de nouvelles découvertes, souvent passionnantes. Je définirais mon rôle de chercheuse comme celui d'une « exploratrice de génomes ». Je m'estime très chanceuse de pouvoir jouer ce rôle, surtout en cette période où la recherche « fondamentale » vraiment « exploratoire » n'a plus vraiment le vent en poupe. La recherche pour moi consiste à faire des hypothèses et à imaginer des expériences pour les tester, certes, mais également à se laisser guider (j'appelle cela dérouler la pelote en tirant sur le fil) là où les résultats nous mènent. Les expériences peuvent avoir des résultats très différents de l'hypothèse de départ, et permettre de formuler une nouvelle hypothèse (puis une autre, …), qu'il faut ensuite tester. Dans le domaine de la bioinformatique, il y a peu de limites aux expériences que l'on peut faire pour tester des

hypothèses : il est somme-toute assez simple d'imaginer des méthodes pour répondre aux questions qui se posent (quand elles ne sont pas déjà développées) et de se donner les moyens de les mettre en œuvre. Il s'agit là aussi d'un avantage énorme que les analyses *in silico* procurent par rapport à la biologie expérimentale, où le design des expériences est bien plus complexe et coûteux. De nos jours, dans les demandes de financement de projets de recherche, il est souvent nécessaire d'exposer précisément ce que l'on va trouver. C'est une philosophie à laquelle je n'ai pas besoin d'adhérer pleinement, et je mesure ma chance. Lorsque je me lance dans un projet, j'ai une idée très claire de ce que je vais chercher (et de comment je vais m'y prendre), mais par contre, je ne sais pas à l'avance ce que je vais trouver, c'est là le côté stimulant de la chose ! Je souhaite transmettre cette façon de voir (ouverture d'esprit et optimisme) à de jeunes chercheurs : je déplore que les sources d'optimisme soient de plus en plus rares pour eux. Etant très attachée à cette vision exploratoire de la recherche, cela n'a pas été chose aisée pour moi, dans un premier temps, d'identifier un axe de recherche à long terme : j'espère en effet continuer à aller là où le « vent de mes nouvelles découvertes » me portera, sans idées préconçues. La rédaction de ce document m'a été très bénéfique car elle m'a permis d'énoncer clairement des thématiques de recherche qui transcendent les projets et les génomes particuliers, et vont constituer mes futures thématiques de recherche personnelles.

Ces axes que je propose cherchent à élucider les relations entre structure et fonction, en mettant à profit la masse de données de séquençage auxquelles j'ai accès. Je peux proposer deux axes principaux (il y en aura certainement d'autres, qui découleront de futures découvertes). Le premier est l'étude des régulations génomiques de la quantité de protéines. On s'est beaucoup focalisé sur les régulations transcriptionnelles (que ce soit au niveau de la synthèse d'ARN messager, ou de leur stabilité/dégradation) et post-transcriptionnelles (stabilité et dégradation des protéines) pour expliquer la présence de protéines en plus ou moins grande quantité dans les cellules. Cependant, l'importance du nombre de copies de gènes présentes sur le génome n'a pas beaucoup été mise en avant. Ce mécanisme semble cependant être le plus « simple » (voire simpliste ?) et il est probable que la sélection naturelle y joue un rôle important. En particulier il est frappant de constater que chez les plantes, les histones, une famille de protéines extrêmement conservée, c'est-à-dire extrêmement contrainte évolutivement « utilisent » possiblement ce mécanisme de régulation par nombre de copies génomiques. J'aimerais poursuivre des investigations dans ce domaine, et les domaines connexes. Mon approche de quantification de familles de gènes par alignement de

lectures ouvre de nombreuses perspectives d'étude, même pour des données fragmentaires (métagénomique, genome skimming). La question de la régulation du nombre de copies est extrêmement intéressante : est-ce seulement le jeu de la sélection ou d'autres mécanismes entrent-ils en jeu ? J'aimerais pouvoir essayer de répondre à cette question, en élargissant mon analyse à un ensemble de jeux de données allant au-delà des projets dans lesquels j'ai été impliquée directement.

A court terme, je pourrais tout d'abord mettre en évidence des familles de gènes amplifiées dans les plantes alpines échantillonnées par le projet PhyloAlps. Ce projet fera suite aux stages de M2 de Lina Alferkh et Daniel De Murat et pourrait être proposé à un nouvel étudiant. L'objectif sera d'identifier des familles de gènes dans lesquelles des régulations par amplification génomique peuvent survenir. Il s'agira de mesurer le nombre de copies de différentes familles de gènes (KO) dans les génomes (par l'approche d'alignement de lectures préalablement décrite). Des résultats prometteurs ont déjà été obtenus par Lina et Daniel : en particulier, les espèces de différentes branches phylogénétiques ont des signatures métaboliques spécifiques. Par la suite, il s'agira de rechercher des corrélations entre l'amplification génomique de certaines familles de gènes et le mode de vie des plantes (altitude, rayonnement UV, niveau de stress, forme de vie…). A ces fins, je prévois d'utiliser la base de données de traits élaborée par les collaborateurs du LECA sur le projet PhyloAlps, qui apporte des informations précises sur les échantillons prélevés, et sera je l'espère plus adaptée que la base « Flora indicativa » utilisée précédemment, qui s'est révélée trop imprécise. En outre, la phylogénie précise des plantes alpines obtenue grâce aux assemblages chloroplastiques provenant de ce projet de « genome skimming » permettra de chercher à identifier des phénomènes de co-évolution entre différents « traits » (les traits pouvant ici être les conditions de vie des espèces, le nombre de copies de certains gènes, ou même la taille des génomes) au long de l'arbre phylogénétique. Lorsque des gènes d'intérêt auront été identifiés, il serait intéressant d'aller plus loin en étudiant en détail les familles de gènes amplifiées, en particulier les variations entre les différentes copies des gènes (taux de mutations synonymes et non synonymes), ce qui permettra de comprendre les mécanismes d'évolution et de sélection survenus dans les différentes lignées. Dans ce but, il faudra développer des méthodes pour reconstituer les séquences de gènes d'intérêt (malgré la faible couverture de séquençage) par assemblage direct ou par alignement sur une séquence de référence. Nous pourrons ainsi étudier plus finement les mécanismes d'amplification de certaines familles de gènes, ainsi que les allèles fonctionnellement importants. Ce projet

apportera ainsi un nouvel éclairage sur l'évolution et l'adaptation des plantes alpines, en faisant des liens entre évènements de duplications géniques et l'adaptation aux conditions extrêmes de la haute montagne, et ce à l'échelle d'un biome végétal entier.

L'approche de quantification des familles de gènes par alignement de lectures courtes s'est aussi montrée utile pour valider l'annotation de nombreux gènes dupliqués en tandem dans les génomes de coraux que nous avons analysés. Je propose de l'appliquer aussi pour les génomes d'algues brunes séquencées dans le cadre du projet Phaeoexplorer. En effet, l'extraction d'ADN de haut poids moléculaire à partir des cellules visqueuses des algues brunes peut s'avérer très difficile, ce qui impacte grandement le rendement de séquençage longues lectures par Oxford Nanopore. Pour certaines espèces, nous n'avons pu obtenir que des lectures courtes (Illumina), et les répertoires de gènes annotés sur ces assemblages « draft » risquent d'être incomplets. Des analyses de familles de gènes amplifiées ou présentes/absentes parmi les algues brunes et avec les espèces sœurs vont être menées dans le cadre du projet Phaeoexplorer. Il serait intéressant de s'affranchir des biais d'assemblage et d'annotation pour quantifier les familles de gènes d'une façon homogène dans toutes les espèces (ou d'identifier des profils de présence/absence de certaines familles). Je prévois donc d'utiliser la méthode de quantification des familles par alignement de courtes lectures sur ce projet. J'accueille cette année un stagiaire de dernière année d'école d'ingénieur (Pierre Guenzi-Tibéri) qui va travailler sur l'amélioration de cette méthode, appliquée au projet PhyloAlps dans un premier temps. S'il le souhaite, il pourra poursuivre en thèse sur le projet Phaeoexplorer, pour lequel j'ai déposé une demande de financement.

A plus long terme, je souhaite appliquer cette approche pour des jeux de données plus conséquents. Je pense en effet que l'approche de quantification par alignement de courtes lectures, moyennant quelques développements, pourrait être particulièrement adaptée pour les jeux de données métagénomiques (et paléogénomiques). Parmi ces développements, il serait utile de mesurer les biais de quantification inhérents à cette méthode (en fonction de la distance phylogénétique entre les consensus utilisés et les lectures alignées, et de la quantité de domaines « non spécifiques » -présents dans plusieurs familles de gènes- par exemple), afin de l'adapter au mieux aux jeux de données à explorer. Il pourrait être intéressant pour certaines applications sur des génomes plus divergents de quantifier directement les domaines plutôt que les familles protéiques (en alignant par exemple les lectures sur la séquence de ces domaines ou sur leurs profils HMM). Je souhaiterais appliquer cette méthode aux échantillons métagénomiques collectés lors des projets TARA, ce qui pourrait permettre de quantifier des « fonctions » présentes/amplifiées à l'échelle

d'écosystèmes entiers. Le même type d'approche pourrait évidemment être appliqué à la détection et à la quantification d'éléments transposables dans tous types d'échantillons, en particulier pour les génomes pour lesquels on ne dispose pas d'un assemblage complet.

Je souhaite également approfondir mes recherches concernant la corrélation entre le nombre de copies de gènes d'histones et la taille des génomes. J'ai pu mettre cette relation en évidence chez les plantes, mais j'aimerais aussi étudier d'autres branches du Vivant. Une analyse préliminaire que j'ai menée en utilisant les banques de données publiques (KO de KEGG) a montré une légère corrélation entre le nombre de certaines protéines d'histones et la taille des génomes chez les animaux, mais celle-ci est beaucoup moins forte que chez les plantes. Il semble donc bien qu'un mécanisme différent soit en jeu pour la régulation des copies de protéines d'histones dans les cellules entre ces différents règnes. Il serait très intéressant d'investiguer l'origine de ces modes de régulation par l'analyse d'être vivants plus primitifs (organismes marins ?). Les méthodes que j'ai développées, de quantification de nombre de copies de protéines (dont les histones) ainsi que d'estimation de tailles des génomes, sont applicables dès lors que l'on dispose de courtes lectures et de séquences de référence pour les protéines à étudier. En effet, les histones ont l'avantage d'être des séquences extrêmement conservées, donc un alignement devrait être possible entre des lectures et leur consensus pour un large éventail d'espèces. De même, nous avons accès à des jeux de gènes conservés dans tous les eucaryotes (BUSCO) qui pourraient servir à calibrer la méthode d'estimation de la taille des génomes. Même si cette méthode nécessite que l'ensemble des lectures provienne d'un seul organisme et n'est donc pas directement applicable à des données métagénomiques, plusieurs approches ont été développées récemment pour reconstruire des génomes à partir de métagénomes. Jean-Marc Aury et Benjamin Istace ont soumis un projet ANR sur ce sujet, et mes collègues du LAGE ont généré des « MAG » (metagenome-assembled genomes) en catégorisant les contigs obtenus suite à l'assemblage de données métagénomiques (Delmont, 2021). Il est donc théoriquement possible d'estimer la quantité de copies d'histones et la taille du génome pour énormément d'organismes (tous ceux pour lesquels on dispose de courtes lectures, même à de faibles profondeurs). Ces estimations pourraient permettre de mettre au jour des corrélations entre taille de génomes et quantité d'histones pour différentes branches du vivant et de mieux comprendre les mécanismes évolutifs en jeu dans cette famille multigénique hors du commun.

Le second axe qui me tient à cœur est de retourner vers l'étude des introns. J'aimerais ici encore mettre à profit l'énorme quantité de données dont nous disposons, pour rechercher et caractériser les introns non canoniques et investiguer leurs mécanismes d'épissage. Dans un premier temps, je propose d'utiliser les données (méta)génomiques et (méta)transcriptomiques dont nous disposons au Genoscope pour un grand nombre d'eucaryotes marins (provenant par exemple des projets TARA) afin de rechercher des introns atypiques et de chercher à expliciter leurs origines et les mécanismes d'évolution des introns.

En particulier, j'aimerais tester des hypothèses que j'ai formulées lorsque j'ai observé les introns d'*Oikopleura dioica* : je propose l'existence d'un lien entre la présence d'un spliceosome « permissif » et l'invasion d'introns non-canoniques dans les génomes. J'ai également observé un lien entre les sites donneur et accepteurs des introns *d'O. dioica* (certains sites accepteurs étant préférentiellement associés à certains sites donneurs). Il semble donc que ces introns soient subi au mécanisme d' « intron definition » (dans lequel les jonctions d'épissages forment un pont à travers l'intron) qui a déjà été proposé comme le principal mode de reconnaissance des introns courts (Berget, 1995). Je souhaite généraliser ces observations à d'autres espèces.

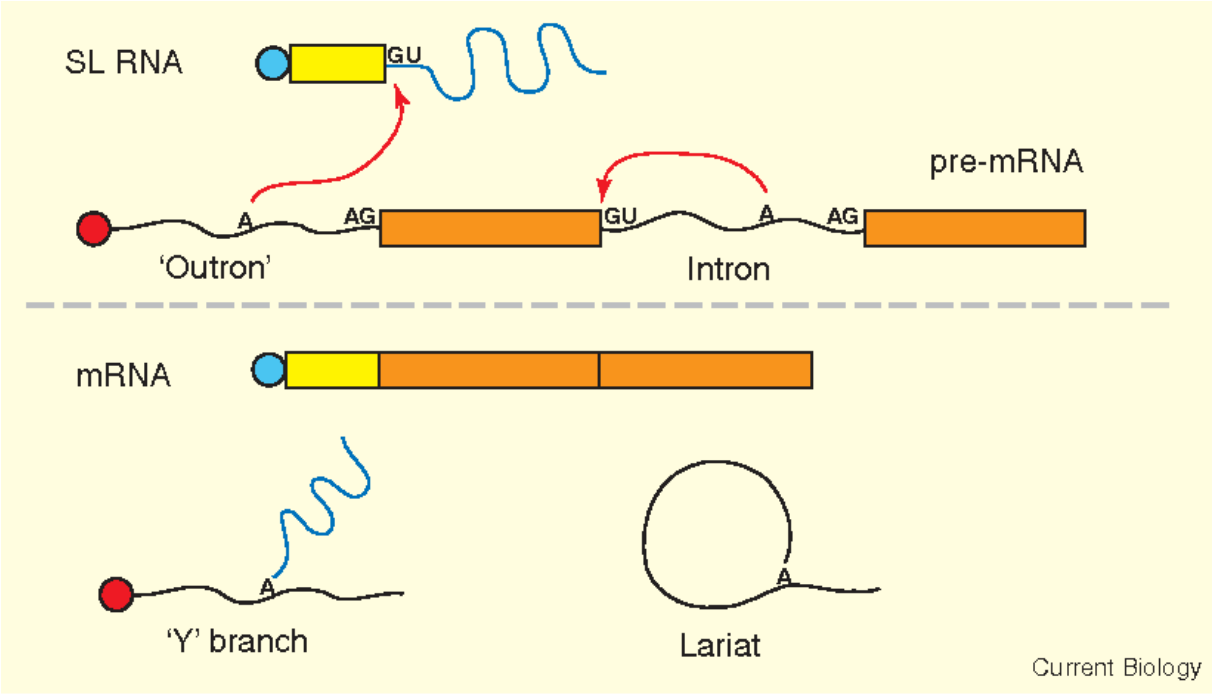

Figure 25 : mécanisme de « spliced leader trans-splicing » (d'après Stover, 2006).

Enfin, j'ai l'intuition que le mécanisme de « transsplicing », peut entrer en concurrence avec l'épissage d'introns en cis. Le transsplicing est un trans-épissage entre deux

molécules d'ARN, à la suite duquel un "spliced leader" est ajouté en 5' des ARN messagers **(Figure 25**). Ce phénomène est fréquent chez *Oikopleura diocia* (mais également chez le nématode *Caenorhabditis elegans* par exemple) et permet la résolution de transcrits polycistroniques. J'ai observé chez *O. dioica* que les gènes trans-épissés avaient des caractéristiques d'introns, en termes de sites d'épissage et de longueur, différentes des gènes non trans-épissés. L'alignement stringeant des séquences métatranscriptomiques sur les séquences métagénomiques, devrait permettre de caractériser les introns, ainsi que d'identifier des phénomènes de trans-épissage. Ces observations pourront ensuite être mises en relation pour vérifier à plus grande échelle la corrélation entre certaines caractéristiques d'introns et la présence de trans-épissage et de tester l'hypothèse d'une compétition entre les deux machineries d'épissage. Elle pourrait aussi apporter des réponses sur la question encore controversée de l'origine du trans-épissage : est-il apparu précocement dans l'évolution puis a été perdu dans certaines lignées « SL transsplicing early » ou bien est-il apparu indépendamment dans différentes lignées «SL transsplicing late » ? (Krchňáková, 2017). Enfin, la détection de spliced leaders dans des données métatranscriptomiques pourrait avoir une application pratique : celle d'assigner taxonomiquement les transcrits, en regroupant entre eux les transcrits partageant les mêmes spliced leader ou en utilisant les séquences de spliced leaders déjà connues (Bitar, 2013). Ces propositions ne sont pour l'instant que des pistes, qui nécessitent encore réflexion, mais je suis convaincue qu'elles pourraient donner lieu à des découvertes passionnantes.

Conclusion

En travaillant au Genoscope, j'ai eu la chance d'avoir accès à des données de séquençage pour des génomes très variés, qui m'ont donné l'occasion de répondre à des questions biologiques et évolutives très variées également. Certaines de ces questions étaient amenées par nos collaborateurs extérieurs, mais j'ai très souvent moi-même été à l'initiative des investigations que j'ai menées. Deux exemples marquants sont, d'une part, l'analyse des introns d'*Oikopleura*, que j'ai menée intégralement, en « tirant sur le fil » des introns non canoniques que j'ai observés dans cette espèce, dès le début de ma carrière au Genoscope. D'autre part, plus récemment, j'ai développé des sujets d'analyse personnels sur les séquences provenant du projet PhyloAlps (dont le but initial était d'obtenir les gènes chloroplastiques), visant à mettre à profit cette grande quantité de séquences pour dire quelque chose des génomes nucléaires et surtout de la fonction des gènes. J'ai pu montrer que la quantité d'histones et corrélée à la taille du génome monoploïde des plantes, ce qui n'avait jamais été décrit. Je suis dans l'attente de la soumission des données de séquençage pour publier un article sur ce sujet. En parallèle, je travaille avec un étudiant en dernière année d'école d'ingénieur sur la méthode de prédiction de tailles de génomes, qui peut être également appliquée pour la quantification de voies métaboliques, et nous publierons cette méthode avant la fin de son stage (septembre 2022). Entre ces accomplissements personnels, j'ai eu la chance de participer à des efforts collectifs, au sein de consortiums d'analyse ayant donné lieu à des publications de grande ampleur, en particulier pour différents génomes de plantes. J'ai ainsi développé mes compétences en gestion de projets, et acquis une grande expertise en analyse de génomes (au Genoscope, certains me voient comme l' « experte plantes » mais je me vois plutôt comme une « exploratrice de génomes »). Pour les raisons évoquées en introduction, inhérentes à mon profil particulier de chercheuse « exploratrice » au service de grands projets de séquençage, je suis longtemps restée dans la zone de transition entre les chercheurs « juniors » et « senior ». Pour être tout à fait honnête il faut que j'avoue que j'adore mettre les « mains dans le cambouis » et que cette position m'a longtemps bien convenu, en dépit de quelques frustrations passagères. Cependant, ces dernières années, j'ai beaucoup apprécié d'encadrer des stagiaires de M2, et je souhaite maintenant prendre davantage de responsabilités en termes d'encadrement. Je souhaiterais diriger des étudiants

en thèse afin de participer à la formation des chercheurs, et parce que je pense que j'ai des valeurs à leur apporter. Des valeurs scientifiques, de curiosité et de rigueur, et aussi des valeurs humaines, comme l'humilité et l'optimisme. Lorsque je dirigerai des thésards (puis des jeunes chercheurs), ma plus grande satisfaction, et mon sentiment de réussite, sera de constater qu'en fin de thèse ils sont capables d'explorer par eux-mêmes et de faire leurs propres découvertes, en débobinant le fil à leur façon, vers des questions (et des réponses !) que je n'aurais pas forcément choisi de creuser moi-même : que leur cheminement ne soit pas strictement identique à celui que j'aurais suivi à leur place. C'est cette diversité de points de vue qui enrichira la Recherche. Evidemment, je ne manquerai pas d'idées pour proposer une ou plusieurs directions à suivre. Mais réussir à laisser l'autonomie au jeune chercheur est je pense bien plus délicat à faire, et c'est ce défi-là que je souhaite relever, et ce rôle-là que je pense devoir jouer dans la suite de ma carrière scientifique lorsque j'aurai obtenu mon HDR. Je pourrai ainsi former de nouveaux « explorateurs » qui pourront à ma suite continuer à investiguer les implications fonctionnelles et évolutives de modifications structurales des génomes.

Références

*NB : ci-dessous sont listées les références externes, ma liste de publications étant déjà disponible à la section « **liste des publications** », p69.*

Alsos (2020) : voir ma liste de publications

Alsos (soumis) : voir ma liste de publications

Amoutzias, G. D., He, Y., Gordon, J., Mossialos, D., Oliver, S. G., & Van de Peer, Y. (2010). Posttranslational regulation impacts the fate of duplicated genes. *Proceedings of the National Academy of Sciences*, *107*(7), 2967-2971.

Andersen, E. J., Nepal, M. P., Purintun, J. M., Nelson, D., Mermigka, G., & Sarris, P. F. (2020). Wheat disease resistance genes and their diversification through integrated domain fusions. *Frontiers in genetics*, 898.

Aury, J. M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B. M., ... & Wincker, P. (2006). Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature*, *444*(7116), 171-178.

Arrigo, N., & Barker, M. S. (2012). Rarely successful polyploids and their legacy in plant genomes. *Current opinion in plant biology*, *15*(2), 140-146.

Ballington, J. R. (2008, July). The role of interspecific hybridization in blueberry improvement. In *IX International Vaccinium Symposium 810* (pp. 49-60).

Beaulieu, J. M., Moles, A. T., Leitch, I. J., Bennett, M. D., Dickie, J. B., & Knight, C. A. (2007). Correlated evolution of genome size and seed mass. *New Phytologist*, *173*(2), 422.

Belser (2018) : voir ma liste de publications

Bennett, M. D., & Leitch, I. J. (2005). Plant genome size research: a field in focus. *Annals of botany*, *95*(1), 1-6.

Bennetzen J, Ma J, Devos KM. (2005). Mechanisms of recent genome size variation in flowering plants. *Annals of Botany* 95: 127–132.

Berget, S. M. (1995). Exon Recognition in Vertebrate Splicing (∗). *Journal of biological Chemistry*, *270*(6), 2411-2414.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., ... & Guiguen, Y. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature communications*, *5*(1), 1-10.

Bitar, M., Boroni, M., Macedo, A. M., Machado, C. R., & Franco, G. R. (2013). The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles. *Frontiers in genetics*, *4*, 199.

Birchler, J. A., Riddle, N. C., Auger, D. L., & Veitia, R. A. (2005). Dosage balance in gene regulation: biological implications. *Trends in Genetics*, *21*(4), 219-226.

Blanc, G., & Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The plant cell*, *16*(7), 1667-1678.

Brodie, J., Chan, C. X., De Clerck, O., Cock, J. M., Coelho, S. M., Gachon, C., ... & Bhattacharya, D. (2017). The algal revolution. *Trends in plant science*, *22*(8), 726-738.

Brown, D. D., Wensink, P. C., & Jordan, E. (1972). A comparison of the ribosomal DNA's of Xenopus laevis and Xenopus mulleri: the evolution of tandem genes. *Journal of molecular biology*, *63*(1), 57-73.

Buitrago-López, C., Mariappan, K. G., Cárdenas, A., Gegner, H. M., & Voolstra, C. R. (2020). The genome of the cauliflower coral Pocillopora verrucosa. *Genome biology and evolution*, *12*(10), 1911-1917.

Buschmann, A. H., Camus, C., Infante, J., Neori, A., Israel, Á., Hernández-González, M. C., ... & Critchley, A. T. (2017). Seaweed production: overview of the global state of exploitation, farming and emerging research activity. *European Journal of Phycology*, *52*(4), 391-406.

Buggs, R. J., Zhang, L., Miles, N., Tate, J. A., Gao, L., Wei, W., ... & Soltis, D. E. (2011). Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. *Current Biology*, *21*(7), 551-556.

Carilli, J., Donner, S. D., & Hartmann, A. C. (2012). Historical temperature variability affects coral response to heat stress. *PloS one*, *7*(3), e34418.

Casneuf, T., De Bodt, S., Raes, J., Maere, S., & Van de Peer, Y. (2006). Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana. *Genome biology*, *7*(2), 1-11.

Chalhoub (2014) : voir la liste de mes publications

Cock, J. M., Sterck, L., Rouzé, P., Scornet, D., Allen, A. E., Amoutzias, G., ... & Wincker, P. (2010). The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature*, *465*(7298), 617-621.

Cock, J. M., Peters, A. F., & Coelho, S. M. (2011). Brown algae. *Current Biology*, *21*(15), R573-R575.

Cock, J. M., Godfroy, O., Macaisne, N., Peters, A. F., & Coelho, S. M. (2014). Evolution and regulation of complex life cycles: a brown algal perspective. *Current opinion in plant biology*, *17*, 1-6.

Coelho, S. M., Mignerot, L., & Cock, J. M. (2019). Origin and evolution of sex-determination systems in the brown algae. *New Phytologist*, *222*(4), 1751-1756.

Coissac, E., Hollingsworth, P. M., Lavergne, S., & Taberlet, P. (2016). From barcodes to genomes: extending the concept of DNA barcoding.

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature reviews genetics*, *6*(11), 836-846.

Cunning, R., Bay, R. A., Gillette, P., Baker, A. C., & Traylor-Knowles, N. (2018). Comparative analysis of the Pocillopora damicornis genome highlights role of immune system in coral evolution. *Scientific reports*, *8*(1), 1-10.

De Bodt, S., Maere, S., & Van de Peer, Y. (2005). Genome duplication and the origin of angiosperms. *Trends in ecology & evolution*, *20*(11), 591-597.

Dehal, P., & Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS biology*, *3*(10), e314.

Delmont, T. O., Gaia, M., Hinsinger, D. D., Fremont, P., Vanni, C., Guerra, A. F., ... & Jaillon, O. (2021). Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. *BioRxiv*, 2020-10.

Dharia, A. (2015). Investigating the Role of Gene Duplication in Ribosomal Protein Evolution and Testing a Model of Duplicate Gene Retention in Mammals.

D'hont (2012) : voir ma liste de publications

Draizen, E. J., Shaytan, A. K., Mariño-Ramírez, L., Talbert, P. B., Landsman, D., & Panchenko, A. R. (2016). HistoneDB 2.0: a histone database with variants—an integrated resource to explore histones and their variants. *Database*, *2016*.

Dubcovsky, J., & Dvorak, J. (2007). Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science*, *316*(5833), 1862-1866.

Edger, P. P., Smith, R., McKain, M. R., Cooley, A. M., Vallejo-Marin, M., Yuan, Y., ... & Puzey, J. R. (2017). Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *The Plant Cell*, *29*(9), 2150-2167.

Eirín-López, J. M., Rebordinos, L., Rooney, A. P., & Rozas, J. (2012). The birth-and-death evolution of multigene families revisited. *Repetitive DNA*, *7*, 170-196.

Farhat S., Florent I., Noel B., Kayal E., Da Silva C., Bigeard E., Alberti A., Labadie K., Corre E., Aury JM, Rombauts S., Wincker P., Guillou L and Porcel BM. Expression Analysis Highlights the Infection Processes of Two Amoebophrya Strains. *Front Microbiol. 2018 Oct 2*;9:2251.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, *151*(4), 1531-1545.

Freeling, M., & Thomas, B. C. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome research*, *16*(7), 805-814.

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual review of plant biology*, *60*, 433-453.

Gadgil, R., Barthelemy, J., Lewis, T., & Leffak, M. (2017). Replication stalling and DNA microsatellite instability. *Biophysical chemistry*, *225*, 38-48.

Glasauer, S. M., & Neuhauss, S. C. (2014). Whole-genome duplication in teleost fishes and its evolutionary consequences. *Molecular genetics and genomics*, *289*(6), 1045-1060.

Greilhuber, J., DOLEŽEL, J., Lysák, M. A., & Bennett, M. D. (2005). The origin, evolution and proposed stabilization of the terms 'genome size'and 'C-value'to describe nuclear DNA contents. *Annals of botany*, *95*(1), 255-260.

Hamada, M., Shoguchi, E., Shinzato, C., Kawashima, T., Miller, D. J., & Satoh, N. (2013). The complex NOD-like receptor repertoire of the coral Acropora digitifera includes novel domain combinations. *Molecular biology and evolution*, *30*(1), 167-176.

Hao, L., & Nei, M. (2004). Genomic organization and evolutionary analysis of Ly49 genes encoding the rodent natural killer cell receptors: rapid evolution by repeated gene duplication. *Immunogenetics*, *56*(5), 343-354.

Harrow (2006) :voir ma liste de publications

Hood, L., Campbell, J. H., & Elgin, S. C. R. (1975). The organization, expression, and evolution of antibody genes and other multigene families. *Annual review of genetics*, *9*(1), 305-353.

Hughes, A. L., & Nei, M. (1989). Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proceedings of the National Academy of Sciences*, *86*(3), 958-962.

Hughes, G. M., Boston, E. S., Finarelli, J. A., Murphy, W. J., Higgins, D. G., & Teeling, E. C. (2018). The birth and death of olfactory receptor gene families in mammalian niche adaptation. *Molecular biology and evolution*, *35*(6), 1390-1406.

Hurgobin, B., Golicz, A. A., Bayer, P. E., Chan, C. K. K., Tirnaz, S., Dolatabadian, A., ... & Edwards, D. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. *Plant biotechnology journal*, *16*(7), 1265-1274.

Ingram, V. M. (1961). Gene evolution and the haemoglobins. *Nature*, *189*(4766), 704-708.

Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., Cassagrande, A., ... & Wincker, P. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *nature*, *449*(7161), 463-7.

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J. E., McKain, M. R., McNeal, J., ... & Claude, W. D. (2012). A genome triplication associated with early diversification of the core eudicots. *Genome biology*, *13*(1), 1-14.

Jiao, Y., Li, J., Tang, H., & Paterson, A. H. (2014). Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *The Plant Cell*, *26*(7), 2792-2802.

Kellis, M., Birren, B. W., & Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature*, *428*(6983), 617-624.

Knowlton, N., & Rohwer, F. (2003). Multispecies microbial mutualisms on coral reefs: the host as a habitat. *the american naturalist*, *162*(S4), S51-S62.

Kozik, A., Rowan, B. A., Lavelle, D., Berke, L., Schranz, M. E., Michelmore, R. W., & Christensen, A. C. (2019). The alternative reality of plant mitochondrial DNA: One ring does not rule them all. *PLoS genetics*, *15*(8), e1008373.

Krchňáková, Z., Krajčovič, J., & Vesteg, M. (2017). On the possibility of an early evolutionary origin for the spliced leader trans-splicing. *Journal of Molecular Evolution*, *85*(1), 37-45.

Langham, R. J., Walsh, J., Dunn, M., Ko, C., Goff, S. A., & Freeling, M. (2004). Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics*, *166*(2), 935-945.

Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., Deyholos, M. K., Gitzendanner, M. A., Graham, S. W., ... & Li, X. (2019). One thousand plant transcriptomes and the phylogenomics of green plants.

Leitch, A. R., & Leitch, I. J. (2008). Genomic plasticity and the diversity of polyploid plants. *Science*, *320*(5875), 481-483.

Li, W. H., Yang, J., & Gu, X. (2005). Expression divergence between duplicate genes. *TRENDS in Genetics*, *21*(11), 602-607.

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *science*, *290*(5494), 1151-1155.

Marzluff, W. F., Wagner, E. J., & Duronio, R. J. (2008). Metabolism and regulation of canonical histone mRNAs: life without a poly (A) tail. *Nature Reviews Genetics*, *9*(11), 843-854.

Masterson, J. (1994). Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science*, *264*(5157), 421-424.

Mayrose, I., Zhan, S. H., Rothfels, C. J., Magnuson-Ford, K., Barker, M. S., Rieseberg, L. H., & Otto, S. P. (2011). Recently formed polyploid plants diversify at lower rates. *Science*, *333*(6047), 1257-1257.

McKain, M. R., Johnson, M. G., Uribe-Convers, S., Eaton, D., & Yang, Y. (2018). Practical considerations for plant phylogenomics. *Applications in plant sciences*, *6*(3), e1038.

McSteen, P. (2010). Auxin and monocot development. *Cold Spring Harbor perspectives in biology*, *2*(3), a001479.

Murat, F., Armero, A., Pont, C., Klopp, C., & Salse, J. (2017). Reconstructing the genome of the most recent common ancestor of flowering plants. *Nature genetics*, *49*(4), 490.

Nei, M., Gu, X., & Sitnikova, T. (1997). Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences*, *94*(15), 7799-7806.

Nei, M., Rogozin, I. B., & Piontkivska, H. (2000). Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proceedings of the National Academy of Sciences*, *97*(20), 10866-10871.

Nei, M., & Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.*, *39*, 121-152.

Niimura, Y., & Nei, M. (2003). Evolution of olfactory receptor genes in the human genome. *Proceedings of the National Academy of Sciences*, *100*(21), 12235-12240.

Ohno, S. (2013). *Evolution by gene duplication*. Springer Science & Business Media.

Ota, T., & Nei, M. (1994). Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. *Molecular biology and evolution*, *11*(3), 469-482.

Otto, S. P. (2007). The evolutionary consequences of polyploidy. *Cell*, *131*(3), 452-462.

Panchy, N., Lehti-Shiu, M., & Shiu, S. H. (2016). Evolution of gene duplication in plants. *Plant physiology*, *171*(4), 2294-2316.

Pandit, M. K., White, S. M., & Pocock, M. J. (2014). The contrasting effects of genome size, chromosome number and ploidy level on plant invasiveness: a global analysis. *New Phytologist*, *203*(2), 697-703.

Pâques, F., Leung, W. Y., & Haber, J. E. (1998). Expansions and contractions in a tandem repeat induced by double-strand break repair. *Molecular and cellular biology*, *18*(4), 2045-2054.

Piontkivska, H., Rooney, A. P., & Nei, M. (2002). Purifying selection and birth-and-death evolution in the histone H4 gene family. *Molecular biology and evolution*, *19*(5), 689-697.

Planes, S., Allemand, D., Agostini, S., Banaigs, B., Boissin, E., Boss, E., ... & Tara Pacific Consortium. (2019). The Tara Pacific expedition—A pan-ecosystemic approach of the "-omics" complexity of coral reef holobionts across the Pacific Ocean. *PLoS biology*, *17*(9), e3000483.

Plomion, C., Aury, J. M., Amselem, J., Leroy, T., Murat, F., Duplessis, S., ... & Salse, J. (2018). Oak genome reveals facets of long lifespan. *Nature Plants*, *4*(7), 440-452.

Pogoreutz, C., Voolstra, C. R., Rädecker, N., Weis, V., Cardenas, A., & Raina, J. B. (2020). The coral holobiont highlights the dependence of cnidarian animal hosts on their associated microbes. In *Cellular dialogues in the holobiont* (pp. 91-118). CRC Press.

Poole, A. Z., & Weis, V. M. (2014). TIR-domain-containing protein repertoire of nine anthozoan species reveals coral–specific expansions and uncharacterized proteins. *Developmental & Comparative Immunology*, *46*(2), 480-488.

Pouchon (2022) : voir ma liste de publications

Ren, R., Wang, H., Guo, C., Zhang, N., Zeng, L., Chen, Y., ... & Qi, J. (2018). Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. *Molecular Plant*, *11*(3), 414-428.

Roman, I., Stănilă, A., & Stănilă, S. (2013). Bioactive compounds and antioxidant activity of Rosa canina L. biotypes from spontaneous flora of Transylvania. *Chemistry Central Journal*, *7*(1), 73.

Rooney, A. P., Piontkivska, H., & Nei, M. (2002). Molecular evolution of the nontandemly repeated genes of the histone 3 multigene family. *Molecular biology and evolution*, *19*(1), 68-75.

Rooney, A. P. (2004). Mechanisms underlying the evolution and maintenance of functionally heterogeneous 18S rRNA genes in apicomplexans. *Molecular Biology and Evolution*, *21*(9), 1704-1711.

Rousseau-Gueutin (2020): voir ma liste de publications

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., ... & Jackson, S. A. (2010). Genome sequence of the palaeopolyploid soybean. *nature*, *463*(7278), 178-183.

Schranz, M. E., Mohammadin, S., & Edger, P. P. (2012). Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Current opinion in plant biology*, *15*(2), 147-153.

Schwager, E. E., Sharma, P. P., Clarke, T., Leite, D. J., Wierschin, T., Pechmann, M., ... & McGregor, A. P. (2017). The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC biology*, *15*(1), 1-27.

Shinzato, C., Shoguchi, E., Kawashima, T., Hamada, M., Hisata, K., Tanaka, M., ... & Satoh, N. (2011). Using the Acropora digitifera genome to understand coral responses to environmental change. *Nature*, *476*(7360), 320-323.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210-3212.

Smyčka (sous presse) : voir ma liste de publications

Soltis, D. E., Soltis, P. S., Bennett, M. D., & Leitch, I. J. (2003). Evolution of genome size in the angiosperms. *American Journal of Botany*, *90*(11), 1596-1603.

Soltis, P. S., & Soltis, D. E. (2016). Ancient WGD events as drivers of key innovations in angiosperms. *Current opinion in plant biology*, *30*, 159-165.

Stebbins, G. L. (1950). *Variation and evolution in plants*. Columbia University Press.

Stover, N. A., Kaye, M. S., & Cavalcanti, A. R. (2006). Spliced leader trans-splicing. *Current Biology*, *16*(1), R8-R9.

Tank, D. C., Eastman, J. M., Pennell, M. W., Soltis, P. S., Soltis, D. E., Hinchliff, C. E., ... & Harmon, L. J. (2015). Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytologist*, *207*(2), 454-467.

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., ... & Rokhsar, D. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *science*, *313*(5793), 1596-1604.

Vanneste, K., Baele, G., Maere, S., & Van de Peer, Y. (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome research*, *24*(8), 1334-1347.

Visger, C. J., Germain-Aubrey, C. C., Patel, M., Sessa, E. B., Soltis, P. S., & Soltis, D. E. (2016). Niche divergence between diploid and autotetraploid Tolmiea. *American Journal of Botany*, *103*(8), 1396-1406.

Voolstra, C. R., Li, Y., Liew, Y. J., Baumgarten, S., Zoccola, D., Flot, J. F., ... & Aranda, M. (2017). Comparative analysis of the genomes of Stylophora pistillata and Acropora digitifera provides evidence for extensive differences between species of corals. *Scientific reports*, *7*(1), 1-14.

Wang, X., Zoccola, D., Liew, Y. J., Tambutte, E., Cui, G., Allemand, D., ... & Aranda, M. (2021). The evolution of calcification in reef-building corals. *Molecular biology and evolution*, *38*(9), 3543-3555.

Wang (2021) : voir ma liste de publications

Wolfe, K. H., & Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, *387*(6634), 708-713.

Wu, S., Han, B., & Jiao, Y. (2020). Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. *Molecular Plant*, *13*(1), 59-71.

Ye, N., Zhang, X., Miao, M., Fan, X., Zheng, Y., Xu, D., ... & Zhao, F. (2015). Saccharina genomes provide novel insight into kelp biology. *Nature communications*, *6*(1), 1-11.

Ying, H., Cooke, I., Sprungala, S., Wang, W., Hayward, D. C., Tang, Y., ... & Miller, D. J. (2018). Comparative genomics reveals the distinct evolutionary trajectories of the robust and complex coral lineages. *Genome biology*, *19*(1), 1-24.

Ying, H., Hayward, D. C., Cooke, I., Wang, W., Moya, A., Siemering, K. R., ... & Miller, D. J. (2019). The whole-genome sequence of the coral Acropora millepora. *Genome biology and evolution*, *11*(5), 1374-1379.

Yokoyama, S., & Yokoyama, R. (1989). Molecular evolution of human visual pigment genes. *Molecular Biology and Evolution*, *6*(2), 186-197.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in ecology & evolution*, *18*(6), 292-298.

Zhou, B., Li, Y., Xu, Z., Yan, H., Homma, S., & Kawabata, S. (2007). Ultraviolet A-specific induction of anthocyanin biosynthesis in the swollen hypocotyls of turnip (Brassica rapa). Journal of experimental botany, 58(7), 1771-1781.

Zimmer, E. A., Martin, S. L., Beverley, S. M., Kan, Y. W., & Wilson, A. C. (1980). Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proceedings of the National Academy of Sciences*, *77*(4), 2158-2162.

Zoratti, L., Jaakola, L., Häggman, H., & Giongo, L. (2015). Anthocyanin profile in berries of wild and cultivated Vaccinium spp. along altitudinal gradients in the Alps. *Journal of agricultural and food chemistry*, *63*(39), 8641-8650.

ANNEXES

Liste des publications/communications

(ORCID : https://orcid.org/0000-0001-8819-7634)

Vergnaud G, Denoeud F. **Minisatellites: mutability and genome architecture.** Genome Res 2000 Jul;10(7):899-907.

Le Flèche P, Hauck Y, Onteniente L, Prieur A, Denoeud F, Ramisse V, Sylvestre P, Benson G, Ramisse F, Vergnaud G,.**A tandem repeats database for bacterial genomes: application to the genotyping of Yersinia pestis and Bacillus anthracis.** BMC Microbiol 2001;1(1):2.

Le Flèche P, Fabre M, Denoeud F, Koeck JL, Vergnaud G.**High resolution, on-line identification of strains from the Mycobacterium tuberculosis complex based on tandem repeat typing.** BMC Microbiol 2002 Nov 27;2(1):37.

Denoeud F, Vergnaud G, Benson G. **Predicting human minisatellites polymorphism** Genome Research 2003 May;13(5):856-867.

Denoeud F, Vergnaud G. **Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains : a web-based resource.** BMC Bioinformatics 2004 Jan ; 5:4.

Ramisse V, Houssu P, Hernandez E, Denoeud F, Hilaire V, Lisanti O, Ramisse F, Cavallo JD, Vergnaud G. **Variable number of tandem repeats in Salmonella enterica subsp. enterica for typing purposes.** J Clin Microbiol. 2004 Dec;42(12):5722-30.

ENCODE Project Consortium. **The ENCODE (ENCyclopedia Of DNA Elements) Project.** Science. 2004 Oct 22;306(5696):636-40.

Guigó R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E, Castelo R, Eyras E, Ucla C, Gingeras TR, Harrow J, Hubbard T, Lewis SE, Reese MG. **EGASP: the human ENCODE Genome Annotation Assessment Project.** Genome Biol. 2006;7 Suppl 1:S2.1-31. Epub 2006 Aug 7. Review.

Harrow J*, Denoeud F*, Frankish A*, Reymond A*, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R. **GENCODE: producing a reference annotation for ENCODE.** Genome Biol. 2006;7 Suppl 1:S4.1-9. Epub 2006 Aug 7.

Tress ML, Martelli PL, Frankish A, Reeves GA, Wesselink JJ, Yeats C, Olason PL, Albrecht M, Hegyi H, Giorgetti A, Raimondo D, Lagarde J, Laskowski RA, López G, Sadowski MI, Watson JD, Fariselli P, Rossi I, Nagy A, Kai W, Størling Z, Orsini M, Assenov Y, Blankenburg H, Huthmacher C, Ramírez F, Schlicker A, Denoeud F, Jones P, Kerrien S, Orchard S, Antonarakis SE, Reymond A, Birney E, Brunak S, Casadio R, Guigo R, Harrow J, Hermjakob H, Jones DT, Lengauer T, Orengo CA, Patthy L, Thornton JM, Tramontano A, Valencia A. **The implications of alternative splicing in the ENCODE protein complement.** Proc Natl Acad Sci U S A. 2007 Mar 27;104(13):5495-500. Epub 2007 Mar 19.

Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J, Dike S, Wyss C, Henrichsen CN, Holroyd N, Dickson MC, Taylor R, Hance Z, Foissac S, Myers RM, Rogers J, Hubbard T, Harrow J, Guigó R, Gingeras TR, Antonarakis SE, Reymond A. **Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions.** Genome Res. 2007 Jun;17(6):746-59.

Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo SW, Lu Y, Denoeud F, Antonarakis SE, Snyder M, Ruan Y, Wei CL, Gingeras TR, Guigó R, Harrow J, Gerstein MB. **Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution.** Genome Res. 2007 Jun;17(6):839-51.

Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermüller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A, Ucla C, Wyss C, Antonarakis SE, Denoeud F, Lagarde J, Drenkow J, Kapranov P, Gingeras TR, Guigó R, Snyder M, Gerstein MB, Reymond A, Hofacker IL, Stadler PF. **Structured RNAs in the ENCODE selected regions of the human genome.** Genome Res. 2007 Jun;17(6):852-64.

ENCODE Project Consortium. **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** Nature. 2007 Jun 14;447(7146):799-816.

Djebali S, Kapranov P, Foissac S, Lagarde J, Reymond A, Ucla C, Wyss C, Drenkow J, Dumais E, Murray RR, Lin C, Szeto D, Denoeud F, Calvo M, Frankish A, Harrow J, Makrythanasis P, Vidal M, Salehi-Ashtiani K, Antonarakis SE, Gingeras TR, Guigó R. **Efficient targeted transcript discovery via array-based normalization of RACE libraries.** Nat Methods. 2008 May 25.

Makrythanasis P, Kapranov P, Bartoloni L, Reymond A, Deutsch S, Guigó R, Denoeud F, Drenkow J, Rossier C, Ariani F, Capra V, Excoffier L, Renieri A, Gingeras TR, Antonarakis SE. **Variation in novel exons (RACEfrags) of the MECP2 gene in Rett syndrome patients and controls.** Hum Mutat. 2009 Sep;30(9):E866-79.

Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M,Valle G, Wincker P, Scarpelli C, Jaillon O, Artiguenave F. **Annotating genomes with massive-scale RNA sequencing.** Genome Biol. 2008;9(12):R175.

Martin F, Kohler A, Murat C, Balestrini R, Coutinho PM, Jaillon O, Montanini B, Morin E, Noel B, Percudani R, Porcel B, Rubini A, Amicucci A, Amselem J,Anthouard V, Arcioni S, Artiguenave F, Aury JM, Ballario P, Bolchi A, Brenna A, Brun A, Buée M, Cantarel B, Chevalier G, Couloux A, Da Silva C, Denoeud F, Duplessis S, Ghignone S, Hilselberger B, Iotti M, Marçais B, Mello A, Miranda M, Pacioni G, Quesneville H, Riccioni C, Ruotolo R, Splivallo R, Stocchi V, Tisserant E, Viscomi AR, Zambonelli A, Zampieri E, Henrissat B, Lebrun MH, Paolocci F, Bonfante P, Ottonello S, Wincker P. **Périgord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis.** Nature. 2010 Apr 15;464(7291):1033-8.

Denoeud F, Henriet S, Mungpakdee S, Aury JM, Da Silva C, Brinkmann H, Mikhaleva J, Olsen LC, Jubin C, Cañestro C, Bouquet JM, Danks G, Poulain J, Campsteijn C, Adamski M, Cross I, Yadetie F, Muffato M, Louis A, Butcher S, Tsagkogeorga G, Konrad A, Singh S, Jensen MF, Huynh Cong E, Eikeseth-Otteraa H, Noel B, Anthouard V, Porcel BM, Kachouri-Lafond R, Nishino A, Ugolini M, Chourrout P, Nishida H, Aasland R, Huzurbazar S, Westhof E, Delsuc F, Lehrach H, Reinhardt R, Weissenbach J, Roy SW,

Artiguenave F, Postlethwait JH, Manak JR, Thompson EM, Jaillon O, Du Pasquier L, Boudinot P, Liberles DA, Volff JN, Philippe H, Lenhard B, Roest Crollius H, Wincker P, Chourrout D. **Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate.** Science. 2010 Dec 3;330(6009):1381-5.

Denoeud F, Roussel M, Noel B, Wawrzyniak I, Da Silva C, Diogon M, Viscogliosi E, Brochier-Armanet C, Couloux A, Poulain J, Segurens B, Anthouard V, Texier C, Blot N, Poirier P, Ng GC, Tan KS, Artiguenave F, Jaillon O, Aury JM, Delbac F, Wincker P, Vivarès CP, El Alaoui H. **Genome sequence of the stramenopile Blastocystis, a human anaerobic parasite.** Genome Biol. 2011;12(3):R29.

Yadetie F, Butcher S, Førde HE, Campsteijn C, Bouquet JM, Karlsen OA, Denoeud F, Metpally R, Thompson EM, Manak JR, Goksøyr A, Chourrout D. **Conservation and divergence of chemical defense system in the tunicate Oikopleura dioica revealed by genome wide response to two xenobiotics.** BMC Genomics. 2012 Feb 2;13:55.

D'Hont A*, Denoeud F*, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, Da Silva C, Jabbari K, Cardi C, Poulain J, Souquet M, Labadie K, Jourda C, Lengellé J, Rodier-Goud M, Alberti A, Bernard M, Correa M, Ayyampalayam S, Mckain MR, Leebens-Mack J, Burgess D, Freeling M, Mbéguié-A-Mbéguié D, Chabannes M, Wicker T, Panaud O, Barbosa J, Hribova E, Heslop-Harrison P, Habas R, Rivallan R, Francois P, Poiron C, Kilian A, Burthia D, Jenny C, Bakry F, Brown S, Guignon V, Kema G, Dita M, Waalwijk C, Joseph S, Dievart A, Jaillon O, Leclercq J, Argout X, Lyons E, Almeida A, Jeridi M, Dolezel J, Roux N, Risterucci AM, Weissenbach J, Ruiz M, Glaszmann JC, Quétier F, Yahiaoui N, Wincker P. **The banana (Musa acuminata) genome and the evolution of monocotyledonous plants.** Nature. 2012 Aug 9;488(7410):213-7.

Collén J, Porcel B, Carré W, Ball SG, Chaparro C, Tonon T, Barbeyron T, Michel G, Noel B, Valentin K, Elias M, Artiguenave F, Arun A, Aury JM, Barbosa-Neto JF, Bothwell JH, Bouget FY, Brillet L, Cabello-Hurtado F, Capella-Gutiérrez S, Charrier B, Cladière L, Cock JM, Coelho SM, Colleoni C, Czjzek M, Da Silva C, Delage L, Denoeud F, Deschamps P, Dittami SM, Gabaldón T, Gachon CM, Groisillier A, Hervé C, Jabbari K, Katinka M, Kloareg B, Kowalczyk N, Labadie K, Leblanc C, Lopez PJ, McLachlan DH, Meslet-Cladiere L, Moustafa A, Nehr Z, Nyvall Collén P, Panaud O, Partensky F, Poulain J, Rensing SA, Rousvoal S, Samson G, Symeonidi A, Weissenbach J, Zambounis A, Wincker P, Boyen C. **Genome structure and metabolic features in the red seaweed Chondrus crispus shed light on evolution of the Archaeplastida.** Proc Natl Acad Sci U S A. 2013 Mar 26;110(13):5247-52.

Porcel BM, Denoeud F, Opperdoes F, Noel B, Madoui MA, Hammarton TC, Field MC, Da Silva C, Couloux A, Poulain J, Katinka M, Jabbari K, Aury JM, Campbell DA,Cintron R, Dickens NJ, Docampo R, Sturm NR, Koumandou VL, Fabre S, Flegontov P, Lukeš J, Michaeli S, Mottram JC, Szöőr B, Zilberstein D, Bringaud F, Wincker P,Dollet M. **The streamlined genome of Phytomonas spp. relative to human pathogenic kinetoplastids reveals a parasite tailored for plants**. PLoS Genet. 2014 Feb 6;10(2):e1004007.

Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IA, Zhao M, Ma J, Yu J, Huang S, Wang X, Wang J, Lu K, Fang Z, Bancroft I, Yang TJ, Hu Q, Wang X, Yue Z, Li H, Yang L, Wu J, Zhou Q, Wang W, King GJ, Pires JC, Lu C, Wu Z, Sampath P, Wang Z, Guo H, Pan S, Yang L, Min J, Zhang D, Jin D, Li W, Belcram H, Tu J, Guan M, Qi C, Du D, Li J, Jiang L, Batley J, Sharpe AG, Park BS, Ruperao P, Cheng F, Waminal NE, Huang Y, Dong C, Wang L, Li J, Hu Z, Zhuang M, Huang Y, Huang J, Shi J, Mei D, Liu J, Lee TH, Wang J, Jin H, Li Z, Li X, Zhang J, Xiao L, Zhou Y, Liu Z, Liu X, Qin R, Tang X, Liu W, Wang Y, Zhang Y, Lee J, Kim HH, Denoeud F, Xu X, Liang X, Hua W, Wang X, Wang J, Chalhoub B, Paterson AH. **The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes.** Nat Commun. 2014 May 23;5:3930.

Parkin IA, Koh C, Tang H, Robinson SJ, Kagale S, Clarke WE, Town CD, Nixon J, Krishnakumar V, Bidwell SL, Denoeud F, Belcram H, Links MG, Just J, Clarke C, Bender T, Huebert T, Mason AS, Pires JC, Barker G, Moore J, Walley PG, Manoli S, Batley J, Edwards D, Nelson MN, Wang X, Paterson AH, King G, Bancroft I, Chalhoub B, Sharpe AG. **Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea**. Genome Biol. 2014 Jun 10;15(6):R77.

Chalhoub B*, Denoeud F*, Liu S, Parkin IA, Tang H, Wang X, Chiquet J, Belcram H,Tong C, Samans B, Corréa M, Da Silva C, Just J, Falentin C, Koh CS, Le Clainche I, Bernard M, Bento P, Noel B, Labadie K, Alberti A, Charles M, Arnaud D, Guo H, Daviaud C, Alamery S, Jabbari K, Zhao M, Edger P, Chelaifa H, Tack D, Lassalle G, Mestiri I, Schnel N, Le Paslier MC, Fan G, Renault V, Bayer PE, Golicz A, Manoli S, Lee T, Thi V, Chalabi S, Hu Q, Fan C, Tollenaere R, Lu Y, Battail C, Shen J, Sidebottom C, Wang X, Canaguier A, Chauveau A, Bérard A, Deniot G, Guan M, Liu Z, Sun F, Lim YP, Lyons E, Town C, Bancroft I, Wang X, Meng J, Ma J, Pires JC, King GJ, Brunel D, Delourme R, Renard M, Aury JM, Adams KL, Batley J, Snowdon RJ, Tost J, Edwards D, Zhou Y, Hua W, Sharpe AG, Paterson AH, Guan C, Wincker P. Plant genetics. **Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome.** Science. 2014 Aug 22;345(6199):950-3.

Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, ZhengC, Alberti A, Anthony F, Aprea G, Aury JM, Bento P, Bernard M, Bocs S, Campa C,Cenci A, Combes MC, Crouzillat D, Da Silva C, Daddiego L, De Bellis F, Dussert S,Garsmeur O, Gayraud T, Guignon V, Jahn K, Jamilloux V, Joët T, Labadie K, Lan T, Leclercq J, Lepelley M, Leroy T, Li LT, Librado P, Lopez L, Muñoz A, Noel B,Pallavicini A, Perrotta G, Poncet V, Pot D, Priyono, Rigoreau M, Rouard M, Rozas J, Tranchant-Dubreuil C, VanBuren R, Zhang Q, Andrade AC, Argout X, Bertrand B,de Kochko A, Graziosi G, Henry RJ, Jayarama, Ming R, Nagai C, Rounsley S, Sankoff D, Giuliano G, Albert VA, Wincker P, Lashermes P. **The coffee genome provides insight into the convergent evolution of caffeine biosynthesis.** Science. 2014 Sep 5;345(6201):1181-4.

Belser C, Istace B, Denis E, Dubarry M, Baurens FC, Falentin C, Genete M,Berrabah W, Chèvre AM, Delourme R, Deniot G, Denoeud F, Duffé P, Engelen S,Lemainque A, Manzanares-Dauleux M, Martin G, Morice J, Noel B, Vekemans X, D'Hont A, Rousseau-Gueutin M, Barbe V, Cruaud C, Wincker P, Aury JM. **Chromosome-scaleassemblies of plant genomes using nanopore long reads and optical maps.** NatPlants. 2018 Nov;4(11):879-887.

Ferreira de Carvalho J, Lucas J, Deniot G, Falentin C, Filangi O, Gilet M,Legeai F, Lode M, Morice J, Trotoux G, Aury JM, Barbe V, Keller J, Snowdon R, He Z, Denoeud F, Wincker P, Bancroft I, Chèvre AM, Rousseau-Gueutin M. **Cytonuclear interactions remain stable during allopolyploid evolution despite repeated whole-genome duplications in Brassica**. Plant J. 2019 Jan 3.

Alsos IG, Lavergne S, Merkel MKF, Boleda M, Lammers Y, Alberti A, Pouchon C, Denoeud F, Pitelkova I, Puşcaş M, Roquet C, Hurdu BI, Thuiller W, Zimmermann NE, Hollingsworth PM, Coissac E. **The Treasure Vault Can be Opened: Large-Scale Genome Skimming Works Well Using Herbarium and Silica Gel Dried Material.** Plants (Basel). 2020 Apr 1;9(4):432.

Rousseau-Gueutin M, Belser C, Da Silva C, Richard G, Istace B, Cruaud C, Falentin C, Boideau F, Boutte J, Delourme R, Deniot G, Engelen S, de Carvalho JF, Lemainque A, Maillet L, Morice J, Wincker P, Denoeud F, Chèvre AM, Aury JM**. Long-read assembly of the Brassica napus reference genome Darmor-bzh**. Gigascience. 2020 Dec 15;9(12):giaa137

Wang Y, Pedersen MW, Alsos IG, De Sanctis B, Racimo F, Prohaska A, Coissac E, Owens HL, Merkel MKF, Fernandez-Guerra A, Rouillard A, Lammers Y, Alberti A, Denoeud F, Money D, Ruter AH, McColl H, Larsen NK, Cherezova AA, Edwards ME, Fedorov GB, Haile J, Orlando L, Vinner L, Korneliussen TS, Beilman DW, Bjørk AA, Cao J, Dockter C, Esdale J, Gusarova G, Kjeldsen KK, Mangerud J, Rasic JT, Skadhauge B, Svendsen JI, Tikhonov A, Wincker P, Xing Y, Zhang Y, Froese DG, Rahbek C, Nogues DB, Holden PB, Edwards NR, Durbin R, Meltzer DJ, Kjær KH, Möller P, Willerslev E. **Late Quaternary dynamics of Arctic biota from ancient environmental genomics**. Nature. 2021 Dec;600(7887):86-92.

Pouchon C, Boyer F, Roquet C, Denoeud F, Chave J, Coissac E, Alsos IG; PhyloAlps Consortium; PhyloNorway Consortium, Lavergne S. **ORTHOSKIM: In silico sequence capture from genomic and transcriptomic libraries for phylogenomic and barcoding applications**. Mol Ecol Resour. 2022 Jan 11.

Smyčka J, Roquet C, Boleda M, Alberti A, Boyer F, Douzet R, Perrier C, Rome M, Valay J-G, Denoeud F, Šemberová K, Zimmermann NE, Thuiller W, Wincker P, Alsos IG, Coissac E, *the PhyloAlps consortium*, Lavergne S. **Tempo and drivers of plant diversification in the European mountain system**. (accepté, Nature communications)

Alsos IG, Rijal DP, Ehrich D, Karger N, Yoccoz NG, Heintsman PD, Brown AG, Lammers Y, Pellissier L, Alm T, Brathen KA, Coissac Z, Merkel MKF, Alberti A, Denoeud F, Bakke J, PhyloNorway consortium. **Post-glacial species arrival and diversity build-up of northern ecosystems took millennia** (soumis).

Noel B*, Denoeud F*, Rouan A, Buitrago-López C, Capasso L, Poulain J, Boissin E, Pousse M, Da Silva C, Couloux A, Armstrong E, Carradec Q, Cruaud C, Labadie K, Lê-Hoang J, Tambutté S, Barbe V, Moulin C, Bourdin G, Iwankow G, Romac S, *Tara Pacific Consortium Coordinators*, Allemand D, Planes S, Gilson E, Zoccola D, Wincker P, Voolstra CR, Aury JM. **High-quality genome assemblies highlight pervasive gene duplication as a major evolutionary driving force in coral biology**. *En preparation.*

Communications scientifiques

Je liste les 5 présentations faites dans des congrès internationaux les plus marquantes.

ECCB 2003 (European Conference on Computational Biology), 27-30 September 2003, Paris, France. Short paper and talk: « Resources for bacterial strain identification using polymorphic tandem repeats ».

Journée « Nouvelles technologies de séquençage », 9 mars **2009**, université de Lille. Présentation « Les nouvelles technologies de séquençage au Genoscope: assemblage et annotation de génomes » (conférence invitée)

PAG (Plant and Animal Genomes), 12-16 jan **2013**, San Diego, USA. Talk "Coffee Genomics: The Coffea canephora genome"

RECOMB Comparative Genomics, 17-19 oct **2013**, Lyon Villeurbanne: **Keynote talk**: Whole genome duplications in plant genomes (conférence invitée)

EAGS (Environemental and Agronomical Genomics Symposium), 27-29 oct **2021**, Tours. Talk: "High-quality genome assemblies of corals highlight the specifics of their long lifespan"

Liste des encadrés ou co-encadrés d'étudiants

Durant ma thèse (2001-2003), Institut de Génétique et Microbiologie (université Paris Sud XI, Orsay) :

2002:

Encadrement d'une stagiaire de licence (V Douchin) : 3 mois. Etude du polymorphisme de minisatellites humains (génétique moléculaire : PCR, et bioinformatique).

Encadrement d'un stagiaire de licence (J Petit) : 3 mois. Etude du polymorphisme de minisatellites humains (génétique moléculaire : PCR, et bioinformatique).

2002 et 2003 :

Mise en place et encadrements de TD et de TP à l'université d'Orsay (12h) . Approches bioinformatiques pour l'identification de souches bacterienne après typage de minisatellites (profils de typage, matrices de distances et arbres phylogénétiques).

Durant mon postdoc (2004-2006) :

Encadrement d'un ingénieur sur le projet ENCODE dans le groupe de Roderic Guigo.

Depuis 2006 au Genoscope :

La plupart de mes projets de recherche ont été menés en collaboration avec plusieurs laboratoires. J'ai eu l'occasion de jouer le rôle de conseillère, que j'affectionne particulièrement, auprès de plusieurs stagiaires et thésards du laboratoire. En particulier, j'ai eu l'occasion d'apporter des éléments de réflexion à une étudiante en thèse, Sarah Farhat, qui étudiait deux organismes marins (*Amoebophrya*) dont les introns sont très particuliers (énormément d'introns non canoniques, très divers, et différents entre les deux espèces étudiées) (Fahrat, 2018).

Janvier-juillet 2018 : encadrement d'une étudiante de Master 2 BiB (Biologie informatique / Bioinformatique, Université Paris Diderot), Lina Alferkh : « Analyse de génomes de plantes séquencés à faible couverture à l'échelle d'un écosystème entier ». Dépôt d'un sujet de thèse en 2018 à l'école doctorale SDSV (« Génomique comparative à l'échelle d'un biome entier : évolution et contingence des voies de biosynthèse de métabolites secondaires chez les plantes de haute montagne »), mais l'étudiante a obtenu une bourse dans un autre laboratoire avant la date de soutenance pour l'école doctorale SDSV et a préféré se désister.

Octobre 2018 : Membre d'un jury de validation des acquis de l'expérience (VAE), en tant que professionnelle, pour le Master BiB (Biologie informatique / Bioinformatique, Université Paris Diderot), étudiante Dina Zeliewski. Cette expérience très intéressante m'a permis de prendre du recul sur le métier de chercheur dans le domaine de la bioinformatique.

Janvier-Juillet 2019 : encadrement d'un étudiant de Master 2 Bib (Biologie informatique / Bioinformatique, Université Paris Diderot), Daniel De Murat **:** « Analyse de génomes de plantes séquencés à l'échelle d'un écosystème entier : détection de familles de gènes amplifiées dans les milieux alpins ». Daniel n'a pas souhaité poursuivre en thèse.

Mars-Aout 2021 : encadrement d'une étudiante de Master 2 AMIB (Université Paris Saclay), Tolulope Apanishile : « Prédiction de la taille du génome et identification de familles de gènes dont l'abondance est corrélée à la taille : étude sur la famille des astéracees ». Tolulopé n'a pas souhaité poursuivre en thèse.

Mars-Aout 2022 : encadrement d'un étudiant de troisième année d'école d'ingénieur (Ecole Nationale Supérieure Agronomique de Toulouse), Pierre Guenzi-Tibéri : « Développement d'un outil de prédiction de taille de génomes à partir de séquençage à faible couverture ». Ce stage pourrait se poursuivre par une thèse sur le projet Phaeoexplorer.

Bilan de mon expérience d'encadrement

Mon expérience d'encadrement de 4 étudiants de Master2 ou dernière année d'école d'ingénieur a été particulièrement enrichissante. Effectivement, j'ai beaucoup apprécié d'avoir le rôle de conseiller et d'expert, en déléguant la mise en œuvre technique des analyses, ce qui a permis que les étudiants apportent (et testent) des idées nouvelles, qui ont enrichi notre perspective. J'ai eu la chance de tomber sur des étudiants autonomes et en qui j'ai rapidement eu toute confiance. J'avais d'ailleurs souhaité garder Lina Alferkh pour une thèse, pour laquelle j'avais déposé un sujet à l'école doctorale SDSV de l'Université d'Evry, Paris Saclay, mais ma candidate a obtenu une bourse dans une autre école doctorale avant les concours et a préféré accepter cette offre. Les étudiants que j'ai accueillis par la suite ne souhaitaient pas poursuivre en thèse, et n'avaient pas tout à fait acquis la maturité pour le faire. J'espère pouvoir garder mon étudiant de cette année, Pierre Guenzi-Tibéri, sur le sujet de thèse que je propose sur l'analyse des génomes d'algues brunes. J'avais eu précédemment l'occasion d'effectuer le rôle de conseiller que j'affectionne particulièrement auprès de plusieurs stagiaires et thésards du laboratoire. En particulier, j'ai eu l'occasion d'apporter des éléments de réflexion à une étudiante en thèse, Sarah Fahrat, qui étudiait deux organismes

marins (*Amoebophrya*) dont les introns sont très particuliers (énormément d'introns non canoniques, très divers, et différents entre les deux espèces étudiées) (Fahrat, 2018). Ayant acquis une grande expertise dans le domaine des introns lors du projet portant sur *Oikopleura dioica*, c'est tout naturellement que j'ai proposé différents axes d'analyse pour sa thèse. Etant ensuite partie en congé parental pour près d'un an, je n'ai malheureusement pas pu suivre les progrès de ses recherches autant que je l'aurais souhaité, mais cette expérience m'a confortée dans mon goût pour l'encadrement. Enfin, j'ai eu l'occasion de participer en octobre 2018 à un jury de validation des acquis de l'expérience (VAE), en tant que professionnelle, pour le Master BiB (Biologie informatique Bioinformatique) de l'université Paris Diderot. Cette expérience très intéressante m'a permis de prendre du recul sur le métier de chercheur dans le domaine de la bioinformatique.

Formation en lien avec l'HDR

2021 : Formation « Etre pédagogue dans le management de la recherche »

- Connaitre les grandes étapes du processus d'apprentissage de thèse, leurs caractéristiques et les outils à mettre en place

- Développer les postures d'encadrement adaptées à chaque étape et à l'ensemble de la thèse

- Prendre conscience de la formalisation d'une démarche d'accompagnement adaptée à chaque situation de doctorat

Formation organisée par l'école doctorale SDSV. Trois modules de deux jours chacun (42h, en avril, juin et septembre 2021).

Liste des collaborations / contrats

Pendant ma thèse à l'Institut de Génétique et Mircrobiologie d'Orsay, j'ai collaboré avec Gary Benson, de la Mount Sinaï School of Medicine New York, développeur de l'outil TRF (tandem repeats finder). Nous avons mis au point une méthode pour prédire le polymorphisme des minisatellites à partir du profil de conservation des unités répétées en tandem.

Durant mon post-doctorat, j'ai été impliquée dans le projet ENCODE, qui était mené par un consortium international de grande envergure, ce qui a été l'occasion de nombreuses collaborations internationales (entre-autres avec les laboratoires de Tom Gingeras, chez Affymetrix, Alexandre Reymond à l'Université de Lausanne, Jennifer Harrow au Sanger Center, Ewan Birney à l'EBI) (Harrow J*, Denoeud F* et al, 2006, Denoeud et al, 2007). Etant en charge de l'analyse dans le groupe de Roderic Guigo, j'ai participé à de nombreux workshops et « data fairs » ce qui m'a familiarisée avec la gestion de projet au sein d'un grand consortium (contraintes variées pour atteindre les dates limites de production et traitement de données, organisation de l'analyse de ces données, organisation de la rédaction des articles…).

Depuis que je suis au Genoscope, j'ai été impliquée dans les collaborations suivantes :

- Collaboration internationale avec l'équipe de Daniel Chourrout (SARS, Bergen, Norvège) sur l'analyse du génome d'*Oikopleura dioica*. (2006-2009)
- Collaboration avec l'équipe de Hicham El Alaoui Université de Clermont) sur l'analyse du génome de *Blastocystis* (parasite intestinal humain).
- Collaboration avec l'équipe de Francis Martin (INRA de Nancy) sur l'analyse du génome de la truffe.
- Collaboration avec l'équipe de Catherine Boyen (Station Biologique de Roscoff) sur l'analyse du génome de l'algue rouge *Chondrus crispus*.
- Collaboration avec l'équipe de Michel Dollet (CIRAD, Montpellier) sur l'analyse du génome de deux *Phytomonas* (trypanosomes de plantes).
- Collaboration internationale, principalement avec l'équipe dAngélique D'hont (CIRAD Montpellier) sur l'analyse du génome du bananier.
- Collaboration internationale au sein du consortium *Coffea canephora* (P Lashermes : IRD Montpellier, Victor Albert : Université de Buffalo, Giovanni Giuliano : ENEA Italie, David Sankoff : université d'Ottawa) sur l'analyse du génome du caféier.
- Collaboration internationale au sein du consortium *Brassica* (Boulos Chalhoub : INRA, Isobel Parkins : Agri-Food, Canada), Dave Edwards : univ of Western Australia, Rod Snowdon : Justus Liebig University, Allemagne, Haibao Tang, Craig Venter Institute…) sur l'analyse du génome du colza.

- Collaboration dans le cadre du projet France Génomique Polysuccess (« how a polyploid becomes a new species, the Brassica model ») – porteuse de projet: Anne-Marie Chèvre, INRA de Rennes.

Je suis actuellement en charge au Genoscope de piloter les projets France Génomique suivants :

- PhyloAlps (étude de la biodiversité de la flore alpine) –porteur de projet : Sébastien Lavergne, LECA Grenoble. Je suis également responsable du le projet « PhyloNorway » (exploration de la Flore arctique), en lien avec PhyloAlps, en collaboration avec l'équipe de Inger Greeves Alsos, au Museum Arctic University de TromsØ, Norvège.

-TARA Pacific : je fais partie du consortium TARA Pacific à l'occasion de l'analyse des génomes de référence de coraux.

- Phaeoexplorer (étude des génomes d'algues brunes) – porteur de projet : Mark Cock, CNRS Station Biologique de Roscoff. Je suis co-PI de ce projet avec Mark Cock (c'est moi qui représente le Genoscope) et également responsable du groupe d'analyse « General Genome Features » au sein du consortium international Phaeoexplorer. Il s'agit d'animer les réunions du groupe de travail, et de coordonner l'écriture des sections de l'article qui décrira les génomes d'algues brunes concernant cette thématique. Je me concentrerai en particulier sur les familles de gènes amplifiées dans/entre les différentes algues brunes, et je prévois d'encadrer un doctorant sur ce thème.

- J'ai également été la **porteuse d'un projet ANR**, écrit conjointement avec Mark Cock, nommé « Phaeodiscovery » déposé à l'AAPG2020 (fin 2019). Ce projet avait pour but de poursuivre l'effort de séquençage entrepris lors du projet Phaeoexplorer par un effort d'analyse. Ce projet n'a pas été retenu : le reproche principal était qu'il dépendait trop de données déjà générées. Nous avions pourtant mentionné clairement que le but principal était l'analyse des génomes d'algues brunes généré par le projet Phaeoexplorer grâce au financement ANR via France Génomique. Il nous semblait logique que l'ANR soutienne l'exploitation de ces données compte tenu de son fort investissement (très apprécié) dans le séquençage et assemblage de ces génomes. Nous étions donc un peu surpris par les commentaires des évaluateurs car c'était justement tout l'objectif du projet. Nous avons soumis à nouveau ce projet à l'AAPG2021, en y ajoutant un volet de production de nouvelles données expérimentales. Nous proposions de valider la fonction de gènes candidats (impliqués dans la multicellularité ou l'adaptation au milieu intertidal entre autres) identifiés par des approches bioinformatiques par une approche expérimentale de knock-out de ces gènes par CRISPR-Cas9 dans l'espèce cultivée *Ectocarpus*. Ce projet n'a malheureusement pas non-plus été retenu.

- Enfin, j'ai participé à la conception et à l'écriture d'un projet soumis à l'appel à projets « **Biodiversa** » par Olivier Evrard, du Laboratoire des Sciences du Climat et de l'Environnement (LSCE) du CEA de Saclay. Ce projet, nommé « REBIOS » visait à analyser l'ADN contenu dans des archives sédimentaires aquatiques afin d'identifier les sols (types de cultures) les plus sujets à l'érosion et de proposer des voies d'amélioration pour les pratiques agricoles dans

différentes régions du monde (entre-autres la Guyane et le Brésil). L'hypothèse sous-jacente est que l'ADN provenant de plantes poussant sur les sols qui s'érodent le plus se retrouve en plus grande quantité dans les sédiments du bassin versant correspondant. Le Genoscope proposait de prendre en charge les expériences de metabarcoding et les analyses subséquentes pour identifier les espèces de plantes présentes dans différents environnements aquatiques. Un projet pilote est actuellement en cours.