Phylogénie des néogastéropodes : développements méthodologiques et évaluation du succès de l'approche par capture d'exons

Thomas Lemarcis

MUSEUM NATIONAL D'HISTOIRE NATURELLE
Ecole Doctorale Sciences de la nature et de l'Homme – ED 227

Année 2024

N°attribué par la bibliothèque
|_|_|_|_|_|_|_|_|_|_|_|_|

THESE

Pour obtenir le grade de

DOCTEUR DU MUSEUM NATIONAL D'HISTOIRE NATURELLE

Spécialité : Biologie moléculaire

Présentée et soutenue publiquement par

Thomas Lemarcis

Le 26 Avril 2024

Phylogénie des néogastéropodes : développements méthodologiques et évaluation du succès de l'approche par capture d'exons

---

Sous la direction de **: M. Nicolas PUILLANDRE, Maître de Conférences HDR, MNHN**

JURY :

Mme. Jousselin, Emmanuelle	**Directrice de recherche, CBGP, Montferrier-sur-Lez, France**	**Rapportrice**
M. Oliverio, Marco	**Professor, Sapienza University, Roma, Italia**	**Rapporteur**
Mme. Cariou, Marie	**Ingénieure de Recherche, Muséum national d'Histoire naturelle, Paris, France**	**Examinatrice**
M. Simion, Paul	**Maître de conférences, Université de Rennes, Rennes, France**	**Examinateur**
Mme. Nicolas-Colin, Violaine	**Professeure, Muséum national d'Histoire naturelle, Paris, France**	**Examinatrice**
M. Puillandre, Nicolas	**Maître de conférences, HDR, Muséum national d'Histoire naturelle, Paris, France**	**Directeur de Thèse**

À mon père et mon grand-père, partis bien trop tôt.

Remerciements

Il paraît que la vie c'est comme une boîte de chocolat. Je dirais plutôt que c'est un enchaînement d'opportunités saisies ou non. Pour ma part c'est cet e-mail envoyé à Nico pour discuter d'un projet de thèse éventuel. Je me souviendrai toujours des mots que tu m'as dit à la fin de l'entretien « Je suis soulagé que tu acceptes ce projet de thèse ». Dans cette phrase il y avait toute la confiance que tu m'as donnée depuis le premier jour et jusqu'au dernier. Cela sans jamais la remettre en cause, quelques soient les circonstances. Pour ça je te remercie et je pense que tu ne peux pas savoir à quel point cela a compté pendant ces trois années. Merci pour tout ce que tu m'as appris pendant ces années de thèse, cela est un réel plaisir de travailler avec toi.

Merci à l'Institut de Systématique, Évolution, Biodiversité et sa directrice, Violaine Nicolas-Colin pour son accueil au sein de l'unité. Merci aux membres de l'équipe 3E ainsi que sa directrice Sarah Samadi pour leur accueil au sein de l'équipe.

Merci aux membres de mon jury : Emmanuelle Jousselin, Marco Oliverio, Marie Cariou, Paul Simion et Violaine Nicolas-Colin. Merci également aux membres du comité de suivi de thèse : Alexander Fedosov, Jawad Abdelkrim, Rodolphe Rougerie et Violaine Llaurens. Merci à Nathalie Machon et Jérôme Sueur ainsi que tous les membres de l'école doctorale 227 pour leur aide et leurs conseils précieux.

Merci aux membres de l'équipe ERC HYPERDIVERSE, merci à Tanya pour ton aide dans toutes les circonstances. Merci à Dario, Mélanie, Allan, Sarah, Claudia et Alessandro.

Merci à toutes les personnes du troisième étage du bâtiment 51 de votre accueil durant les dernières semaines de ma thèse. Merci à Barbara, Pierre, Priscillia, Magalie, Laure, Virginie, Philippe et Philippe.

Merci aux taxonomistes pour leur aide d'identification ainsi que de prêt de matériel pour mon travail de thèse : Yuri Kantor, Alexander Fedosov, Giulia Fassio, Marco Oliverio, Maria Vittoria Modica, Roland Houart, Emmanuel Tardy, Olivier Crabos, Kevin Monsecour, Richard Salisbury, Aart Dekkers, Sandro Gori, Shih I Huang, Max Marrow, Gary Rosenberg, Richard Salisbury, Girogio Strano, Fred Vervaet, Peter Stahlschmidt, André Verhecken, Yasunori Kano.

Merci à Yuri et Sasha pour nos échanges et votre savoir précieux sur les néogastéropodes. J'ai pu en apprendre beaucoup sur ce groupe grâce à vous.

Merci aux membres de la plateforme de séquençage de l'ICM : Yannick Marie, Delphine Bouteiller et Agnès Rastetter. Merci pour votre disponibilité et vos compétences qui ont grandement contribué à ce projet de thèse.

Merci à Julie Vasseur pour ton aide, tes compétences et ta disponibilité de chaque instant. J'aurais aimé pouvoir plus travailler au laboratoire avec toi. Le SSM reposait entièrement sur tes épaules.

Merci à Marie, Amandine, Jawad et Fayçal pour votre aide et vos réponses à toutes mes interrogations.

Merci à Gaspard, Manon et Thomas pour votre accueil dans le bureau après mon déménagement en urgence. Nous avons pu enrichir nos connaissances de la géographie mondiale tous ensemble. Merci également à Henri, Estelle, Valentin, Nisha d'avoir complété la team du bureau de la convivialité lors des pauses thé, très rares mais toujours importantes.

Merci à Jawad, tu m'as fait un speedrun de l'apprentissage du langage Python. Sans toi je n'aurais jamais appris si rapidement. Merci également pour nos moments de détente raquette en main ô combien importants.

Merci à l'Agent F pour ton aide tant dans le travail qu'en dehors. J'espère que j'aurai la chance à l'avenir de travailler avec des personnes comme toi.

Merci à Paul pour ton accompagnement pendant toute cette thèse, et tes réponses à toutes mes questions, qui étaient particulièrement nombreuses et variées.

Merci à Antoine, tu nourris la B.O qui m'accompagne presque tous les jours, depuis maintenant 15 ans.

Merci aux amis du baby-foot pour ces moments de détente. Le Gamelle Comedy Club a un bel avenir devant lui.

Merci aux copains de la cantine pour tous ces repas partagés.

Merci à Paul, Jawad, Damien et Gaspard pour ces moments sportifs partagés, dans un souci d'entretenir les corps autant que l'esprit.

Merci à Arnaud d'être à mes côtés, à plus ou moins longue distance, depuis toutes ces années.

Merci à mes parents d'avoir allumé cette étincelle de curiosité qui est toujours avec moi chaque jour. Merci à ma grand-mère, mon oncle et mon frère pour votre soutien.

Merci à Claire, la vie est bien plus douce avec toi.

La présente thèse ne constitue pas une publication au sens
du Code International de Nomenclature Zoologique.
[Code, Recommandation 8E]

Articles et communications réalisés pendant la thèse ou en cours de réalisation

(* : inclus dans la thèse)

Articles :

*** Lemarcis T**, Fedosov A, Kantor Y, Abdelkrim J, Zaharias P, Puillandre N. (2022) Neogastropod (Mollusca, Gastropoda) phylogeny : a step forward with mitogenomes. Zoologica Scripta, 51: 550-561.

* Fedosov A, Zaharias P, **Lemarcis T**, Modica MV, Holford M, Oliverio M, Kantor Y, Bouchet P, Puillandre N. Phylogenomics of the Neogastropoda: the backbone hidden in the bush. Syst. Biol., in press.

*** Lemarcis T**, Blin A, Derzelle A, Farhat S, Fedosov A, Zaharias P, Zuccon D, Puillandre N. Too far from relatives? Impact of the genetic distance on the success of exon capture phylogeny. In prep.

Communications :

Fassio G, Chiappa G, Nucella E, **Lemarcis T**, Takano T, Kano Y, Puillandre N, Bouchet P, Treneman N, Malaquias MA, Modica MV and OliverioM. Shell evolution in Velutinoidea: a phylogenetic approach. Congrès EVOLMAR, 14-17 novembre 2023, online.

Fedosov A, Zaharias P, **Lemarcis T**, Modica MV, Holford M, Kantor Y, Oliverio M, Bouchet P, Puillandre N. Exon-capture-based phylogeny of the Neogastropoda. World Congress of Malacology, 1-5 août 2022, Munich, Allemagne.

Lemarcis T, Abdelkrim J, Derzelle A, Zaharias P, Kantor Y, Fedosov A, Puillandre N. Exon design for large-scale phylogeny of the Neogastropoda. World Congress of Malacology, 1-5 août 2022, Munich, Allemagne. Poster.

Puillandre N, **Lemarcis T,** Ratti C, Fedosov A, Kantor Y, Modica MV, Oliverio M, Bouchet. Sampling the known and unknown diversity in hyperdiverse groups for molecular phylogeny. World Congress of Malacology, 1-5 août 2022, Munich, Allemagne. Poster.

Lemarcis T. Impact of phylogenetic distance on exon capture success in Neogastropoda, Journée des doctorants de l'ISYEB, 17 mars 2023, Paris, France. Présentation orale.

Lemarcis T. Impact of phylogenetic distance on exon capture success in Neogastropoda, Congrès des Jeunes Chercheurs du Muséum, 3-5 mai 2023, Paris, France. Présentation orale.

Liste des scripts informatiques déposés sur le Github du projet ERC HYPERDIVERSE

https://github.com/Hyperdiverseproject/Exon_capture

1cleanreads.py

2trinityWrapper.py

2spadesWrapper.py

2.5rename_spades_contigs.py

2.5rename_trinity_contigs.py

2.5merging_contigs.py

3clusteringWrapper.py

4mapping.py

5recipblastWrapper.py

6cov_hetWrapper.py

7filter_cov_het.py

8cut_fasta_and_1stalign.py

8.5merging_doublons.py

9blastx.py

10sequences_translation.py

11sorting_samples_NEO700.py

11sorting_samples_NEO50.py

11sorting_samples_Raphito700.py

11sorting_samples_Raphito50.py

Table des matières

PROLOGUE 15

Chapitre 1 - Introduction 19
1. Présentation générale des Neogastropoda 19
2. Phylogénies et classifications 26
3. Phylogénie basée sur les mitogénomes 29
4. Les méthodes de séquençage de génomes réduits 43

Chapitre 2 - Développements méthodologiques 49
1. Pré-séquençage : design des sondes de capture 49
1.1. Jeu de données 50
1.2. Recherche des exons 53
1.3. Design des sondes 59
2. Extraction et quantification d'ADN 62
2.1. Extraction d'ADN à partir des coquilles : 63
2.2. Quantifications ADN : Tests, Choix du protocole Qubit 67
2.3. Choix de la méthode de quantification 77
3. Stratégie d'échantillonnage pour la capture d'exons 78
3.1. Sélection des échantillons et séquençage en 3 batchs 83
3.2. Quantification des ADN et préparation des banques 85
4. Post-séquençage : des données brutes vers les exons. 93
4.1. Description du pipeline 93
4.2. Identification des différents points à améliorer dans le pipeline. 100
5. Résultats bruts 105

Chapitre 3 : Distance génétique et succès de la capture d'exon 107

Chapitre 4 : Phylogénies 145
1. Neogastropoda 145
1.1. Jeux de données et analyses phylogénétiques 145
1.2. Arbre phylogénétique NT123-700 149
1.3. Comparaison avec les autres arbres 152
1.4. Changements potentiels dans la classification 159

1.5. Conclusions 163
2. Raphitomidae 165
2.1. Jeux de données et analyses phylogénétiques 165
2.2. Arbre phylogénétique AA-700-NP 169
2.3. Changements potentiels dans la classification 178
2.4. Conclusions 179

Conclusions et perspectives 181
1. Phylogénies des Neogastropoda et des Raphitomidae 181
2. Bilan et perspectives méthodologiques 184

Bibliographie 189

ANNEXE 1 : 207
ANNEXE 2 : 244
ANNEXE 3 : 245

PROLOGUE

Ce travail de thèse s'inscrit au sein du projet HYPERDIVERSE financé par l'ERC (European Research Council), un projet européen débuté en Octobre 2020 pour une durée de 5 années. Le projet HYPERDIVERSE vise à identifier les mécanismes qui vont contribuer à l'hyper-diversification de certains groupes taxonomiques, par la recherche de facteurs génétiques (innovations clés) (Hunter, 1998; Van Valen, 1971) ou écologiques (« Ecological Release ») (Duda & Lee, 2009; Van Valen, 1965) qui vont jouer un rôle sur l'augmentation du taux de diversification dans un groupe donné. Les taxa hyper-diversifiés sont largement méconnus, comme les insectes, les araignées, les annélides, les nématodes et les mollusques. Tester les hypothèses liées aux dynamiques de diversification de ces groupes demeure un challenge en biologie évolutive et en systématique. En effet, les mécanismes sous-jacents du déterminisme génétique et des processus évolutifs à l'œuvre, et en particulier le rôle précis de la variation d'un trait et la capacité de l'organisme qui porte ce trait à conquérir de nouvelles niches écologiques, restent mal connus.

L'objectif principal du projet HYPERDIVERSE est donc d'identifier les processus évolutifs à l'origine de la variation des traits, les mécanismes avec lesquels ces traits interagissent avec l'environnement, et leur impact sur les dynamiques de diversification chez un organisme non modèle.

Pour répondre à cet objectif, il est nécessaire de connaître l'histoire évolutive du groupe étudié, et d'en reconstruire une phylogénie robuste. C'est sur la base de cette phylogénie que nous pourrons émettre des hypothèses sur la dynamique de diversification du groupe, à savoir quelles lignées sont plus ou moins diversifiées et l'échelle de temps à laquelle la diversification s'est mise en place. Enfin cela permettra de tester la présence de corrélations entre la dynamique de diversifications et les traits d'histoire de vie.

Le groupe des néogastéropodes (Mollusca, Gastropoda, Neogastropoda) est le groupe modèle qui a été choisi pour ce projet ERC. C'est un taxon qui inclut de nombreuses espèces économiques et culturelles importantes. Les néogastéropodes sont pour la grande majorité des prédateurs, et certaines espèces sont connues pour produire des toxines, des analgésiques ou des anticoagulants (Farhat et al., 2023), ce qui en fait un groupe prometteur pour la découverte de composants biologiques et d'applications dans le domaine des biotechnologies par exemple.

Les néogastéropodes regroupent à l'heure actuelle plus de 15000 espèces décrites mais on estime entre 30000 et 50000 le nombre total d'espèces supposées, ce qui en fait un des groupes de mollusques les plus diversifiés. Ce sont des prédateurs marins actifs dont la phylogénie reste encore mal connue, notamment à l'échelle des super-familles. De plus il reste donc encore un grand nombre d'espèces à décrire dans ce groupe.

Le Muséum National d'Histoire Naturelle de Paris (MNHN) regroupe la plus grande collection de néogastéropodes au monde. Les expéditions menées par les équipes du muséum permettent de collecter un grand nombre de spécimens sur le terrain, qui couvrent une grande partie de la diversité des néogastéropodes. Cette grande diversité de spécimens disponibles nous permet d'envisager d'inclure dans une phylogénie la totalité des super-familles et familles décrites à l'heure actuelle chez les néogastéropodes. Ce vaste échantillonnage permettra également de décrire de nouvelles espèces, voire de nouveaux genres ou peut-être même de nouvelles familles.

Le projet HYPERDIVERSE comprend 3 objectifs majeurs, (i) Améliorer les connaissances taxonomiques au sein du groupe, notamment avec la production d'une phylogénie globale ; (ii) Établir des corrélations entre la variabilité des traits clés identifiés et les taux de diversification ; (iii) Identifier les déterminants génétiques et les mécanismes évolutifs qui sous-tendent les valeurs adaptatives et les traits clés. Ces 3 objectifs sont répartis en 8 tâches (Figure 1).

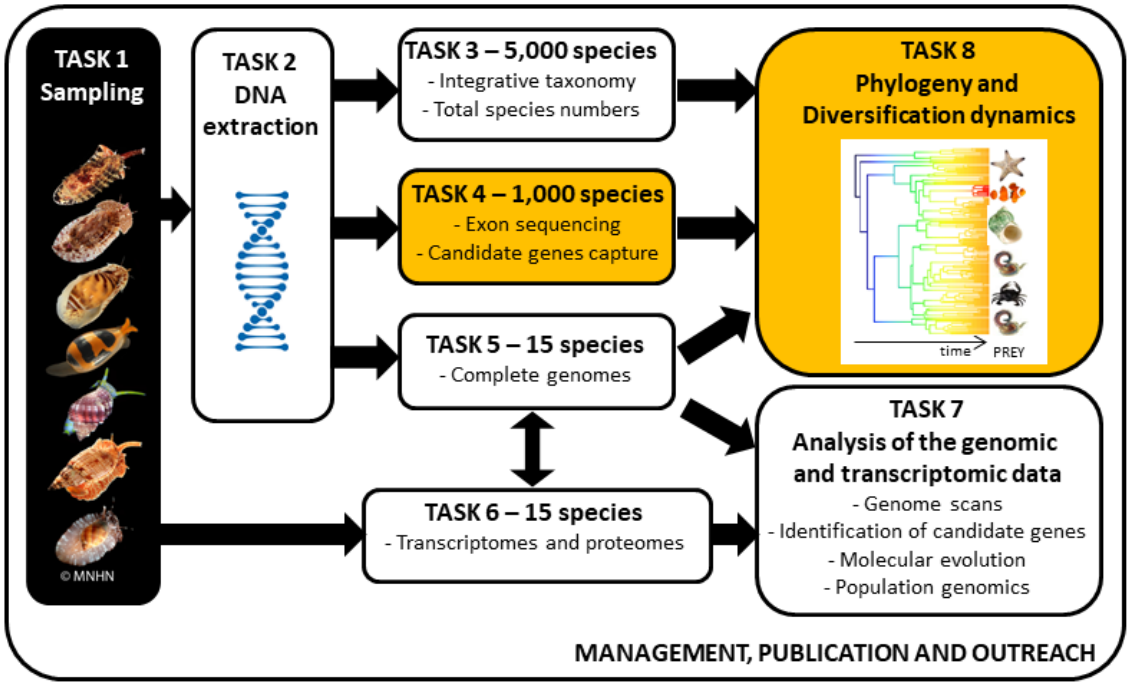


Figure 1 : Schéma global des différentes tâches du projet HYPERDIVERSE.

Ma thèse s'inscrit donc directement dans l'objectif 1 du projet HYPERDIVERSE, avec l'application d'une approche de capture d'exons pour produire la phylogénie la plus exhaustive à l'heure actuelle chez les néogastéropodes. Cela permettra de clarifier la classification à plusieurs échelles taxonomiques, au niveau des superfamilles, des familles et des genres. C'est sur la base de cette phylogénie que les hypothèses de diversification et de modification des vitesses d'évolution des traits d'histoire de vie seront testées.

La méthode de capture d'exons choisie pour produire la phylogénie des néogastéropodes a déjà été appliquée au sein de l'équipe et est maintenant largement citée et utilisée dans la littérature scientifique (Bi et al., 2012; Bragg et al., 2016; Teasdale et al., 2016). Cependant, son efficacité à de larges échelles taxonomiques n'a que rarement été testée de façon rigoureuse avec un protocole spécifique. En plus de la reconstruction phylogénétique au sens strict, afin de clarifier la systématique du groupe, cette thèse permettra d'aborder des questions méthodologiques liées à la capture d'exons et à son application à de larges échelles phylogénétiques. L'étude de l'efficacité de la capture d'exons en fonction de la distance phylogénétique nous permettra de vérifier qu'un même jeu de sondes de capture peut capturer des taxa qui ont divergé depuis plusieurs dizaines de millions d'années et donc de produire des phylogénies à de grandes profondeurs phylogénétiques.

Ma thèse au sein de l'équipe HYPERDIVERSE m'a permis de travailler dans un environnement d'étude avec des modèles biologiques variés, avec des questions communes mais parfois des méthodes différentes. C'est ce qui m'a encouragé à apprendre et à perfectionner mes compétences en bioinformatique pour l'analyse des données. J'ai aussi pu contribuer et répondre à des questions relatives à d'autres axes du projet, notamment pour la réalisation d'expérimentations de laboratoire.

Cette thèse va suivre un plan en 4 grandes parties. Tout d'abord, une partie introductive permettra de présenter le modèle biologique, et notamment l'état des connaissances relatives à son histoire évolutive. Cette partie incluera la présentation d'une phylogénie des néogastéropodes basées sur les mitogénomes que nous avons publiée. Dans le chapitre 2, je détaillerai le protocole que nous avons suivi, puis les différents tests expérimentaux que j'ai réalisés au cours de ma thèse, liés notamment à la préparation des échantillons. Je présenterai, toujours dans le chapitre 2, les méthodes bioinformatiques utilisées pour analyser les données issues du séquençage. Dans le chapitre 3, je présenterai les résultats des tests réalisés pour estimer l'efficacité de la capture d'exons en fonction de l'éloignement phylogénétique, puis,

dans le chapitre 4, les phylogénies reconstruites à l'échelle de l'ordre (Neogastropoda) mais aussi pour une famille en particulier (Raphitomidae). Cela sera suivi par une conclusion générale de la thèse et les perspectives qui s'ouvrent après le travail que j'ai réalisé.

Chapitre 1 - Introduction

1. PRESENTATION GENERALE DES NEOGASTROPODA

Les néogastéropodes représentent un ordre de gastéropodes marins, au sein de la sous-classe des Caenogastropoda. À l'heure actuelle, le groupe-frère des néogastéropodes parmi les Caenogastropoda n'a pas été identifié. De plus, la monophylie des néogastéropodes n'est pas confirmée (Osca et al., 2015). Les Caenogastropoda représentent une classe qui comprend 157 familles décrites et acceptées à l'heure actuelle (Bouchet et al., 2017), dont la classification n'a pour l'instant pas réellement été testée par des approches phylogénétiques. De nombreux clades au sein des Caenogastropoda se sont diversifiés rapidement et en particulier à la fin du Mézosoïque et Paléogène (Ponder et al., 2008), ce qui pourrait expliquer en partie la difficulté à résoudre les relations phylogénétiques dans ce groupe, et en particulier au moment de l'apparition des néogastéropodes.

Au sein des Caenogastropoda, l'ordre des Neogastropoda est composé de 7 super-familles acceptées actuellement, incluant les Buccinoidea, Conoidea, Mitroidea, Muricoidea, Olivoidea, Turbinelloidea et Volutoidea. Un total de 60 familles actuelles a été décrit (Bouchet et al., 2017), avec des nombres d'espèces très variables selon les familles (Tableau 1).

Tableau 1 : Nombre d'espèces et de genres actuels pour chaque famille et super-famille au sein des Neogastropoda (MolluscaBase, disponible avec le lien suivant : https://www.molluscabase.org/), Tonnoidea et Ficoidea, phylogénétiquement proches des néogastéropodes (voir partie suivante).

Super-famille	*Famille*	*Nombre de genres*	*Nombre d'espèces*
BUCCINOIDEA	Austrosiphonidae	4	31
	Belomitridae	1	30
	Buccinanopsidae	2	7
	Buccinidae	42	442
	Busyconidae	6	16
	Chauvetiidae	1	42
	Colidae	3	27
	Colubrariidae	9	115
	Columbellidae	75	950
	Cominellidae	3	28
	Dolicholatiridae	3	31
	Eosiphonidae	10	55
	Fasciolariidae	66	559
	Melongenidae	8	36
	Nassariidae	30	669
	Pisaniidae	19	203
	Prodotiidae	7	22
	Prosiphonidae	28	122
	Retimohniidae	3	31
	Tudiclidae	9	84
CONOIDEA	Borsoniidae	30	232
	Bouchetispiridae	1	2
	Clathurellidae	16	231
	Clavatulidae	14	113
	Cochlespiridae	7	63
	Conidae	8	1054
	Conorbidae	2	7
	Drilliidae	35	558
	Fusiturridae	1	6
	Horaiclavidae	26	210
	Mangeliidae	69	788
	Marshallenidae	1	3
	Mitromorphidae	7	156
	Pseudomelatomidae	53	554
	Raphitomidae	78	875
	Terebridae	19	568
	Turridae	15	209

MITROIDEA	Charitodoronidae	1	7
	Mitridae	34	444
	Pyramimitridae	3	10
MURICOIDEA	Muricidae	194	1968
OLIVOIDEA	Ancillariidae	10	176
	Bellolividae	3	17
	Benthobiidae	2	7
	Olividae	11	284
	Pseudolividae	5	10
TURBINELLOIDEA	Columbariidae	5	61
	Costellariidae	18	621
	Ptychatractidae	5	33
	Turbinellidae	3	10
	Vasidae	10	33
	Volutomitridae	8	65
VOLUTOIDEA	Cancellariidae	50	362
	Cystiscidae	13	483
	Granulinidae	6	133
	Marginellidae	25	1251
	Marginellonidae	3	4
	Volutidae	49	423
[unassigned]	Babyloniidae	2	20
	Harpidae	3	62
	Strepsiduridae	1	2
	[unassigned]	1	1
TONNOIDEA	Bursidae	15	69
	Cassidae	13	100
	Charoniidae	1	5
	Cymatiidae	23	137
	Laubierinidae	4	5
	Personidae	2	24
	Ranellidae	3	5
	Thalassocyonidae	2	5
	Tonnidae	3	35
FICOIDEA	Ficidae	1	12
TOTAL		1243	16013

La répartition des néogastéropodes est mondiale, dans toutes les mers et les océans du globe. On peut les retrouver à des profondeurs variées, des côtes jusqu'aux profondeurs abyssales, comme la famille des Raphitomidae (Criscione et al., 2021). La majeure partie de la diversité spécifique de cet ordre est concentrée dans la zone intertropicale. Les néogastéropodes ont une grande diversité de formes et de couleurs (Figure 2), et certains groupes (comme les cônes) sont particulièrement prisés des collectionneurs.

Leur taille est également très variable, allant d'une taille inférieure à 10 mm, voire 5 mm (Cystiscidae, Marginellidae, Raphitomidae) à une taille supérieure à 200 mm comme chez les familles des Muricidae ou Buccinidae par exemple (Y. I. Kantor et al., 2013; Oliverio, 2009).

Les néogastéropodes sont des prédateurs actifs, dont la plupart des espèces sont carnivores. Ce régime alimentaire est associé à l'évolution de traits morphologiques tels que l'allongement du canal siphonal, un retournement de l'ouverture de la bouche à l'avant de la tête ainsi que la formation d'un proboscis très développé (Cunha et al., 2009; Oliverio & Modica, 2010; Ponder et al., 2008; Ponder & Lindberg, 1997; Strong, 2003). L'évolution du proboscis chez les gastéropodes s'est produite à plusieurs reprises, de manière indépendante et elle est associée au régime alimentaire carnivore (Taylor, 1993). Certains néogastéropodes peuvent avoir des phases alimentaires de charognards comme les membres de la famillle des Nassariidae (Morton, 2003; Morton & Jones, 2003), et une lignée au sein de la famille des Columbellidae a même évolué secondairement de la carnivorie vers l'herbivorie (deMaintenon, 1999; Russini et al., 2017).

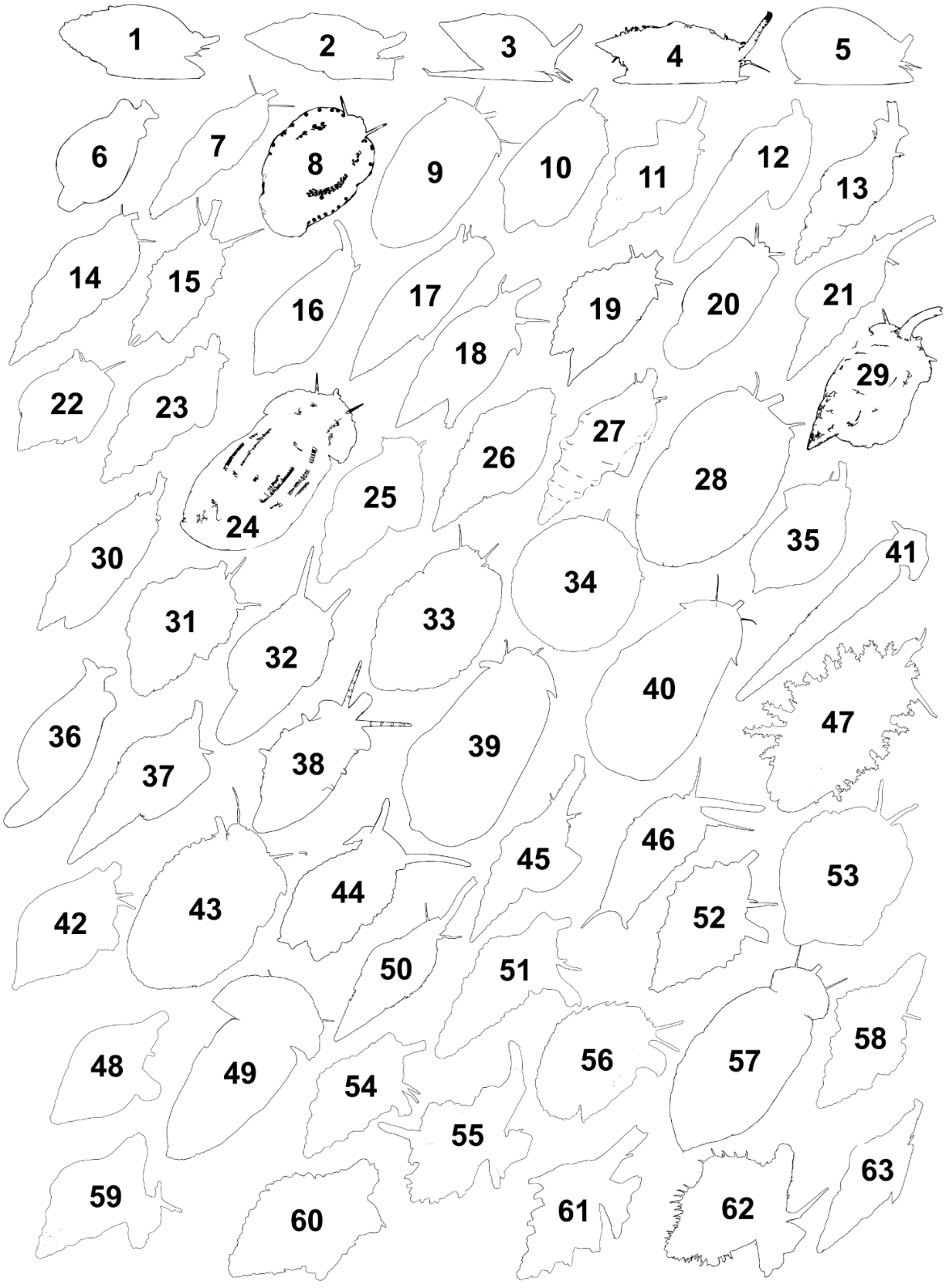

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
41
36
37
38
39
40
47
42
43
44
45
46
53
48
49
50
51
52
54
55
56
57
58
59
60
61
62
63

Figure 2 : Planche représentant les néogastéropodes vivants.

Cette planche inclut également des spécimens de taxons de non-néogastéropodes proches : Tonnoidea (Tonnidae, Cassidae, Bursidae, Personidae) ; Calyptraeoidea (Calyptraeidae) ; Cypraeoidea (Cypraeidae, Ovulidae) ; Capuloidea (Capulidae). 1 : Mitridae, 2 : Columbellidae, 3 : Nassariidae, 4 : Costellariidae, 5 : Mitridae, 6 : Marginellidae, 7 : Colubrariidae, 8 : Ovulidae, 9 : Tonnidae, 10 : Volutidae, 11 : Clathurellidae, 12 : Terebridae, 13 : Clathurellidae, 14 : Colubrariidae, 15 : Nassariidae, 16 : Conidae, 17 : Mitridae, 18 : Raphitomidae, 19 : Raphitomidae, 20 : Cystiscidae, 21 : Pseudomelatomidae, 22 : Cassidae, 23 : Mangeliidae, 24 : Harpidae, 25 : Horaiclavidae, 26 : Borsoniidae, 27 : Drilliidae, 28 : Tonnidae, 29 : Volutidae, 30 : Mangeliidae, 31 : Bursidae, 32 : Cancellariidae, 33 : Ovulidae, 34 : Calyptraeidae, 35 : Mitridae, 36 : Cystiscidae, 37 : Turridae, 38 : Marginellidae, 39 : Olividae, 40 : Olividae, 41 : Terebridae, 42 : Columbellidae, 43 : Tonnidae, 44 : Cancellariidae, 45 : Turridae, 46 : Nassariidae, 47 : Cypraeidae, 48 : Mitromorphidae, 49 : Harpidae, 50 : Columbellidae, 51 : Chauvetiidae, 52 : Muricidae, 53 : Velutinidae, 54 : Pisaniidae, 55 : Muricidae, 56 : Capulidae, 57 : Harpidae, 58 : Fasciolariidae, 59 : Tudiclidae, 60 : Buccinidae, 61 : Muricidae, 62 : Personidae, 63 : Fasciolariidae. Crédits photos : MNHN. Auteurs : Delphine Brabant, Barbara Buge, Laurent Charles, Gilles Devauchelle, Anne Lise Fleddum, Jean-Jacques Lemasson, Philippe Maestrati.

Les modifications de la radula chez les néogastéropodes, qui est un organe constitué d'une lame basale munie de nombreuses dents chitineuses (Bandel, 1984), ont également été influencées par le régime alimentaire carnivore, qui est une adaptation aux ressources alimentaires disponibles dans le milieu. La radula a des formes très variables, ces variations étant *a priori* liées à cette adaptation, avec des proies différentes selon les groupes (Bouchet et al., 2011; A. Fedosov et al., 2018; A. E. Fedosov, Malcolm, et al., 2019; Y. Kantor et al., 2017, 2022; Y. I. Kantor, Puillandre, et al., 2012; Y. I. Kantor, Strong, et al., 2012; Y. I. Kantor et al., 2013, 2022). On constate une réduction du nombre de rangées de dents, avec plusieurs configurations principales, comme le type rachiglosse, comprenant entre une à trois dents, ou le type toxoglosse qui est présent seulement dans la super-famille des Conoidea (Fedesov & Kantor, 2008; Holford et al., 2009; Y. I. Kantor & Taylor, 1991), allant jusqu'à une forme similaire à une seringue hypodermique que les cônes, par exemple, utilisent pour injecter le venin dans leurs proies (Bondarev I., 2001; Puillandre et al., 2014).

Il a été montré que plusieurs lignées de néogastéropodes produisent des composés moléculaires pour capturer leurs proies, comme les neurotoxines relativement bien connues chez les cônes, mais aussi des anesthésiants ou encore des anticoagulants dans d'autres groupes (Bose et al., 2017; M. V. Modica et al., 2018; Olivera et al., 2017). Cette adaptation constitue l'hypothèse principale pour expliquer le succès évolutif du groupe, hypothèse qui est testée dans le cadre du projet HYPERDIVERSE. La production de ces différents composés moléculaires par les néogastéropodes leur donne accès à une grande variété de proies dont ils vont se nourrir, comme des crustacés, des échinodermes, d'autres mollusques, de vers marins et parfois même de poissons (Duda et al., 2001; Duda & Palumbi, 2004; Kraus et al., 2011; Phuong et al., 2016; Vallejo, 2005). Le groupe des néogastéropodes a longtemps été défini sur la base du partage de certains caractères comme la nature de la radula, et certains caractères anatomiques qui sont partagés par de nombreux groupes, tels que les glandes salivaires accessoires tubulaires, la valve et la glande de Leiblein ainsi que la glande rectale (Y. Kantor et al., 2017, 2022; Y. I. Kantor et al., 2013, 2014; Y. I. Kantor, Puillandre, et al., 2012; Y. I. Kantor, Strong, et al., 2012).

2. PHYLOGENIES ET CLASSIFICATIONS

Au début des années 2000, les premières phylogénies moléculaires des néogastéropodes se basent sur un seul marqueur moléculaire : ARN 12S ou 16S ou cytochrome oxydase I (Duda & Kohn, 2005; Duda & Rolán, 2004; Espiritu et al., 2001; Hayashi, 2005; Oliverio et al., 2002). Ces phylogénies ciblaient en général une seule famille chacune, mais l'utilisation d'un seul marqueur ne permettait pas d'établir des phylogénies sur de plus grandes profondeurs phylogénétiques. À partir de la fin des années 2000 et le début des années 2010, des phylogénies multi marqueurs ont été proposées (Barco et al., 2010; Claremont et al., 2011; Claremont, Houart, et al., 2013; Claremont, Vermeij, et al., 2013; Cunha et al., 2005, 2007; A. Fedosov et al., 2015; Galindo et al., 2016; Y. I. Kantor, Strong, et al., 2012; Kraus et al., 2011; H. Li et al., 2010; M. V. Modica et al., 2009, 2011; Nam et al., 2009; Oliverio & Modica, 2010; Pereira et al., 2010; Puillandre et al., 2015; Zou et al., 2011). Ces phylogénies ont permis d'établir des hypothèses taxonomiques sur de nombreuses familles voire des super-familles. Cependant, le faible nombre de marqueurs ne permettaient toujours pas d'avoir des résolutions phylogénétiques suffisantes pour résoudre plus précisément certaines relations phylogénétiques

complexes. À la même période, le séquençage de mitogénomes complets a permis de produire des phylogénies à des échelles taxonomiques parfois plus larges (Abalde et al., 2017; Bandyopadhyay et al., 2006; Choi et al., 2021; Cunha et al., 2009; Harasewych et al., 2019; Uribe et al., 2018; Vaux et al., 2018). C'est le cas pour la phylogénie publiée par Osca et al. (Osca et al., 2015), basée sur un échantillonnage à l'échelle de l'ensemble des Caenogastropoda. Plus récemment, l'approche de capture d'exons (voir ci-dessous et chapitres suivants) a été utilisée dans différents groupes pour produire des phylogénies à des échelles taxonomiques plus profondes (Abdelkrim et al., 2018; Phuong & Mahardika, 2018; Zaharias et al., in press).

Au cours de la dernière décennie, ces phylogénies moléculaires, accompagnées de révisions taxonomiques pour certaines, ont permis d'améliorer la compréhension des relations phylogénétiques entre les groupes au sein des néogastéropodes à différentes échelles taxonomiques. Cela a été le cas au sein de différentes super-familles telles que les Buccinoidea (Couto et al., 2016; Galindo et al., 2016; Y. I. Kantor et al., 2022, p. 2), les Conoidea (Abdelkrim et al., 2018; Bouchet et al., 2011; Y. I. Kantor, Strong, et al., 2012; Puillandre et al., 2011; M. Yang et al., 2021), les Mitroidea (A. Fedosov et al., 2015, 2018; Y. I. Kantor et al., 2014), les Muricoidea (Barco et al., 2010) et les Olivoidea (Y. Kantor et al., 2017). Les phylogénies moléculaires ont permis également de réviser la taxonomie au niveau familial, par exemple chez les Cancellariidae (M. V. Modica et al., 2011), Costellariidae (A. Fedosov et al., 2015) et les marginelles (Cystiscidae, Granulinidae, Marginellidae, Marginellonidae — (A. E. Fedosov, Caballer Gutierrez, et al., 2019). Tous les groupes que j'ai cité plus haut se retrouvent monophylétiques dans les différentes études et cela a permis de réviser en profondeur les classifications à des niveaux familiaux et génériques.

Il reste néanmoins de nombreuses relations phylogénétiques au niveau super-familial et/ou familial à démêler. C'est le cas, par exemple, des Turbinelloidea dont la monophylie reste à confirmer à l'heure actuelle. Plusieurs autres familles n'ont pas encore été étudiées sur la base de phylogénies moléculaires, c'est le cas des Babyloniidae, des Harpidae, des Strepsiduridae et des Volutidae en particulier.

De plus, les phylogénies à l'échelle des néogastéropodes que je viens de présenter sont réalisées à partir d'échantillonnages limités ; soit parce qu'elles se concentrent seulement sur une famille ou un groupe de familles en particulier, avec quelques autres néogastéropodes en groupes externes, soit parce que les phylogénies à l'échelle des néogastéropodes et des Caenogastropoda

n'incluaient que peu de représentants, y compris aux échelles des familles et des super-familles. C'est le cas des phylogénies multi-marqueurs, comme par exemple la phylogénie des Mitriformes (A. Fedosov et al., 2015), dans laquelle certains clades à l'échelle des super-familles sont supportés (Buccinoidea, Mitroidea et Olivoidea), alors que d'autres, comme les Turbinelloidea, ne le sont pas. Dans la phylogénie des Olivoidea (Y. Kantor et al., 2017), le clade Olivoidea est supporté mais les autres groupes de néogastéropodes ainsi que les groupes externes ne le sont pas. La phylogénie des Conoidea à partir de la capture d'exons (Abdelkrim et al., 2018) ne permet pas de résoudre les relations phylogénétiques avec certaines super-familles utilisées en groupes externes, comme les Mitroidea et les Olivoidea. Pour la phylogénie des marginelles publiée en 2019 (A. E. Fedosov, Caballer Gutierrez, et al., 2019), on constate que la super-famille des Volutoidea (sauf les Cancellariidae) se retrouve bien comme un clade supporté. En revanche, les super-familles des Tonnoidea et la famille des Cancellariidae se retrouvent dans le même clade alors qu'ils font partie de deux super-familles distinctes à l'heure actuelle. Enfin, tous les autres clades retrouvés ne sont pas supportés, bien qu'ils semblent se regrouper dans l'arbre, comme pour les Mitroidea, les Olivoidea et les Turbinelloidea. De même, la phylogénie des Buccinoidea publiée en 2021 (Y. I. Kantor et al., 2022) retrouve bien les clades des Buccinoidea et des Turbinelloidea supportés.

Comme expliqué précédemment, le groupe-frère des néogastéropodes n'est pas connu. Des phylogénies récentes suggèrent que les groupes des Tonnoidea et Ficoidea pourraient être très proches des néogastéropodes (Figure 3), voire inclus dans ce groupe (Colgan et al., 2007; Fourdrilis et al., 2018; Harasewych et al., 2019; Machkour-M'Rabet et al., 2021; Osca et al., 2015; J.-G. Wang et al., 2017; Q. Wang et al., 2021). Pourtant, d'un point de vue anatomique, il semblerait que ces deux super-familles soient bien différentes des autres néogastéropodes. Pour cette raison, nous avons choisi d'ajouter les Tonnoidea et les Ficoidea dans l'étude, au même titre (c'est-à-dire avec le même effort d'échantillonnage – voir chapitre 2) que les néogastéropodes. Plus généralement, nous avons aussi décidé d'inclure un grand nombre de groupes proches parmi les Caenogastropoda.

Par ailleurs, plusieurs phylogénies basées sur les mitogénomes ne parviennent pas à résoudre les relations phylogénétiques entre certaines super-familles et regroupent même entre elles des super-familles éloignées. C'est le cas avec la phylogénie des néogastéropodes publiée en 2009 (Cunha et al., 2009) qui se base sur la traduction en acides aminés des 13 gènes mitochondriaux. Dans cette phylogénie, les spécimens de la super-famille des Conoidea ne sont pas monophylétiques, or plusieurs études ont confirmé la monophylie de cette super-famille

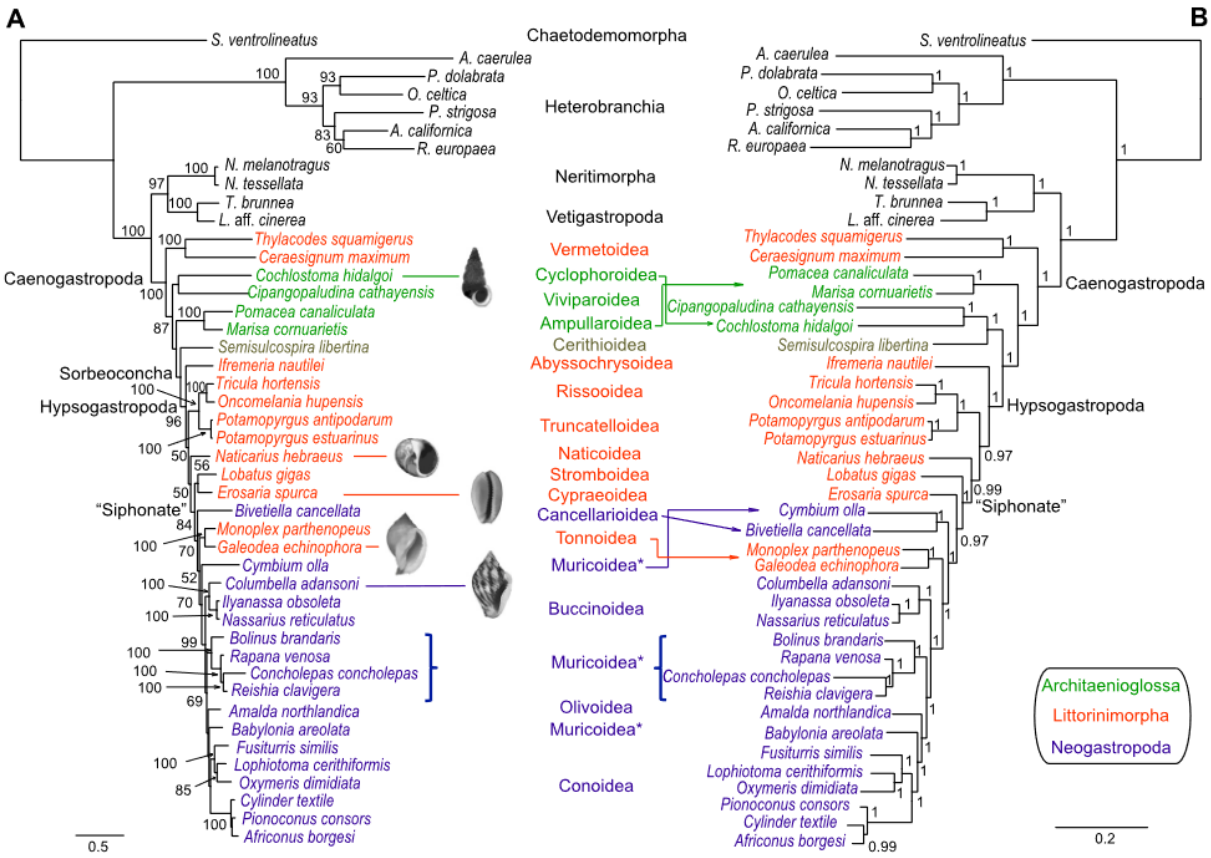


Figure 3 : Phylogénie des Caenogastropoda reconstruite à partir des données mitogénomiques. Modifiée à partir de Osca et al. 2015.

(Abdelkrim et al., 2018; A. E. Fedosov et al., in press; Puillandre et al., 2011). En outre, les Buccinoidea se retrouvent intercalés au sein des deux groupes de Conoidea retrouvés. C'est également le cas de la phylogénie des Caenogastropoda (Osca et al., 2015) : dans cette phylogénie les néogastéropodes ne sont pas monophylétiques, car la super-famille des Tonnoidea se retrouve au sein des Neogastropoda. De plus, la super-famille des Muricoidea n'est pas monophylétique dans la phylogénie bayésienne.

3. PHYLOGENIE BASEE SUR LES MITOGENOMES

Nous avons fait face à des contretemps au début de ma thèse. Les expérimentations au laboratoire prévues au cours des 6 premiers mois ont été repoussées, en premier lieu à cause de travaux dans le laboratoire moléculaire, mais aussi à cause de déménagement du Service de Systématique Moléculaire du MNHN. Cependant, j'avais la volonté de me former à la

bioinformatique et à l'apprentissage de langages de programmation (Python), et j'ai profité de ces contretemps pour le faire. J'ai pu commencer cette formation dans le cadre de l'identification des exons nécessaires à l'approche par capture d'exons (voir chapitre 2). Sachant que les mitogénomes sont une source de données pertinentes à ce niveau phylogénétique (Zaharias et al., 2020), j'ai décidé de mettre à profit cette période de ma thèse pour proposer une phylogénie moléculaire des néogastéropodes à partir des mitogénomes, celle-ci pouvant être réalisée à partir de données déjà disponibles, et ne nécessitant donc pas d'accès à un laboratoire de biologie moléculaire, tout en poursuivant ainsi ma formation à la bioinformatique. En effet, de nombreux mitogénomes de néogastéropodes et de groupes proches étaient déjà présents dans la base de données de GenBank, et nous avions aussi à notre disposition des transcriptomes qui avaient été produits dans le cadre d'autres projets au sein de l'équipe. Nous avons donc décidé de mettre en place un pipeline bioinformatique afin d'extraire les mitogénomes les plus complets possibles à partir de ces transcriptomes, avec pour objectif d'obtenir des séquences pour des familles qui n'étaient pas disponibles dans Genbank et donc enrichir taxonomiquement le jeu de données. Ce travail a été valorisé dans un article publié dans la revue Zoologica Scripta.

Received: 22 March 2022 | Revised: 24 May 2022 | Accepted: 13 June 2022

DOI: 10.1111/zsc.12552

Zoologica Scripta | WILEY

ORIGINAL ARTICLE

Neogastropod (Mollusca, Gastropoda) phylogeny: A step forward with mitogenomes

Thomas Lemarcis[1] | **Alexander E. Fedosov**[1,2] | **Yuri I. Kantor**[1,2] | **Jawad Abdelkrim**[3] | **Paul Zaharias**[1] | **Nicolas Puillandre**[1]

[1]Institut Systématique Evolution Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, CNRS, EPHE, Sorbonne Université, Université des Antilles, Paris, France

[2]A. N. Severtsov Institute of Ecology and Evolution, Russian Academy of Sciences, Moscow, Russia

[3]UAR 2700 'Acquisition et Analyse de Données pour l'Histoire naturelle', Service d'Analyse de Données, CNRS, Muséum National d'Histoire Naturelle, Sorbonne Universités, Paris Cedex, France

Correspondence
Thomas Lemarcis, Institut Systématique Evolution Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, 57 rue Cuvier, CP 51, Paris 75005, France.
Email: thomas.lemarcis@mnhn.fr

Funding information
H2020 European Research Council, Grant/Award Number: 865101; French Agence Nationale de la Recherche, France, Grant/Award Number: ANR-13-JSV7-0013-01

Abstract

The Neogastropoda (Mollusca, Gastropoda) encompass more than 15,000 described species of marine predators, including several model organisms in toxinology, embryology and physiology. However, their phylogenetic relationships remain mostly unresolved and their classification unstable. We took advantage of the many mitogenomes published in GenBank to produce a new molecular phylogeny of the neogastropods. We completed the taxon sampling by using an in-house bioinformatic pipeline to retrieve mitochondrial genes from 13 transcriptomes, corresponding to five families not represented in GenBank, for a final dataset of 113 taxa. Because mitogenomic data are prone to reconstruction artefacts, eight different evolutionary models were applied to reconstruct phylogenetic trees with IQTREE, RAxML and MrBayes. If the over-parametrization of some models produced trees with aberrant internal long branches, the global topology of the trees remained stable over models and softwares, and several relationships were revealed or found supported here for the first time. However, even if our dataset encompasses 60% of the valid families of neogastropods, some key taxa are missing and should be added in the future before proposing a revision of the classification of the neogastropods. Our study also demonstrates that even complex models struggle to satisfactorily handle the evolutionary history of mitogenomes, still leading to long-branch attractions in phylogenetic trees. Other approaches, such as reduced-genome strategies, must be envisaged to fully resolve the neogastropod phylogeny.

KEYWORDS
mitogenomes, Neogastropoda, phylogeny, systematics

1 | INTRODUCTION

Mitogenomes constitute one of the best compromises between informativeness and cost (both in terms of time and money) when it comes to selecting a suitable set of characters to resolve phylogenies (Zaharias et al., 2020). If it remains not informative enough for very deep relationships and is sometimes prone to long-branch artefacts (Schrödl

& Stöger, 2014; Uribe et al., 2019), it has been widely used in various groups of organisms, and in particular in molluscs, where it helped to clarify relationships from the species level (Abalde et al., 2017) to deeper relationships between lineages separated since several hundreds of MY (Uribe et al., 2019; Williams et al., 2014). And even in an era when datasets based on transcriptomes or reduced genomes are becoming more and more common, new mitogenomes are being produced regularly. Furthermore, in gastropods, the publication of a new mitogenome(s) provides often the opportunity to produce a new phylogeny by combining the newly produced mitogenome(s) with previously published ones (Barghi et al., 2016; Choi et al., 2021; Harasewych et al., 2019; Huang et al., 2021; Wang et al., 2021).

However, such publications generally limit the scope of the phylogeny to a few mitogenomes among those available in public databases, and the phylogenetic analyses conducted are often limited to one method, and do not test for various methods and/or models in order to more thoroughly reconstruct phylogenetic relationships. One of the groups that would benefit from such a more complete analysis are the neogastropods. This crown group of the caenogastropods is composed of more than 15,000 described species, mostly predators, and their evolutionary success has regularly been suspected to be linked to their capacity to produce various molecular compounds to capture their prey, such as anaesthetics, anticoagulants and neurotoxins (Bose et al., 2017; Modica et al., 2018; Olivera et al., 2017). However, the lack of a clear and robust phylogenetic context for the group, illustrated by their unstable superfamily and family-level classification, makes it difficult to test such hypothesis (Kuznetsova et al., 2022).

The current state-of-the-art of the neogastropod classification, as provided in WoRMS (WORMS, 2018, consulted on the 14 March 2022), is mostly based on the molecular phylogenies published in the last 10 years and the associated taxonomic revisions. Thus, the superfamilies Conoidea (Abdelkrim et al., 2018; Bouchet et al., 2011; Kantor et al., 2012; Puillandre et al., 2011; Yang et al., 2021), Buccinoidea (Couto et al., 2016; Galindo et al., 2016; Kantor et al., 2021; Oliverio & Modica, 2010), Mitroidea (Fedosov et al., 2015, 2018; Kantor et al., 2014), Olivoidea (Kantor et al., 2017), Muricoidea (Barco et al., 2010), the families Cancellariidae (Modica et al., 2011), Costellariidae (Fedosov et al., 2015, 2017) and the marginellids (Cystiscidae, Granulinidae, Marginellidae, Marginellonidae—Fedosov et al., 2019) have been found monophyletic, and the family- and/or genus-level classifications have been deeply revised. In some cases, the reconstructed phylogenetic relationships necessitated the introduction of a large number of new family-level taxa, in particular within the Conoidea and Buccinoidea, that switched from 7 and 9 families, respectively, to 17 and 20. The superfamily Turbinelloidea still awaits to be revised, although Fedosov et al. (2015, 2017) already started the work, and only a few families remain virtually untouched by molecular phylogeneticians (Volutidae, Harpidae, Babyloniidae and Strepsiduridae—but see Ravitchandirane and Sukumar (2013) for the Babyloniidae).

But the main challenge of the neogastropod phylogenetic reconstruction is to clarify the relationships at the superfamily level: three families (Harpidae, Babyloniidae and Strepsiduridae) remain unassigned in a superfamily and the relationships between the seven currently recognized superfamilies (and the three unassigned families) are virtually unknown. All the previously published phylogenetic studies that, at least partly, dealt with the neogastropod superfamily relationships failed to resolve most of these deep relationships. Two notable exceptions are (a) the recovery of the sister-taxa Tonnoidea (also revised at the family level—Strong et al., 2019) and Ficoidea, two superfamilies currently not recognized as a Neogastropoda, within the neogastropod clade (as sister to all the other neogastropods but the Cancellariidae; Colgan et al., 2007; Fourdrilis et al., 2018; Harasewych et al., 2019; Machkour-M'Rabet et al., 2021; Osca et al., 2015; Wang et al., 2017, 2021); (b) the monophyly of a group that includes all the neogastropods except the Volutoidea (Cancellariidae, marginellids, Volutidae), the Tonnoidea and the Ficoidea (Abdelkrim et al., 2018; Choi et al., 2021; Cunha et al., 2009; Fourdrilis et al., 2018; Harasewych et al., 2019; Machkour-M'Rabet et al., 2021; Osca et al., 2015; Uribe et al., 2021; Wang et al., 2017, 2021).

In order to further improve our understanding of the neogastropod phylogenetic relationships, we propose to take advantage of the many neogastropod mitogenomes available in GenBank (359 on January 28th, 2022), complemented with several mitogenomes extracted from transcriptomic data obtained from GenBank or other projects of our team (unpub.), to reconstruct the most complete phylogeny of neogastropods published to date. Even if the sampling remains incomplete, as several families of neogastropod are not represented in this mitogenome dataset, this phylogeny certainly constitutes a step forward, and highlights some previously unnoticed relationships that allow us discussing the evolution of morpho-anatomical characters and elaborating further on the evolutionary success of the neogastropods.

2 | MATERIAL AND METHODS

2.1 | Sampling

All the neogastropod, tonnoidean and ficoidean mitogenomes from GenBank were downloaded on 28 January

2022. A selection of mitogenomes of Littorinimorpha, and in particular the most closely related lineages to Neogastropoda (Osca et al., 2015) were added as outgroups. Following Osca et al. (2015), *Ifremeria nautilei* (Provannidae) was used to root all the phylogenetic trees. The family, genus and species names were updated according to WoRMS. To complement the taxonomic sampling, mitochondrial gene sequences extracted from publicly available transcriptomes or transcriptomic data produced in the framework of other projects by our team were added to the dataset. The complete list of mitogenomes from GenBank and mitochondrial gene sequences extracted from transcriptomes is provided in Appendix S1.

2.2 | Mitogenomes extracted from transcriptomic data

To complete the taxonomic sampling, 13 partial mitogenomes were extracted from transcriptomes (see Appendix S1). We designed a reference-based approach with one to four mitogenomes used as reference for each transcriptome. A suite of custom python scripts to reproduce the pipeline is available at https://github.com/Hyperdiverseproject/Neo_mitogenomes/.

First, a BLASTn (Camacho et al., 2009) search of the transcriptomes against the reference mitogenomes was performed and the contig with the best hit (e-value < 1e-10) was retained. In parallel, a BLASTp search was also performed. Assembled contigs were translated from transcriptomes using open reading frames (ORF)finder (Rombel et al., 2002), and translated genes were retrieved from reference mitogenomes directly from GenBank.

To ensure contig quality and avoid contaminations, raw reads from transcriptomes were mapped against the best hit contigs from BLASTn and BLASTp searches using Bowtie2 (Langmead & Salzberg, 2012). The depth of coverage was calculated with the option 'coverage' of samtools (Danecek et al., 2021). If two contigs from the same transcriptome aligned in the same region of the reference mitogenome, the contig with the highest depth of coverage was selected. Hit contigs were then merged and aligned against the mitogenome references with multiple alignment using Fast Fourier Transform (MAFFT) and 'localpair' and 'addfragments' options (Katoh & Frith, 2012).

A second round of filtering was performed by calculating the number of differences ('genetic distance') between the contig sequence and the sequence of the closest reference mitogenome: if the contig had genetic distance >50% of the sequence length, the contig was removed from the alignment. Remaining contigs were realigned against the reference mitogenomes and contigs not aligned with a mitochondrial gene were removed from the alignment. The pipeline resulted in 13 partial mitogenomes extracted from transcriptomes (Appendix S1).

2.3 | Subsampling and data cleaning

Single mitochondrial genes (13 protein-coding genes and ribosomal genes) were extracted from GenBank mitogenomes using AnnotationBustR (Borstein & O'Meara, 2018). Given that some lineages (species, genera) are overrepresented in the GenBank mitogenome dataset, we first performed a phylogenetic analysis (Neighbour-Joining with MEGA—Kumar et al., 2016) to confirm that conspecific and congeneric mitogenomes were clustering together, and then to select one mitogenome per genus, the one corresponding to the type species and/or with less missing data. One exception is the genus *Turricula*, for which two available (partial) mitogenomes do not cluster together: one (extracted from a transcriptome) cluster with the other Clavatulidae, the other (*Turricula nelliae spurius*, MK251986) within the Pseudomelatomidae. Given the morphological resemblance of *Turricula nelliae* with some members of Pseudomelatomidae (e.g. *Comitas*), we suspect here a misidentification. In any case, both mitogenomes were retained in the analyses. The resulting dataset comprised 114 taxa, including 10 outgroups.

Each gene was then aligned using MAFFT v7.490 (Katoh & Standley, 2013) auto mode and phylogenetic trees were reconstructed with IQ-TREE v2.1.3 (Minh et al., 2020) with General Time Reversible (GTR) + Gamma (G) in order to detect potential misalignments and contaminations. Indels in coding genes (when not a multiple of 3) were manually removed, as well as sequence fragments at the beginning or end of the alignment that were out of the gene ORF. Two gene fragments were also removed from the dataset: the first 600 nucleotides (NTs) of the cytochrome *c* oxidase subunit III (*cox3*) gene of JQ446041 (*Concholepas concholepas*) had no blast hit in GenBank; the second part of the cytochrome *c* oxidase subunit I (*cox1*) of MW316798 (*Aspa marginata*) was obviously misaligned, leading to a very long branch. We also removed the mitogenome of *Ceraesignum maximum* (Vermetidae, HM174253), because it constituted a very long branch and was not crucial for neogastropod phylogeny, as an early branching taxa within Caenogastropoda (Osca et al., 2015; Rawlings et al., 2010), thus leaving 113 taxa in the final dataset.

2.4 | Phylogenetic analyses

Sequences from coding genes were translated from NTs to amino acids (AAs) using the translateNT2AA program

implemented in MACSE (Ranwez et al., 2011) and then aligned with MAFFT using the G-INS-i algorithm. We used MAFFT E-INS-i for non-coding mitochondrially encoded 12S and 16S RNA. The program reportGapsAA2NT in Multiple Alignment of Coding SEquences Accounting (MACSE) was used to derive each NT alignment from each MAFFT AA alignment. Finally, the 15 genes were concatenated in two ways: only NTs (all 15 genes) or AAs for the 13 coding genes + 12S and 16S in NTs. We will refer to the two datasets as the NT matrix or AA + NTs matrix.

We performed a first round of analyses using IQ-TREE to determine the best fit model strategy for the NT matrix. In all cases, ModelFinder Plus (MFP) was used to select the best model of substitution (Kalyaanamoorthy et al., 2017). The dataset was partitioned in eight different ways (Table 1): a single partition for the entire NT ('single-partition + MFP'), partitioning by gene with edge-proportional ('gene-partitioned + MFP) and edge-unlinked ('gene partitioned + MFP + edge-unlinked') partition models (Chernomor et al., 2016), partitioning by codon (except for 12S and 16S) with edge-proportional ('codon partitioned + MFP') and edge-unlinked ('codon partitioned + MFP + edge-unlinked') partition models, partitioning by codon (except for 12S and 16S) with a selection of the best-fit partitioning scheme by merging partitions with edge-proportional ('codon partitioned + MFP + MERGE') and edge-unlinked ('codon partitioned + MFP + MERGE + edge-unlinked') partition models, by using the GHOST model with four classes, unlinked branch lengths, substitution rates and inferred base frequencies ('GHOST'; Crotty et al., 2019). The GHOST model was specifically designed to take into account heterotachy, that is, variation of the evolutionary rate of site through time (Lopez et al., 2002). Heterotachy is a process that is likely to have occurred in large groups and in fast-evolving genomes such as the mitochondrial genome. Each analysis was run with 1000 standard bootstraps, except for GHOST due to numerical underflow issues. Because the number of free parameters can greatly change from one model to another (Table 1), the Log-Likelihood scores alone are not appropriate to compare the different models. Instead, we ranked the Akaike information criterion (AICc) and Bayesian information criterion (BIC) scores outputted by IQ-TREE for each analysis and compared them (Table 1). The best AICc score was found for the codon partitioned + MFP analysis while the best BIC score was found for the codon partitioned + MFP + MERGE. Since there was no strong argument to support one or the other, we decided to run RAxML-ng and MrBayes on the NT matrix using both the codon partitioned + MFP and the codon partitioned + MFP + MERGE.

The AA matrix data were run with the same strategy for IQ-TREE, RAxML and MrBayes; a gene partitioned approach was used and ModelFinder identified the best fit model for each partition. Unfortunately, the AA matrix did not work with 12S and 16S in IQ-TREE due to a software error mentioning running out of RAM (and reported to the IQ-TREE google group). For both the NT and AA matrix, when the model of substitution selected by ModelFinder did not exist in either RAxML or MrBayes, we converted the model into the equivalent model available in each respective software.

RAxML was run using the 'all-in-one' analysis flag implying the use of bootstopping criterion (Pattengale et al., 2010). The bayesian analyses run in MrBayes v3.2 (Ronquist et al., 2012) consisted of three parallel analyses, each with eight Markov chains of 50,000,000 generations and a sampling frequency of one tree each 10,000 generations. The number of swaps was set to 5, and the chain temperature at 0.02. We evaluated the convergence of each analyses using Tracer v1.7.1 (Rambaut et al., 2018).

TABLE 1 Log-likelihood, number of free parameters, AICc and BIC scores for the IQ-tree analyses performed with eight different partitions and models

IQ-TREE model	Log-likelihood	Nb free parameters	AICc	BIC
Single partition + MFP	−664,013.774	249	1,328,534.31	1,330,412.72
Gene partitioned + MFP	−660,006.477	540	1,321,134.938	1,325,185.616
Gene partitioned + MFP + edge-unlinked	−658,844.169	3316	1,326,294.878	1,349,452.313
Codon partitioned + MFP	−638,265.299	817	**1,278,262.591**	1,284,356.644
Codon partitioned + MFP + edge-unlinked	−632,616.015	9102	1,314,381.131	1,352,420.118
Codon partitioned + MFP + MERGE	−639,582.443	510	1,280,222.257	**1,284,050.178**
Codon partitioned + MFP + MERGE + edge-unlinked	−639,462.959	1219	1,281,588.603	1,290,602.724
GHOST	−790,917.6179	927	1,583,816.399	1,590,714.972

The best AICc and BIC scores are in bold.

Abbreviations: AICc, Akaike information criterion; BIC, Bayesian information criterion; MFP, ModelFinder Plus.

Analyses were stopped before reaching 50,000,000 generations to limit computation time, but not before the ESS values were all superior to 200, thus leading to a burnin of 24%, 43% and 4% for the NT codon partitioned + MFP, NT codon partitioned +MFP + MERGE and the AA matrix, respectively.

3 | RESULTS

The final dataset included sequences of 10 outgroups, 11 Tonnoidea (5 families), 1 Ficoidea (1 family) and 91 neogastropod (36 families) representatives. Among them, incomplete mitogenomes (between 8 and 13 genes each—Genbank accession numbers: see Appendix S1) were recovered from transcriptomes for 13 taxa, representing 10 families, including 5 (Colubrariidae, Olividae, Personidae, Pisaniidae, Turbinellidae) that were not represented in GenBank. However, 26 families of neogastropods and 4 families of Tonnoideans considered as valid in WoRMS are still not represented in our dataset.

Eight trees were obtained and will be compared (Figures 1 and S1): NT codon partitioned + MFP with IQ-TREE, RAxML and MrBayes, NT codon partitioned + MFP + MERGE with IQ-TREE, RAxML and MrBayes and AA with RAxML and MrBayes. The trees obtained with the same matrix, method or partition model are almost identical, but more differences (with only a few supported nodes—Bootstraps B < 80, Posterior Probabilities [PP] < 0.95) are found when comparing trees obtained with different matrices, methods and/or substitution models. The two NT IQTREE are identical, except for the position of Pisaniidae and Cominellidae (unsupported in both cases) within the Buccinoidea. Similarly, the two NT RAxML trees are identical, with the only difference being the Tonnoidea + Ficoidea clade sister to the Calyptraeidae + Neogastropoda clade in the codon + MFP + Merge tree whereas the Calyptraeidae are sister to the Cancellariidae in the codon + MFP tree (as in the six other trees), but these nodes are not supported. Except a few other unsupported differences, the RAxML trees are also very similar to the IQTREE trees, and all of them are also similar to the MrBayes trees. In general, the support values for the IQTREE and RAxML trees are lower than for the MrBayes trees. In the following section, the description of the results will focus on the MrBayes trees, and the IQTREE and RAxML trees (Figures 1 and S1) will be discussed only to point at the differences (or similarities) with the MrBayes trees.

In all trees, all the families represented by several taxa are monophyletic, except the Turridae (the genera *Gemmuloborsonia* and *Lucerapex* are sister to the Pseudomelatomidae + Drilliidae clade in the MrBayes AA tree with PP = 0.99, thus not clustering with the rest of the Turridae, as for example, in the MrBayes NT trees with B = 1) and the Tudiclidae (non-monophyly supported in most trees). Within families, the relationships are very consistent among trees. The relationships are less stable within superfamilies, with some unstable taxa such as the Clathurellidae and Mitromorphidae within the Conoidea, and several families within the Buccinoidea (but again, the corresponding nodes are most often not supported). For example, the Columbellidae are sister to the rest of the Buccinoidea in the NT trees (not supported), but sister to the Colubrariidae in the AA trees (PP = 0.99, B = 52).

The Buccinoidea, Conoidea, Tonnoidea, Olivoidea, Muricoidea, Mitroidea, (the latter two being represented by only one family each) are always found monophyletic, with high support. However, the superfamily Turbinelloidea, although represented by two taxa only, is never found monophyletic, with *Vasum* (Turbinellidae) included with high support in a clade together with Conoidea, Mitroidea, Buccinoidea, Olivoidea and Babyloniidae, thus excluding Costellariidae. The position of *Vasum* is not stable among trees: it is sister to the Buccinoidea (e.g. in the MrBayes NT codon + MFP tree − PP = 0.87) or sister to the Babyloniidae (e.g. in the MrBayes AA tree − PP = 1). Similarly, the Volutoidea are never found monophyletic, with the Volutidae being sister to a clade including Conoidea, Buccinoidea, Olivoidea, Turbinelloidea, Muricoidea and Babyloniidae, and the Cancellariidae being sister to the Calyptraeidae with high support in most trees.

In all trees, the clade Conoidea + Mitroidea + Buccinoidea + Olivoidea + Babyloniidae + Turbinellidae is always supported. The Muricidae, Costellariidae, Volutidae and Tonnoidea + Ficoidea are then successively branching as sister clades to the previous one, always with good support with both the AA and NT matrices (0.98 < PP < 1). The only exception is in the RAxML codon + MFP + Merge tree, as detailed above. Finally, the Cancellariidae are always sister to the Calyptraeidae with high support (0.96 < PP < 1), and all the neogastropods (except Cancellariidae) + Tonnoidea/Ficoidea are monophyletic (PP = 1).

4 | DISCUSSION

4.1 | A note on model selection for the NT matrix

While selecting a model of sequence evolution is facilitated by using model-selection methods (e.g. ModelFinder), choosing the right model-selection strategy on complex datasets such as mitochondrial genomes remains challenging. The use of information-theoretic

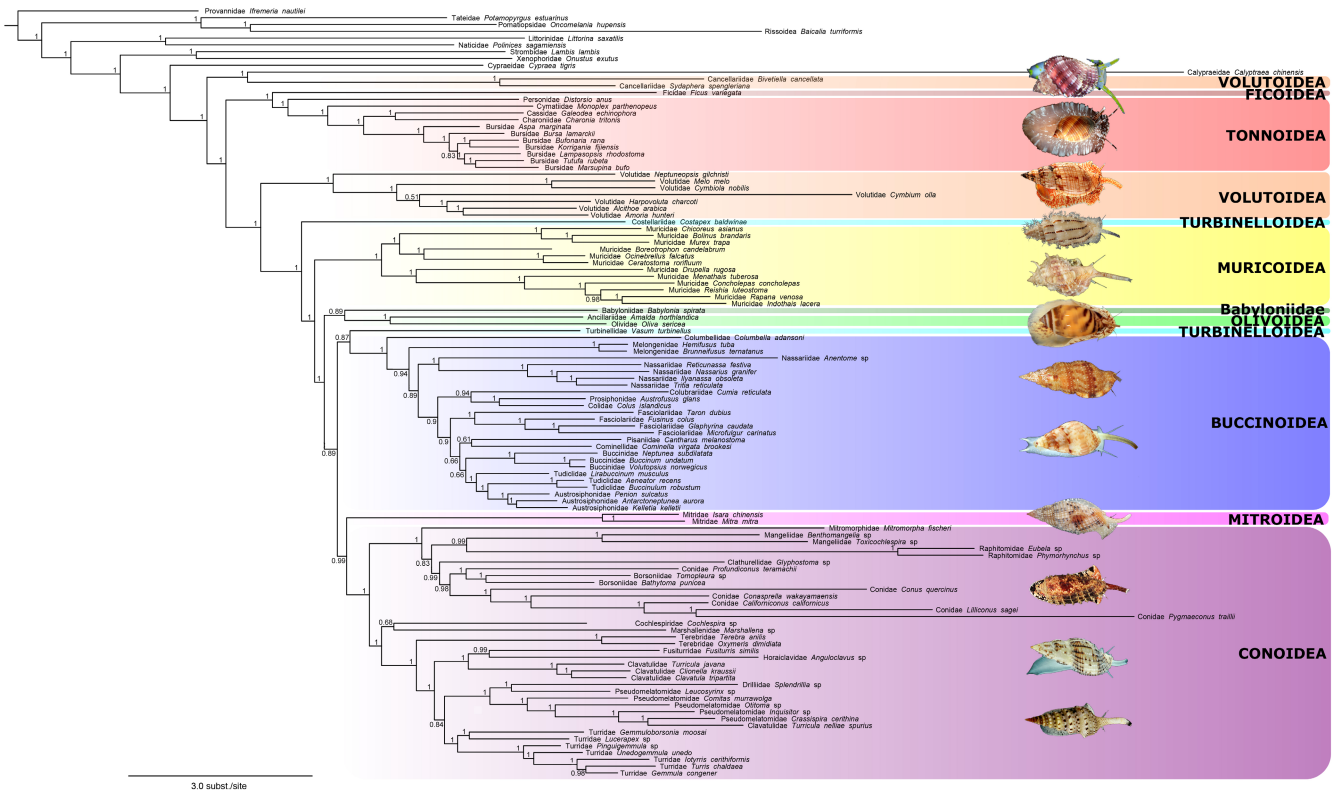
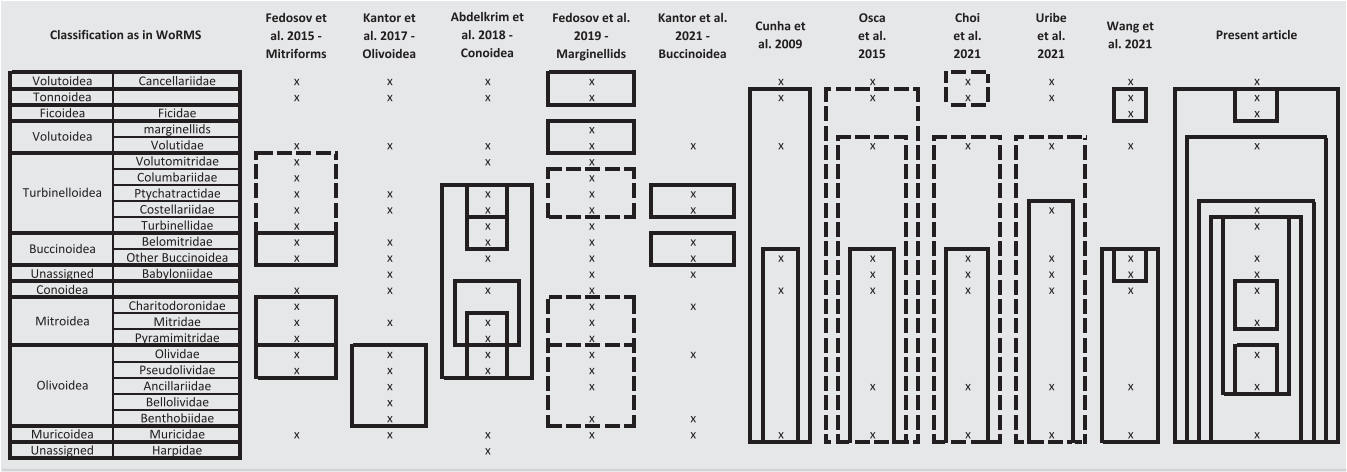


FIGURE 1 Phylogenetic tree obtained with the NT 'codon + MFP' matrix, analysed with MrBayes. The bootstraps values and posterior probabilities (>0.5) are given for each node. Superfamilies are highlighted with colour boxes. Illustrations from top to down: Cancellariidae, Tonnidae, Volutidae, Costellariidae, Muricidae, Olividae, Colubrariidae, Nassariidae, Mitridae, Conidae, Raphitomidae, Turridae *(*Credits: Philippe Maestrati, Laurent Charles /MNHN). MFP, ModelFinder Plus; NT, nucleotide

methods (Posada & Buckley, 2004) such as corrected AICc and BIC is now common practice to distinguish and select several phylogenetic models. In the first phase of our phylogenetic analyses using IQ-TREE, we used eight different strategies for analyzing the NT matrix and rank each phylogenetic model using either AICc or BIC criteria (Table 1). The implementation of the GHOST model has led to the lowest supports and topologies highly different from the others, possibly due to repeated numerical underflow errors. But GHOST is also a mixture model and some recent discussions (Crotty & Holland, 2022) questioned the use of information-theory methods to compare partition and mixture models. Similarly, models involving edge-unlinked partitions have low AICc or BIC scores, likely due to the great amount (several thousands) of parameters to evaluate. Interestingly enough, GHOST and edge-unlinked models are the only two strategies that can take heterotachy (Lopez et al., 2002) into account. Our results stress out the need for alternative heterotachy models that are less parameter-rich and more suited for datasets with high number of taxa. Obviously, using a single partition over the entire NT matrix is unrealistic, even using a FreeRate model (Le et al., 2012) with 10 categories to take more site heterogeneity into account. Thus, our options for model selection were reduced to a medium-level parametrized gene-partition strategy, a highly parametrized codon-partition strategy or an alternate medium-level parametrized strategy where codon-partitions are merged using the MERGE option in IQ-TREE. The 'codon partitioned + MF' had the best AICc score, while the 'codon partitioned + MF + MERGE' strategy had the best BIC score. Hence, we decided to keep both for subsequent analysis with RAxML and MrBayes. While the two codon-partitioned strategies were the ones favoured by information-theory criteria, we must note that all trees inferred with this partition schemes reveal unrealistic branch lengths, that is, with some internal branches having on average more than seven substitutions per site, regardless of the reconstruction method used (i.e. RAxML, IQ-TREE or MrBayes). We suspect that for some partitions, that is, the ones with high-evolving rate sites such as third codon positions, the methods fail to correctly assess the substitution rates. A careful look at some partition parameters in MrBayes show that a lot of them are not converging, even at the end of the analyses. In spite of the obviously overparametrized model that was used and unrealistic branch lengths, the topologies (which

TABLE 2 Comparison of neogastropod molecular phylogenies



A 'x' means that the taxon is represented in the phylogeny; if several samples of this taxon were included, it also means that the taxon has been recovered monophyletic. Full rectangles correspond to supported clades (bootstraps > 80 and/or posterior probabilities > 0.95); dashed rectangles correspond to unsupported clades. The five firsts phylogenies were based on few mitochondrial and nuclear genes (typically: cox-1, 16S, 12S, 28S), and for each of them the focus taxon is indicated; the five others were based on mitogenomes; the last one corresponds to the phylogeny on the Figure 1. Superfamilies for which the monophyly is supported in several phylogenies (Conoidea, Tonnoidea, Buccinoidea except Belomitridae) are not detailed at the family level

are the prime interest of our study) remain stable across analyses or within the tree space explored by MrBayes.

It is now generally accepted that over-parameterization should be favoured over under-parameterization (Abadi et al., 2019; Fabreti & Höhna, 2022), and studies with similar datasets (i.e. mitochondrial genomes spanning a large taxonomic diversity) have also faced the same issues of unrealistic branch lengths but reliable topologies (e.g. Song et al., 2016; Uribe & Zardoya, 2017). Our eight NT matrix topologies (Figures 1 and S1) show few very important differences (i.e. apart from already known unstable regions of the tree) but over-partitioned datasets and overly complex models can return unrealistic branch lengths, even when AICc or BIC scores are good. Caution must be used before running any kind of meta-phylogeny analysis (e.g. dating or diversification analyses) that will use branch length information and we recommend carefully checking the branch lengths to eventually detect unrealistic branch lengths.

4.2 | Input for neogastropod classification

With 42 families of neogastropods and tonnoideans, representing almost 60% of the families currently considered as valid in WoRMS, our dataset is the most complete published so far. Many of the previously recovered relationships are recovered here and new ones are revealed (Table 2).

At the family level, all the families represented by at least two species are recovered monophyletic, except two. The first exception is the Tudiclidae (Buccinoidea), with *Lirabuccinum musculus* not clustering with the two others (*Aeneator recens* and *Buccinulum robustum*). Kantor et al. (2021) already pointed at the radula of *Lirabuccinum musculus* being different from the other Tudiclidae species, suggesting that it might not be a Tudiclidae. The second exception is the Turridae (Conoidea), with *Lucerapex* sp. and *Gemmuloborsonia moosai* not clustering with the others in most trees. This non-monophyly (supported only in the MrBayes AA tree) was already found in Uribe et al. (2018), also based on mitogenomes, but has been contradicted by the exon-based phylogeny of Abdelkrim et al. (2018), with the Turridae being monophyletic and highly supported.

At the superfamily level, some superfamilies are recovered monophyletic and supported, often confirming previously published phylogenies, although based on a much more reduced sampling (Choi et al., 2021; Osca et al., 2015; Uribe et al., 2021; Wang et al., 2021). For example, the Conoidea, represented by many mitogenomes covering almost all the family-level diversity of the group, is found monophyletic. However, some unsupported relationships between families are probably resulting from long-branch attraction (LBA), as suggested in Uribe et al. (2018), and are contradicted both by exon-based phylogenies and anatomical characters (Abdelkrim et al., 2018). For example, the Cochlespiridae, found here closely related to the Marshallenidae in several trees, is more probably sister to all the other Conoidea. The Olivoidea, here for the first time in a mitogenome phylogeny represented by several species, are found monophyletic, confirming the results obtained previously (Kantor et al., 2017). Similarly, the Buccinoidea, recently revised in Kantor et al. (2021), are found monophyletic. However, the Belomitridae, not

included here, might not be closely related to the other Buccinoidea, as suggested by Abdelkrim et al. (2018) and Fedosov et al. (2019). Furthermore, the relationships between the Buccinoidea families are not supported, and sometimes contradict previously published results: for example, the Columbellidae are sister to the Colubrariidae in the MrBayes AA tree with high support, whereas they were found embedded within the Nassariidae in Kantor et al. (2021). The Mitroidea, represented here by a single family (Mitridae), also includes the Pyramimitridae and Charitodoronidae, and these three families were constituting a clade in previously published phylogenies (Fedosov et al., 2018). The Turbinelloidea, represented here by two families (Costellariidae and Turbinellidae), are not monophyletic. The other Turbinelloidea families are sometimes represented by one or a few samples in previously published phylogenies, but their close relationship is never supported, and this superfamily clearly constitutes a potentially non-monophyletic taxon. The Volutoidea, represented here by the Volutidae and Cancellariidae, are not monophyletic. The family Volutidae is sister to all the other neogastropods (except Cancellariidae), and was found to form a clade with the marginellids (Cystiscidae, Granulinidae, Marginellidae and Marginellonidae) in Fedosov et al. (2019). However, as illustrated in our tree, the Cancellariidae never cluster with the rest of the Volutoidea (except in Wang et al. (2021), although not highly supported), and is either found to be sister to all the other Neogastropoda + Tonnoidea/Ficoidea (Cunha et al., 2009; Osca et al., 2015) or to Tonnoidea (Choi et al., 2021; Fedosov et al., 2019). Here, it is generally sister to *Calyptraea chinensis* (Calyptraeidae), a taxon not included in previously published neogastropod phylogenies. Morpho-anatomical data do not support this relationship (Simone, 2002), and the long branch leading to *Calyptraea chinensis* would suggest the occurrence of a phenomenon of LBA. It is important to note that the Cancellariidae have been placed within the Volutoidea only recently (Bouchet et al., 2017), while previously it belonged to its own superfamily Cancellarioidea. Thus, the position of the Cancellariidae, as sister to Calyptraeidae or to the Neogastropoda/Tonnoidea/Ficoidea clade remains to be determined. Finally, the Tonnoidea are found monophyletic and sister to the Ficoidea, a result already supported in Strong et al. (2019) and Wang et al. (2021).

Several supported relationships between the superfamilies (and with the Babyloniidae) are here recovered. The Conoidea and Mitroidea (represented only by the Mitridae) form a clade, as suggested also in Abdelkrim et al. (2018). While the relationships between the clade Conoidea + Mitroidea with Buccinoidea, Turbinellidae, Olivoidea and Babyloniidae are not resolved, they all form a well-supported clade. The deeper relationships are also always supported, with the respective position of the Muricoidea, Costellariidae, Volutidae and Tonnoidea + Ficidae always branching in the same order in our trees, with high support. Nevertheless, a number of phylogenetically important neogastropod taxa remain unrepresented in our dataset (Volutomitridae, Harpidae), or underrepresented (Cancellariidae, Turbinelloidea, Olivoidea). Inclusion of these currently missing lineages may lead to the modification of the tree topology, and therefore, we consider the current tree still too preliminary to induce revisions in systematics or to redefine apomorphies of the currently revealed clades. Further increase in the taxonomic sampling to eventually include all crucial neogastropod lineages will be vital to build a solid basis for the order's reclassification.

4.3 | Future priorities

In future studies, priority should be given to completing the taxon sampling. Among the missing families, some were previously confidently placed in superfamilies in published molecular phylogenies, and do not constitute priority targets: Laubierinidae, Ranellidae, Thalassocyonidae and Tonnidae in Tonnoidea (Strong et al., 2019), Buccinanopsidae, Busyconidae, Chauvetiidae, Eosiphonidae, Prodotiidae and Retimohniidae in Buccinoidea (Kantor et al., 2021), Bouchetispiridae and Conorbidae in Conoidea (Abdelkrim et al., 2018; Kantor et al., 2012), Charitodoronidae and Pyramimitridae in Mitroidea (Fedosov et al., 2015) and Bellolividae, Benthobiidae and Pseudolividae in Olivoidea (Kantor et al., 2017). However, the addition of these families in a phylogeny may improve the overall quality of the tree, either by strengthening the support for the monophyly of the corresponding superfamilies or by clarifying the relationships between the superfamilies. However, priority should be given to those families that have never, or rarely, been included in molecular phylogenies, and whose superfamily membership remain dubious. Within the Buccinoidea, the Belomitridae and Dolicholatiridae have been shown to be sister to the rest of the Buccinoidea in Kantor et al. (2021), but the Belomitridae were not recovered as sister to the other Buccinoidea in Abdelkrim et al. (2018). The four families of marginellids (Cystiscidae, Granulinidae, Marginellidae and Marginellonidae) have been shown to be monophyletic and most probably sister to the Volutidae (thus forming the Volutoidea, excepted Cancellariidae) in Fedosov et al. (2019); however, this remains to be confirmed in a large-scale phylogeny of neogastropods. The likely non-monophyletic Turbinelloidea probably constitutes the main gap in the neogastropod sampling, since several

published molecular phylogenies (including the present one) do not support its monophyly, although three out of five families (Columbariidae, Ptychatractidae, Volutomitridae) have almost never been included in a molecular phylogeny targeting the Neogastropoda. The only exception is Fedosov et al. (2015), in which these five lineages were forming a clade, although not supported. Finally, two unassigned neogastropod families, Harpidae and Strepsiduridae, remain to be included in a molecular phylogeny, and also constitute priorities for future studies.

Furthermore, the monophyly of most, if not all, neogastropod families remain to be tested by including more representatives in each of them, and it is not impossible that some of them would not cluster with the type-genera of the corresponding family. On top of that, and this is true for the whole neogastropods, sequencing more genera, and even more species within each genus, may reveal family-level taxa that remained undetected so far, that is happening regularly when (super)family-level classifications are revised with molecular data (e.g. Buccinoidea – (Kantor et al., 2021); Conoidea – (Abdelkrim et al., 2018)). Thus, as a general rule, sampling effort should first focus on type taxa, in order to ascertain the link with available family-level names, and second on genus or species-level lineages whose family membership is dubious, either because of divergent DNA sequences (typically: *cox-1*) or peculiar morpho-anatomical characters. It should be noted that some of these taxa are rare and difficult to collect, especially if the type taxa are targeted. Given that sampling in the field becomes more and more difficult, because of the legislative barriers reinforced in many countries, museum collections certainly represent the most promising source of material to complete the sampling. However, sequencing DNA from not freshly collected material is more challenging, and requires adequate sequencing strategies (Raxworthy & Smith, 2021).

Indeed, the second priority to improve the phylogeny of neogastropods is to develop next-generation sequencing (NGS)-based methods to recover genetic information of sufficient quality and quantity to resolve the deeper nodes of the phylogeny. As discussed before, sequencing mitogenomes constitutes a good compromise in terms of time and money, but it is prone to artefacts (LBA) and might not be resolutive enough for the deeper nodes of the neogastropod phylogeny. Reduced-genome (UCE, Exon-capture) or transcriptome-based phylogenies have recently shown a high potential in resolving deep nodes in molluscan phylogenies, and are of course promising for neogastropod. If transcriptomic data require fresh samples, exon-capture approaches, by targeting short exons spread over the whole genome, can cope with degraded DNA (e.g. Abdelkrim et al., 2018; Moles & Giribet, 2021). It thus constitutes, in our opinion, the preferential strategy that must be applied to resolve the neogastropod phylogeny.

ACKNOWLEDGEMENTS

The material used to produce the new transcriptomes was collected during the KANACONO expedition in New Caledonia (convention MNHN-Province Sud, APA_NCPS_2016_012; PI N. Puillandre and S. Samadi), with support from the Laboratoire d'Excellence Diversités Biologiques et Culturelles (LabEx BCDiv, ANR-10-LABX-0003-BCDiv) and the project CONOTAX, funded by the French Agence Nationale de la Recherche, France (ANR-13-JSV7-0013-01). This expedition operated under the regulations then in force in the countries in question and satisfy the conditions set by the Nagoya Protocol for access to genetic resources. Additional samples were collected during the expedition to Central Vietnam, supported by the Coastal branch of Russian-Vietnamese Tropical Research and Technology Center (Nha-Trang). The authors are thankful to the staff of the Tropical Center for assistance in the field sampling and for lending laboratory equipment. We thank Laetitia Aznar-Cormano for her help with the RNA extraction and Juliette Gorson and Mandë Holford for their help in the RNA sequencing. The analyses were performed on the Plateforme de Calcul Intensif et Algorithmique PCIA (UAR2700 2AD, MNHN). The present study was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 865101) to N.P.

ORCID

Thomas Lemarcis https://orcid.org/0000-0002-7099-1984
Alexander E. Fedosov https://orcid.org/0000-0002-8035-1403
Yuri I. Kantor https://orcid.org/0000-0002-3209-4940
Jawad Abdelkrim https://orcid.org/0000-0001-7996-8660
Paul Zaharias https://orcid.org/0000-0003-3550-2636
Nicolas Puillandre https://orcid.org/0000-0002-9797-0892

REFERENCES

Abadi, S., Azouri, D., Pupko, T., & Mayrose, I. (2019). Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Communications*, *10*(1), 934. https://doi.org/10.1038/s41467-019-08822-w

Abalde, S., Tenorio, M. J., Afonso, C. M. L., & Zardoya, R. (2017). Mitogenomic phylogeny of cone snails endemic to Senegal. *Molecular Phylogenetics and Evolution*, *112*, 79–87. https://doi.org/10.1016/j.ympev.2017.04.020

Abdelkrim, J., Aznar-Cormano, L., Fedosov, A., Kantor, Y., Lozouet, P., Phuong, M., Zaharias, P., & Puillandre, N. (2018). Exon-capture based phylogeny and diversification of the venomous gastropods (Neogastropoda, Conoidea). *Molecular Biology and Evolution*, *35*, 2355–2374.

Barco, A., Claremont, M., Reid, D. G., Houart, R., Bouchet, P., Williams, S. T., Cruaud, C., Couloux, A., & Oliverio, M. (2010). A molecular phylogenetic framework for the Muricidae, a diverse family of carnivorous gastropods. *Molecular Phylogenetics and Evolution*, *56*(3), 1025–1039.

Barghi, N., Concepcion, G. P., Olivera, B. M., & Lluisma, A. O. (2016). Characterization of the complete mitochondrial genome of *Conus tribblei* Walls, 1977. *Mitochondrial DNA Part A*, *27*(6), 4451–4452. https://doi.org/10.3109/19401736.2015.1089566

Borstein, S. R., & O'Meara, B. C. (2018). *AnnotationBustR*: An R package to extract subsequences from GenBank annotations. *PeerJ*, *6*, e5179. https://doi.org/10.7717/peerj.5179

Bose, U., Wang, T., Zhao, M., Motti, C., Hall, M., & Cummins, S. (2017). Multiomics analysis of the giant triton snail salivary gland, a crown-of-thorns starfish predator. *Scientific Reports*, *7*(1), 6000.

Bouchet, P., Kantor, Y., Sysoev, A., & Puillandre, N. (2011). A new operational classification of the Conoidea (Gastropoda). *Journal of Molluscan Studies*, *77*, 273–308.

Bouchet, P., Rocroi, J.-P., Hausdorf, B., Kaim, A., Kano, Y., Nützel, A., Parkhaev, P., Schroedl, M., & Strong, E. E. (2017). Revised classification, nomenclator and typification of gastropod and Monoplacophoran Families. *Malacologia*, *61*(1–2), 1–526. https://doi.org/10.4002/040.061.0201

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Chernomor, O., von Haeseler, A., & Minh, B. Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology*, *65*(6), 997–1008. https://doi.org/10.1093/sysbio/syw037

Choi, E. H., Choi, N. R., & Hwang, U. W. (2021). The mitochondrial genome of an Endangered freshwater snail *Koreoleptoxis nodifila* (Caenogastropoda: Semisulcospiridae) from South Korea. *Mitochondrial DNA Part B*, *6*(3), 1120–1123. https://doi.org/10.1080/23802359.2021.1901626

Colgan, D. J., Ponder, W. F., Beacham, E., & Macaranas, J. (2007). Molecular phylogenetics of Caenogastropoda (Gastropoda: Mollusca). *Molecular Phylogenetics and Evolution*, *42*, 717–737.

Couto, D. R., Bouchet, P., Kantor, Y. I., Simone, L. R. L., & Giribet, G. (2016). A multilocus molecular phylogeny of Fasciolariidae (Neogastropoda: Buccinoidea). *Molecular Phylogenetics and Evolution*, *99*, 309–322. https://doi.org/10.1016/j.ympev.2016.03.025

Crotty, S. M., & Holland, B. R. (2022). Comparing partitioned models to mixture models: Do information criteria apply? *Systematic Biology*, syac003. https://doi.org/10.1093/sysbio/syac003

Crotty, S. M., Minh, B. Q., Bean, N. G., Holland, B. R., Tuke, J., Jermiin, L. S., & Haeseler, A. V. (2019). GHOST: Recovering historical signal from heterotachously evolved sequence alignments. *Systematic Biology*, *69*, 249–264. https://doi.org/10.1093/sysbio/syz051

Cunha, R. L., Grande, C., & Zardoya, R. (2009). Neogastropod phylogenetic relationships based on entire mitochondrial genomes. *BMC Evolutionary Biology*, *9*, 210.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., SA, M. C., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Fabreti, L. G., & Höhna, S. (2022). Bayesian inference of phylogeny is robust to substitution model over-parameterization. *Evolutionary Biology*. https://doi.org/10.1101/2022.02.17.480861

Fedosov, A., Puillandre, N., Herrmann, M., Kantor, Y., Oliverio, M., Dgebuadze, P., Modica, M. V., & Bouchet, P. (2018). The collapse of Mitra: Molecular systematics and morphology of the Mitridae (Gastropoda: Neogastropoda). *Zoological Journal of the Linnean Society*, *183*(2), 253–337.

Fedosov, A., Puillandre, N., Kantor, Y., & Bouchet, P. (2015). Phylogeny and systematics of mitriform gastropods (Mollusca: Gastropoda: Neogastropoda). *Zoological Journal of the Linnean Society*, *175*(2), 336–359.

Fedosov, A. E., Caballer Gutierrez, M., Buge, B., Sorokin, P. V., Puillandre, N., & Bouchet, P. (2019). Mapping the missing branch on the neogastropod tree of life: Molecular phylogeny of marginelliform gastropods. *Journal of Molluscan Studies*, *85*, 440–452.

Fedosov, A. E., Puillandre, N., Herrmann, M., Dgebuadze, P., & Bouchet, P. (2017). Phylogeny, systematics, and evolution of the family Costellariidae (Gastropoda: Neogastropoda). *Zoological Journal of the Linnean Society*, *179*(3), 541–626.

Fourdrilis, S., de Frias Martins, A. M., & Backeljau, T. (2018). Relation between mitochondrial DNA hyperdiversity, mutation rate and mitochondrial genome evolution in Melarhaphe neritoides (Gastropoda: Littorinidae) and other Caenogastropoda. *Scientific Reports*, *8*(1), 17964. https://doi.org/10.1038/s41598-018-36428-7

Galindo, L. A., Puillandre, N., Utge, J., Lozouet, P., & Bouchet, P. (2016). The phylogeny and systematics of the Nassariidae revisited (Gastropoda, Buccinoidea). *Molecular Phylogenetics and Evolution*, *99*, 337–353.

Harasewych, M. G., Sei, M., Wirshing, H. H., & Uribe, J. E. (2019). The complete mitochondrial genome of Neptuneopsis gilchristi G.B. Sowerby III, 1898 (Neogastropoda: Volutidae: Calliotectinae). *The Nautilus*, *133*, 7.

Huang, H., Yang, H., Li, J., Guo, B., & Ye, Y. (2021). *Complete mitochondrial genomes of Babylonia formosae and Babylonia zeylanica (Neogastropoda: Babyloniidae) and increased sampling give new insights into neogastropoda phylogeny*. In Review. https://doi.org/10.21203/rs.3.rs-948577/v1

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589.

Kantor, Y., Lozouet, P., Puillandre, N., & Bouchet, P. (2014). Lost and found: The Eocene family Pyramimitridae (Neogastropoda) discovered in the Recent fauna of the Indo-Pacific. *Zootaxa*, *3754*(3), 239–276. https://doi.org/10.11646/zootaxa.3754.3.2

Kantor, Y. I., Fedosov, A., Puillandre, N., Bonillo, C., & Bouchet, P. (2017). Returning to the roots: Morphology, molecular phylogeny and classification of the Olivoidea (Gastropoda: Neogastropoda). *Zoological Journal of the Linnean Society*, *180*(3), 493–541.

Kantor, Y. I., Strong, E. E., & Puillandre, N. (2012). A new lineage of Conoidea (Gastropoda: Neogastropoda) revealed by morphological and molecular data. *Journal of Molluscan Studies*, *78*, 246–255.

Kantor, Y. I., Fedosov, A. E., Kosyan, A. R., Puillandre, N., Sorokin, P. A., Kano, Y., Clark, R. N., & Bouchet, P. (2021). Molecular phylogeny and revised classification of the Buccinoidea (Neogastropoda). *Zoological Journal of the Linnean Society*, *194*, 789–857. https://doi.org/10.1093/zoolinnean/zlab031

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Katoh, K., & Frith, M. C. (2012). Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics*, *28*(23), 3144–3146. https://doi.org/10.1093/bioinformatics/bts578

Kumar, S., Stecher, G., & Tamura, K. (2016). mega7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, *33*(7), 1870–1874. https://doi.org/10.1093/molbev/msw054

Kuznetsova, K. G., Zvonareva, S. S., Ziganshin, R., Mekhova, E. S., Dgebuadze, P., Yen, D. T. H., Nguyen, T. H. T., Moshkovskii, S. A., & Fedosov, A. E. (2022). *Vexitoxins: A novel class of conotoxin-like venom peptides from predatory gastropods of the genus* Vexillum [Preprint]. *Biochemistry*. https://doi.org/10.1101/2022.01.15.476460

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Le, S. Q., Dang, C. C., & Gascuel, O. (2012). Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Molecular Biology and Evolution*, *29*(10), 2921–2936. https://doi.org/10.1093/molbev/mss112

Lopez, P., Casane, D., & Philippe, H. (2002). Heterotachy, an important process of protein evolution. *Molecular Biology and Evolution*, *19*(1), 1–7. https://doi.org/10.1093/oxfordjournals.molbev.a003973

Machkour-M'Rabet, S., Hanes, M. M., Martínez-Noguez, J. J., Cruz-Medina, J., & García-De León, F. J. (2021). The queen conch mitogenome: Intra- and interspecific mitogenomic variability in Strombidae and phylogenetic considerations within the Hypsogastropoda. *Scientific Reports*, *11*(1), 11972. https://doi.org/10.1038/s41598-021-91224-0

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Modica, M. V., Bouchet, P., Cruaud, C., Utge, J., & Oliverio, M. (2011). Molecular phylogeny of the nutmeg shells (Neogastropoda, Cancellariidae). *Molecular Phylogenetics and Evolution*, *59*, 685–697.

Modica, M. V., Sánchez, J. R., Pasquadibisceglie, A., Oliverio, M., Mariottini, P., & Cervelli, M. (2018). Anti-haemostatic compounds from the vampire snail Cumia reticulata: Molecular cloning and in-silico structure-function analysis. *Computational Biology and Chemistry.*, *75*, 168–177.

Moles, J., & Giribet, G. (2021). A polyvalent and universal tool for genomic studies in gastropod molluscs (Heterobranchia). *Molecular Phylogenetics and Evolution*, *155*, 106996. https://doi.org/10.1016/j.ympev.2020.106996

Olivera, B. M., Fedosov, A., Imperial, J. S., & Kantor, Y. (2017). Physiology of envenomation by conoidean gastropods. In *Physiology of Molluscs: A Collection of Selected Reviews* (pp. 153–188). Apple Academic Press.

Oliverio, M., & Modica, M. V. (2010). Relationships of the haematophagous marine snail Colubraria (Rachiglossa: Colubrariidae), within the neogastropod phylogenetic framework. *Zoological Journal of the Linnean Society*, *158*, 779–800.

Osca, D., Templado, J., & Zardoya, R. (2015). Caenogastropod mitogenomics. *Molecular Phylogenetics and Evolution*, *93*, 118–128.

Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R., Moret, B. M., & Stamatakis, A. (2010). How many bootstrap replicates are necessary? *Journal of Computational Biology*, *17*(3), 337–354.

Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, *53*, 793–808.

Puillandre, N., Kantor, Y. I., Sysoev, A. V., Couloux, A., Meyer, C. P., Rawlings, T. A., Todd, J. A., & Bouchet, P. (2011). The dragon tamed? A molecular phylogeny of the Conoidea (Gastropoda). *Journal of Molluscan Studies*, *77*(3), 259–272. https://doi.org/10.1093/mollus/eyr015

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, *67*(5), 901–904.

Ranwez, V., Harispe, S., Delsuc, F., & Douzery, E. J. P. (2011). MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *PLoS One*, *6*(9), e22594. https://doi.org/10.1371/journal.pone.0022594

Ravitchandirane, V., Sukumar, V., & Thangaraj, M. (2013). Phylogenetic status of babylonia Zeylanica (family Babyloniidae) based on 18S rRNA gene Fragment. *Annales of West University of Timisoara. Series of Biology*, *16*(2), 135.

Rawlings, T. A., MacInnis, M. J., Bieler, R., Boore, J. L., & Collins, T. M. (2010). Sessile snails, dynamic genomes: Gene rearrangements within the mitochondrial genome of a family of caenogastropod molluscs. *BMC Genomics*, *11*(1), 440. https://doi.org/10.1186/1471-2164-11-440

Raxworthy, C. J., & Smith, B. T. (2021). Mining museums for historical DNA: Advances and challenges in museomics. *Trends in Ecology & Evolution*, *36*(11), 1049–1060. https://doi.org/10.1016/j.tree.2021.07.009

Rombel, I. T., Sykes, K. F., Rayner, S., & Johnston, S. A. (2002). ORF-FINDER: A vector for high-throughput gene identification. *Gene*, *282*(1–2), 33–41. https://doi.org/10.1016/S0378-1119(01)00819-8

Ronquist, F., Teslenko, M., Van Den mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, *61*(3), 539–542.

Schrödl, M., & Stöger, I. (2014). A review on deep molluscan phylogeny: Old markers, integrative approaches, persistent problems. *Journal of Natural History*, *48*(45–48), 2773–2804.

Simone, L. R. L. (2002). Comparative morphological study and phylogeny of representatives of the superfamily Calyptraeoidea (including Hipponicoidea) (Mollusca, Caenogastropoda). *Biota Neotropica*, *2*(2), 1–137. https://doi.org/10.1590/S1676-06032002000200013

Song, F., Li, H., Jiang, P., Zhou, X., Liu, J., Sun, C., Vogler, A. P., & Cai, W. (2016). Capturing the phylogeny of Holometabola with mitochondrial genome data and Bayesian site-heterogeneous mixture models. *Genome Biology and Evolution*, *8*(5), 1411–1426. https://doi.org/10.1093/gbe/evw086

Strong, E. E., Puillandre, N., Beu, A. G., Castelin, M., & Bouchet, P. (2019). Frogs and tuns and tritons–A molecular phylogeny and revised family classification of the predatory gastropod superfamily Tonnoidea (Caenogastropoda). *Molecular Phylogenetics and Evolution*, *130*, 18–34.

Uribe, J. E., Fedosov, A. E., Murphy, K. R., Sei, M., & Harasewych, M. G. (2021). The complete mitochondrial genome of *Costapex baldwinae* (Gastropoda: Neogastropoda: Turbinelloidea: Costellariidae) from the Caribbean Deep-Sea. *Mitochondrial DNA Part B*, *6*(3), 943–945. https://doi.org/10.1080/23802359.2021.1889408

Uribe, J. E., Irisarri, I., Templado, J., & Zardoya, R. (2019). New patellogastropod mitogenomes help counteracting long-branch attraction in the deep phylogeny of gastropod mollusks. *Molecular Phylogenetics and Evolution*, *133*, 12–23.

Uribe, J. E., & Zardoya, R. (2017). Revisiting the phylogeny of Cephalopoda using complete mitochondrial genomes. *Journal of Molluscan Studies*, *83*(2), 133–144. https://doi.org/10.1093/mollus/eyw052

Uribe, J. E., Zardoya, R., & Puillandre, N. (2018). Phylogenetic relationships of the conoidean snails (Gastropoda: Caenogastropoda) based on mitochondrial genomes. *Molecular Phylogenetics and Evolution*, *127*, 898–906.

Wang, J.-G., Zhang, D., Jakovlić, I., & Wang, W.-M. (2017). Sequencing of the complete mitochondrial genomes of eight freshwater snail species exposes pervasive paraphyly within the Viviparidae family (Caenogastropoda). *PLoS One*, *12*(7), e0181699. https://doi.org/10.1371/journal.pone.0181699

Wang, Q., Liu, H., Yue, C., Xie, X., Li, D., Liang, M., & Li, Q. (2021). Characterization of the complete mitochondrial genome of *Ficus variegata* (Littorinimorpha: Ficidae) and molecular phylogeny of Caenogastropoda. *Mitochondrial DNA Part B*, *6*(3), 1126–1128. https://doi.org/10.1080/23802359.2021.1901628

Williams, S. T., Foster, P. G., & Littlewood, D. T. J. (2014). The complete mitochondrial genome of a turbinid vetigastropod from MiSeq Illumina sequencing of genomic DNA and steps towards a resolved gastropod phylogeny. *Gene*, *533*, 38–47.

WoRMS Editorial Board. (2022). *World Register of Marine Species*. Available from www.marinespecies.org. Accessed 14 March, 2022.

Yang, M., Dong, D., & Li, X. (2021). The complete mitogenome of Phymorhynchus sp. (Neogastropoda, Conoidea, Raphitomidae) provides insights into the deep-sea adaptive evolution of Conoidea. *Ecology and Evolution*, *11*, 7518–7531. https://doi.org/10.1002/ece3.7582

Zaharias, P., Pante, E., Gey, D., Fedosov, A. E., & Puillandre, N. (2020). Data, time and money: Evaluating the best compromise for inferring molecular phylogenies of non-model animal taxa. *Molecular Phylogenetics and Evolution*, *142*, 106660.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Lemarcis, T., Fedosov, A. E., Kantor, Y. I., Abdelkrim, J., Zaharias, P., & Puillandre, N. (2022). Neogastropod (Mollusca, Gastropoda) phylogeny: A step forward with mitogenomes. *Zoologica Scripta*, *51*, 550–561. https://doi.org/10.1111/zsc.12552

La phylogénie moléculaire que nous avons produite avec les mitogénomes est la plus exhaustive à ce jour à l'échelle des néogastéropodes (hormis l'article (A. E. Fedosov et al., in press), qui vient d'être accepté quelques jours avant de terminer mon manuscrit de thèse – Annexe 1). Cette phylogénie met en avant de nouvelles hypothèses que nous pourrons tester avec les données obtenues avec la capture d'exons, telles que la non monophylie des Turbinelloidea. L'ajout de nombreux marqueurs à l'aide de la capture d'exons mais aussi la mise en place d'un échantillonnage le plus diversifié et complet possible nous permettront d'améliorer la résolution de la phylogénie à l'échelle des néogastéropodes. En effet, l'utilisation d'un grand nombre de marqueurs devrait nous permettre d'ajouter suffisamment de données pour enrichir la phylogénie à ces profondeurs phylogénétiques.

4. LES METHODES DE SEQUENÇAGE DE GENOMES REDUITS

Les approches transcriptomiques peuvent permettre d'obtenir des informations phylogénétiques sur les relations phylogénétiques profondes. Cela a déjà été utilisé chez les mollusques (Kocot et al., 2011; Smith et al., 2011; Zapata et al., 2014). Cependant, à l'échelle des néogastéropodes, le nombre de taxa à séquencer pour les inclure dans la phylogénie rendrait le coût de ce séquençage exorbitant (Zaharias et al., 2020). De plus, les méthodes que j'ai précédemment citées ne permettent pas d'obtenir un compromis suffisamment intéressant entre le nombre de marqueurs séquencés sur un grand nombre de taxa. C'est ce qui nous a conduit à envisager d'utiliser des approches de génomes réduits (Figure 4). Parmi elles, le « genome skimming » est une approche de séquençage qui consiste à récupérer les parties génomiques en grand nombre de copies. Cela comprend l'ADN ribosomal, le génome mitochondrial et les éléments nucléaires répétés comme les microsatellites ou les éléments transposables (Trevisan et al., 2019). C'est pourtant une méthode qui est encore peu utilisée chez les mollusques, et le nombre de marqueurs obtenus est finalement assez réduit, notamment pour des applications de reconstruction phylogénétique profonde. Pour ces raisons, nous avons opté pour une approche qui utilise des génomes réduits mais enrichis, afin de produire une phylogénie bien plus robuste, à un coût raisonnable (Zaharias et al., 2020). Dans cette catégorie de méthodes, le séquençage des UCEs (Ultra Conserved Elements en anglais) permet de séquencer plusieurs milliers de marqueurs à la fois. C'est également le cas de la méthode de capture d'exons, et je vais décrire

ces deux méthodes dans les paragraphes suivants et expliquer les raisons qui nous ont poussé à choisir une méthode plutôt que l'autre.

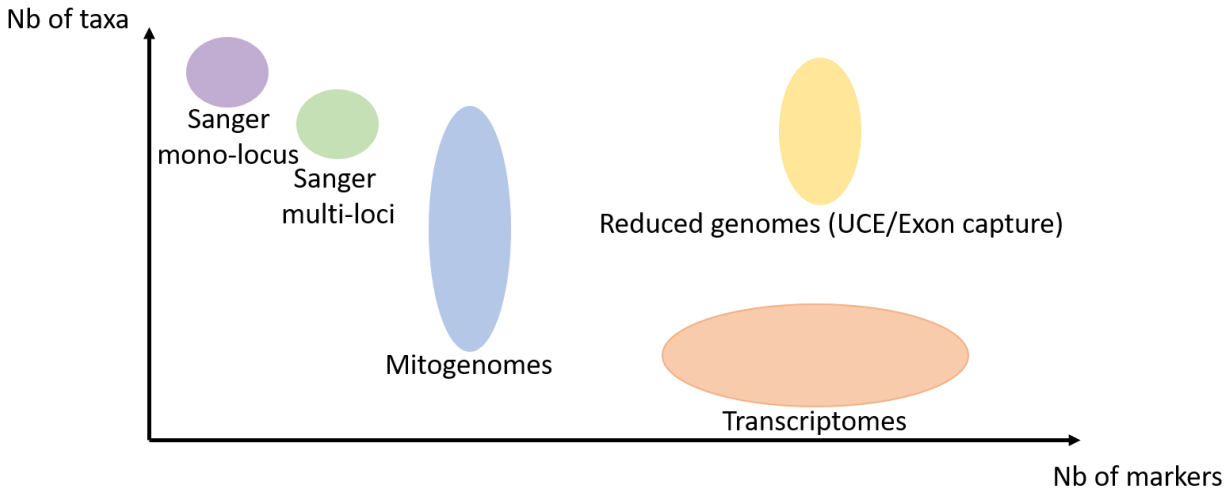


Figure 4 : Représentation graphique du nombre de marqueurs séquencés en fonction du nombre de taxa. *Les différentes méthodes de séquençage pour reconstruire des phylogénies sont représentées : En violet le séquençage Sanger à partir d'un gène ; en vert le séquençage Sanger à partir de plusieurs gènes concaténés ; en bleu le séquençage de mitogénomes complets ; en orange le séquençage de transcriptomes et en jaune le séquençage de génomes réduits.*

Les UCEs sont des éléments très conservés au sein de différents groupes taxonomiques (Ryu et al., 2012). Ils jouent un rôle dans l'amplification des activités protéiques ou les processus de développement, mais aussi des régulations épigénétiques ainsi que des processus immunitaires. Ces éléments se retrouvent chez tous les eucaryotes et peuvent être regroupés entre eux par similarité de séquences. De plus, la composition des séquences adjacentes est différente selon la lignée considérée, ce qui peut apporter de l'information phylogénétique pertinente.

La méthode de séquençage des UCEs consiste en l'utilisation « d'ancres » qui vont venir s'accrocher sur des zones génomiques très conservées afin de séquencer les zones adjacentes, plus variables, de ces portions de génome très conservées. Comme cette méthode permet de cibler des zones très conservées, il devrait être possible, en théorie, de récupérer les zones partagées par des groupes ayant divergé il y a potentiellement plusieurs centaines de millions d'années. L'avantage de ces zones très conservées est qu'elles peuvent être capturées avec les

mêmes sondes pour un taxon très large, ce qui est moins le cas avec des sondes de capture d'exons.

La méthode des UCEs a été pour la première fois publiée en 2012 par Brant C. Faircloth (Faircloth et al., 2012). Cette preuve de concept s'est basée sur trois modèles biologiques différents : la poule (*Gallus gallus),* l'anole vert (*Anolis carolinensis*) et le diamant mandarin (*Taeniopygia guttata*). Le pipeline de dessin des sondes d'ancrages est composé de quatre grandes étapes : (i) Identification des UCEs par méthodes de BLAST (Camacho et al., 2009) ; (ii) Dessin des sondes d'ancrage ; (iii) Alignement de ces sondes contre des génomes afin de conserver les sondes avec un pourcentage d'identité supérieur à 90% ; (iv) Test *in silico* des sondes sur des génomes de primates.

Suite à cet article, de nombreuses équipes ont utilisé cette méthode pour séquencer les génomes réduits dans des groupes différents tels que les oiseaux, les arthropodes, les mammifères et les mollusques (Blaimer et al., 2015, 2016; Bossert et al., 2019; Branstetter et al., 2017; Derkarabetian et al., 2019; Faircloth et al., 2015; McCormack et al., 2016; Moles & Giribet, 2021; Starrett et al., 2017). L'utilisation des UCEs a permis de produire des phylogénies à des échelles taxonomiques plus ou moins profondes au sein de ces différents groupes.

La méthode de capture d'exons utilise des sondes pour récupérer et séquencer les zones d'intérêt. Pour dessiner les sondes de capture, un ou plusieurs génomes et des transcriptomes sont utilisés pour mettre en évidence les zones exoniques qui seront le plus partagées au sein du groupe que l'on veut séquencer. Ce sont les exons qui sont ciblés par les sondes de capture qui seront séquencés. Ces exons doivent être à la fois suffisamment conservés pour être capturés dans des groupes qui ont divergé depuis plusieurs dizaines de millions d'années, et suffisamment variables pour apporter de l'information phylogénétique et permettre de placer les différents groupes les uns par rapport aux autres. La quantité de données informatives qui peuvent être séquencées grâce à la capture d'exons est bien plus importante par rapport à celle des mitogénomes. Enfin, contrairement aux UCEs, la méthode de capture d'exons cible des zones dont la variabilité est mieux connue *a priori*, limitant ainsi la sélection des exons à ceux présentant un niveau de variabilité adapté au taxon d'intérêt.

La méthode de capture d'exons a également permis d'augmenter la quantité de données génomiques séquencées pour un échantillon donné. La méthode a pour la première fois été utilisée par Bi et al. (Bi et al., 2012) sur des tamias (*Tamias* sp.). À partir de séquences de transcriptomes, ils ont identifié des exons dont ils se sont servis pour produire des sondes de

capture. Les exons capturés et séquencés sont alors assemblés afin d'identifier les loci orthologues et de reconstruire une phylogénie. De même que pour les UCEs, cette méthode a été utilisée dans de nombreux groupes différents pour reconstruire des phylogénies (Choquet et al., 2019; Hedtke et al., 2013; Hugall et al., 2016; Jiang et al., 2019; Phuong & Mahardika, 2018; Teasdale et al., 2016). Les sondes utilisées pour capturer les exons d'intérêt sont dessinées à partir d'alignements entre des génomes de référence et des transcriptomes. Le protéome annoté du ou des génomes de référence est utilisé pour aligner les transcriptomes par la méthode de BLAST. Cela permet de repérer les séquences exoniques partagées par plusieurs groupes taxonomiques d'un ensemble plus grand. De plus, contrairement à l'approche UCE, les séquences obtenues sont codantes, ce qui facilite grandement l'identification des zones homologues et l'alignement des séquences.

Au sein de l'équipe, plusieurs études basées sur une approche de captures d'exons ont été effectuées à différentes échelles taxonomiques : à l'échelle de la super-famille chez les Conoidea (Abdelkrim et al., 2018), à l'échelle de la famille chez les Turridae (Zaharias et al., in press) et à l'échelle de l'ordre chez les néogastéropodes (A. E. Fedosov et al., in press). Cette dernière étude sur la phylogénie des néogastéropodes produite à partir de données de capture d'exons était déjà en cours lorsque j'ai débuté ma thèse, et l'article correspondant, dont je suis co-auteur, vient d'être accepté pour publication dans Systematic Biology (Annexe 1). Cette phylogénie se base sur un échantillonnage réduit par rapport à la phylogénie produite au cours de cette thèse. De plus, la proportion de données manquantes dans ce jeu de données, comme dans les études de capture d'exons précédentes, est importante. Enfin, elle est hétérogène entre les spécimens, avec pour résultat des arbres localement peu résolus, avec des groupes ayant des longues branches. Je reviendrai sur les raisons de cette quantité de données manquantes dans le chapitre suivant, raisons qui nous ont conduit à recommencer le design des sondes de capture de zéro, comme je l'exposerai aussi dans le chapitre suivant.

L'apport de la méthode de capture d'exons dans l'amélioration des phylogénies au sein des néogastéropodes malgré les portions de l'arbre non encore résolues nous a donc poussé à utiliser également cette méthode de séquençage dans le cadre du projet HYPERDIVERSE (Voir chapitre 2).

La capture d'exons donne la possibilité de séquencer des exons partagés par des taxa très divergents. Cependant, la perte d'efficacité de la capture due à l'éloignement phylogénétique des taxa ciblés n'a pas été testée de façon robuste. C'est pourquoi, nous avons décidé de tester

jusqu'à quelle distance génétique nos sondes de capture sont capables de s'accrocher et permettent de faire un séquençage de nos spécimens. Afin de limiter les biais qui ne seraient pas liés à la distance génétique, nous avons sélectionné des échantillons qui ont tous été collectés lors de la même mission, conservés dans les mêmes conditions et dont l'ADN a été extrait avec la même méthode par la même personne (Dario Zuccon). La méthodologie utilisée pour réaliser cette expérimentation a mis en évidence l'impact de la distance génétique sur la capacité des sondes à capturer les exons d'intérêts. Mais avant de présenter les résultats de cette étude (chapitre 3) et les phylogénies obtenues (chapitre 4), je vais développer dans le chapitre suivant les protocoles mis en place pour obtenir les séquences.

Chapitre 2 - Développements méthodologiques

1. PRE-SEQUENÇAGE : DESIGN DES SONDES DE CAPTURE

La capture d'exons est la méthode que nous avons choisie pour générer le jeu de données le plus exhaustif possible afin de reconstruire notre phylogénie. C'est une méthode qui consiste à venir capturer spécifiquement des séquences exoniques par la synthèse de sondes qui vont cibler ces zones à capturer. L'avantage de cibler les exons est que nous allons pouvoir capturer des séquences provenant de spécimens qui ont divergé depuis plusieurs millions voire centaines de millions d'années. En effet, capturer un même exon dans des lignées distantes est possible si l'on synthétise suffisamment de sondes pour capturer ces exons dont les séquences sont divergentes.

Pour dessiner les sondes de capture qui vont cibler les zones codantes il est préférable d'avoir un (ou plusieurs) génome de référence le plus complet possible. Grâce au génome de référence, nous pouvons identifier les zones codantes du génome et cela nous servira également à délimiter nos exons. Nous utilisons aussi des transcriptomes qui représentent la part du génome qui est codante. Le coût de production des génomes est bien plus important que pour les transcriptomes, nous avons donc à notre disposition une plus grande quantité de transcriptomes que de génomes. Ces transcriptomes sont choisis dans des groupes taxonomiques différents qui vont représenter le plus possible la diversité au sein des néogastéropodes. C'est un aspect à la fois important et difficile : pour que notre jeu de données soit le plus représentatif possible, il est primordial que nous couvrions le plus de groupes taxonomiques différents. Cependant, il est compliqué d'avoir à la fois des transcriptomes représentatifs de l'ensemble de la diversité du groupe, et de qualité homogène entre toutes les lignées. Ainsi, même si les qualités des transcriptomes sont variables en termes de complétion, il est important d'intégrer également les transcriptomes de moins bonne qualité tant que leur intérêt taxonomique est avéré.

Les étapes de recherche des exons, communs à la plupart des groupes de néogastéropodes, ont nécessité un travail d'équipe. Nous nous sommes réunis à plusieurs reprises afin de déterminer le meilleur protocole bioinformatique possible. Nous avons travaillé avec mon directeur de thèse ainsi qu'avec Alessandro Derzelle qui était post-doctorant au sein de l'équipe HYPERDIVERSE. Alessandro a rédigé les scripts nécessaires aux premières étapes du pipeline

que je vais vous détailler ensuite (étapes 1 à 8). J'ai par ailleurs rédigé plusieurs scripts de filtrations pour les étapes 9 à 11 du pipeline (tous les scripts utilisés au cours de ma thèse sont disponibles sur le Github du projet – https://github.com/Hyperdiverseproject/Exon_capture). Pour la construction des séquences ancestrales, Alexander Fedosov, un chercheur associé aux différents projets de recherche de l'équipe depuis plusieurs années, a rédigé les scripts nécessaires à leur obtention (étapes 12 et 18 du pipeline).

1.1. JEU DE DONNEES

La publication d'un génome complet annoté d'un néogastéropode de la famille des Conidae, *Lautoconus ventricosus* (ci-après nommé « *Conus ventricosus*») par Pardos-Blas et ses collaborateurs en 2021 (Pardos-Blas et al., 2021) nous a permis d'avoir un génome de référence d'un néogastéropode pour dessiner nos sondes de capture. Ceci est une nouveauté par rapport aux captures d'exons précédentes faites dans l'équipe (Abdelkrim et al., 2018). Le génome de référence utilisé à l'époque était celui de *Lottia gigantea* (Simakov et al., 2013), une espèce qui a divergé des néogastéropodes depuis plus de 450 millions d'années (Zapata et al., 2014b). Les jeux de données issus de la capture d'exons qui en ont résulté contenaient une grande proportion de données manquantes. Cela peut s'expliquer par la distance génétique importante entre le génome de *Lottia gigantea* et les néogastéropodes. L'utilisation d'un génome de référence éloigné peut avoir aussi comme impact des difficultés à identifier les limites introns/exons. En effet, la modification des positions de certains exons mais aussi l'apparition de zones introniques au sein de certaines zones exoniques chez *Lottia gigantea*, mais pas chez les néogastéropodes, ou l'inverse, peut compliquer l'identification des exons d'intérêt. De plus, un événement de duplication totale de génome mis en évidence chez les néogastéropodes (Farhat et al., 2023) a pu avoir des conséquences sur l'identification des gènes orthologues entre les néogastéropodes et *Lottia gigantea*. Les précédentes captures d'exons réalisées dans l'équipe ont aussi eu pour résultat la présence en grande quantité de paralogues (Abdelkrim et al., 2018) en plus des données manquantes dans le jeu de données final. Tous ces problèmes ont donc été réduits avec l'utilisation d'un génome de référence au sein des néogastéropodes.

En complément du génome, nous avons utilisé des transcriptomes avec les plus hauts scores BUSCO et/ou qui représentent la diversité des néogastéropodes, parmi ceux disponibles dans GenBank ou produit dans l'équipe dans le cadre d'autres projets de recherche. Cela comprend 15 transcriptomes, dont 12 représentatifs des différentes lignées de néogastéropodes + 1

Charoniidae et 1 Personidae (Tonnoidea), et un groupe externe (Ovulidae). Parmi les 12 transcriptomes de néogastéropodes, 2 transcriptomes de Raphitomidae ont été inclus, cette famille faisant l'objet d'un échantillonnage plus complet que les autres groupes de néogastéropodes (voir Introduction et chapitre ci-dessous).

Nous avons décidé de repartir de zéro afin de dessiner nos sondes de capture. Nous voulions réduire au maximum la présence de paralogues potentiellement ciblés par nos sondes de capture mais également augmenter drastiquement la quantité d'exons capturés par nos sondes. Pour ce faire, nous avons décidé de suivre une méthode déjà utilisée précédemment par plusieurs équipes (Phuong & Mahardika, 2018; Teasdale et al., 2016) par une approche d'alignement de séquences afin de repérer les exons partagés par un grand nombre de taxa au sein des néogastéropodes. Cependant, de nombreuses étapes ont été modifiées, d'autres ont été ajoutées, en s'inspirant en partie des autres pipelines publiés.

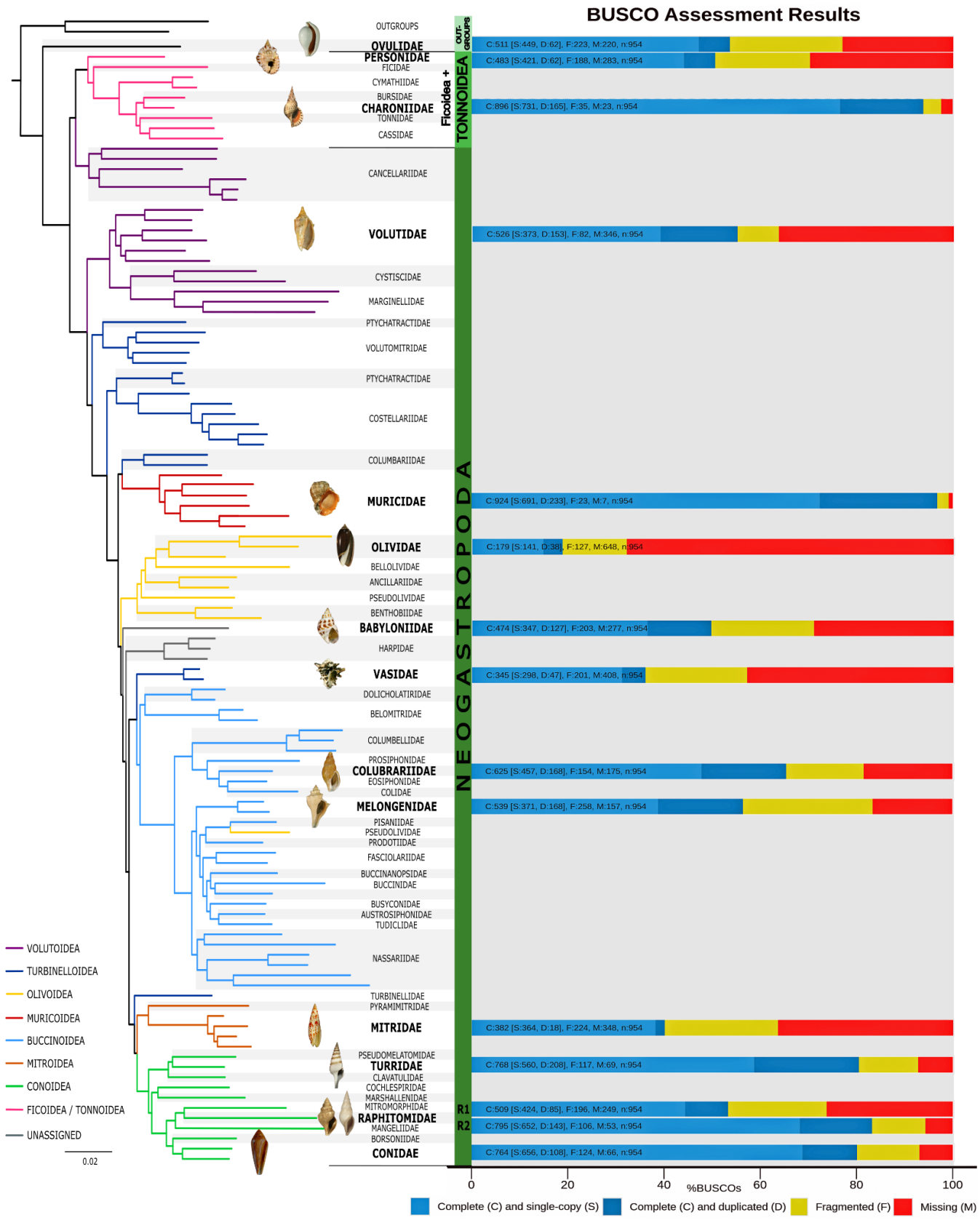

BUSCO Assessment Results
OUTGROUPS
OVULIDAE
PERSONIDAE
FICIDAE
CYMATHIIDAE
BURSIDAE
CHARONIIDAE
TONNIDAE
CASSIDAE
CANCELLARIIDAE
VOLUTIDAE
CYSTISCIDAE
MARGINELLIDAE
PTYCHATRACTIDAE
VOLUTOMITRIDAE
PTYCHATRACTIDAE
COSTELLARIIDAE
COLUMBARIIDAE
MURICIDAE
OLIVIDAE
BELLOLIVIDAE
ANCILLARIIDAE
PSEUDOLIVIDAE
BENTHOBIIDAE
BABYLONIIDAE
HARPIDAE
VASIDAE
DOLICHOLATIRIDAE
BELOMITRIDAE
COLUMBELLIDAE
PROSIPHONIDAE
COLUBRARIIDAE
EOSIPHONIDAE
COLIDAE
MELONGENIDAE
PISANIIDAE
PSEUDOLIVIDAE
PRODOTIIDAE
FASCIOLARIIDAE
BUCCINANOPSIDAE
BUCCINIDAE
BUSYCONIDAE
AUSTROSIPHONIDAE
TUDICLIDAE
NASSARIIDAE
TURBINELLIDAE
PYRAMIMITRIDAE
MITRIDAE
PSEUDOMELATOMIDAE
TURRIDAE
CLAVATULIDAE
COCHLESPIRIDAE
MARSHALLENIDAE
MITROMORPHIDAE
RAPHITOMIDAE
MANGELIIDAE
BORSONIIDAE
CONIDAE
OUT-GROUPS
Ficoidea + TONNOIDEA
NEOGASTROPODA
C:511 [S:449, D:62], F:223, M:220, n:954
C:483 [S:421, D:62], F:188, M:283, n:954
C:896 [S:731, D:165], F:35, M:23, n:954
C:526 [S:373, D:153], F:82, M:346, n:954
C:924 [S:691, D:233], F:23, M:7, n:954
C:179 [S:141, D:38], F:127, M:648, n:954
C:474 [S:347, D:127], F:203, M:277, n:954
C:345 [S:298, D:47], F:201, M:408, n:954
C:625 [S:457, D:168], F:154, M:175, n:954
C:539 [S:371, D:168], F:258, M:157, n:954
C:382 [S:364, D:18], F:224, M:348, n:954
C:768 [S:560, D:208], F:117, M:69, n:954
R1
C:509 [S:424, D:85], F:196, M:249, n:954
R2
C:795 [S:652, D:143], F:106, M:53, n:954
C:764 [S:656, D:108], F:124, M:66, n:954
VOLUTOIDEA
TURBINELLOIDEA
OLIVOIDEA
MURICOIDEA
BUCCINOIDEA
MITROIDEA
CONOIDEA
FICOIDEA / TONNOIDEA
UNASSIGNED
0.02
0
20
40
%BUSCOs
60
80
100
Complete (C) and single-copy (S)
Complete (C) and duplicated (D)
Fragmented (F)
Missing (M)

Figure 5 : Phylogénie des Neogastropoda, issue de Fedosov et al. in press (Annexe 1).
En gras sont représentées les familles pour lesquelles un transcriptome a été utilisé pour réaliser le dessin des exons. Le graphique de la partie droite de la figure représente les scores BUSCO pour chacun des transcriptomes.

1.2. RECHERCHE DES EXONS

1. Les 15 transcriptomes que nous avons utilisés ont été soit téléchargés à partir de GenBank soit récupérés déjà assemblés car obtenus pour d'autres projets de l'équipe. Dans un premier temps, chaque transcriptome est annoté en utilisant TransDecoder v.5.5.0 (Haas, BJ. https://github.com/TransDecoder/TransDecoder) afin d'identifier les séquences codantes potentielles les plus longues et enlever les séquences d'ARNm.
2. Un alignement en utilisant BLASTp est ensuite effectué entre chaque transcriptome annoté et le protéome publié de *Conus*, avec une e-value de 10^{-10}. Seuls les matches uniques, à savoir les séquences codantes de chaque transcriptome qui s'alignent contre une protéine unique de *Conus*, sont conservées (cas A & C de la figure 6, colonne E du tableau 2), avec pour objectif de réduire la probabilité de cibler des paralogues (cas B de la figure 6). Le nombre de hits uniques est corrélé avec le score BUSCO (qui peut être considéré comme une mesure de la complétude d'un transcriptome, et donc de sa qualité) : en effet, plus le score BUSCO (Simão et al., 2015) est élevé, plus le nombre de hits, qui va de 731 à 20249, dans le protéome de *Conus* est élevé. (Tableau 2).
3. Plusieurs protéines de *Conus* sont ciblées par plusieurs séquences codantes pour chaque transcriptome (cas C de la figure 6). On peut faire l'hypothèse que différents transcrits issus d'un transcriptome qui matchent avec la même protéine de *Conus* peuvent être des isoformes du même gène ou des séquences codantes incomplètes (non recouvrantes, ou avec une couverture limitée qui empêche de les assembler correctement). Dans ces cas, la protéine correspondante de *Conus* n'est comptée qu'une seule fois afin de ne pas surreprésenter ces cas particuliers.

Tableau 2 : Tableau listant les transcriptomes sélectionnés.

Famille	***Espèce***	***Référence***	***BUSCO score***	*# de protéines uniques de* Conus *avec au moins un hit dans le transcriptome*
Charoniidae	*Charonia tritonis*	Bose et al. 2017	76.62	4313
Muricidae	*Rapana venosa*	Song et al. 2016	72.43	3467
Conidae	*Profundiconus* cf. *vaubani*	Fassio et al. 2019	68.76	3649
Raphitomidae	*Typhlosyrinx* sp.	[nouveau transcriptome]	68.34	3941
Turridae	*Gemmula* sp.	Zaharias et al. 2020	58.70	3348
Ovulidae	*Ovula ovum*	Lemarcis et al. 2022	47.90	2168
Colubrariidae	*Cumia reticulata*	Modica et al. 2015	47.06	2874
Raphitomidae	*Gymnobela pacifica*	[nouveau transcriptome]	44.44	2226
Personidae	*Distorsio anus*	Lemarcis et al. 2022	44.13	1877
Volutidae	*Alcithoe aillaudorum*	[nouveau transcriptome]	39.10	2610
Melongenidae	*Hemifusus tuba*	Li et al. 2019	38.89	4743
Mitridae	*Mitra mitra*	Lemarcis et al. 2022	38.16	3266
Babyloniidae	*Babylonia areolata*	[nouveau transcriptome]	36.37	2116
Vasidae	*Vasum turbinellum*	Lemarcis et al. 2022	31.24	1522
Olividae	*Oliva sericea*	Lemarcis et al. 2022	14.78	544

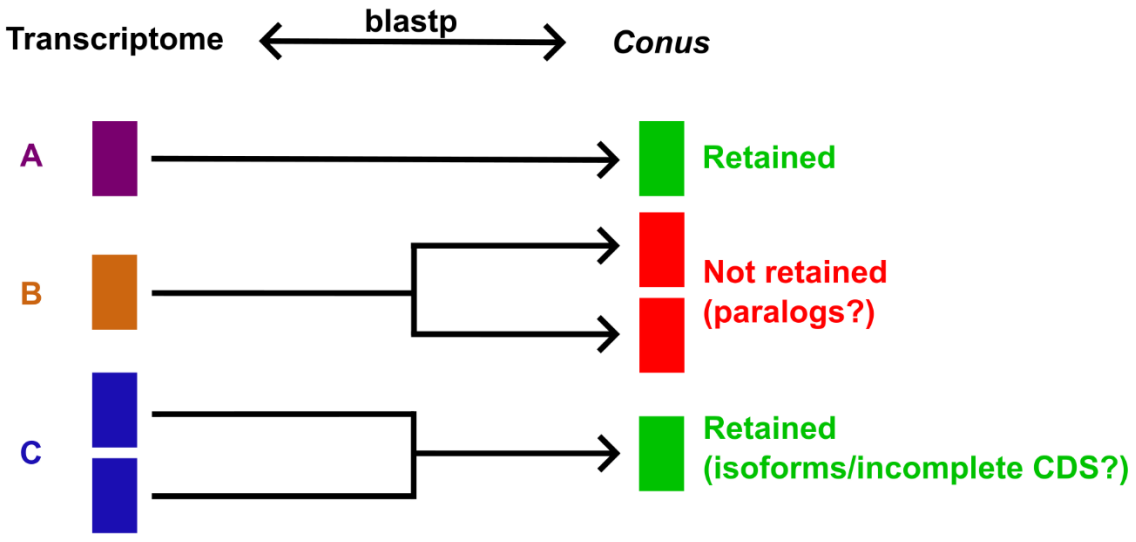


Figure 6 : Schéma représentant les différents cas possibles rencontrés lors du BLASTp entre les transcriptomes traduits en protéines que nous avons sélectionnés et le protéome de notre référence *Conus*.

Cas A : une séquence codante d'un transcriptome correspond à une seule protéine de Conus, elle est alors conservée. Cas B : une séquence codante correspond à plusieurs protéines différentes de Conus, ce sont potentiellement des paralogues, cette séquence est éliminée du jeu de données. La qualité d'assemblage du génome de Conus au niveau chromosomique ainsi que sa complétude nous laissent supposer que si une séquence du transcriptome s'aligne avec plusieurs protéines prédites provenant des séquences génomiques de Conus alors cela a plus de chances de s'expliquer par la présence de paralogues et non un potentiel manque de données ou une erreur d'assemblage. Cas C : plusieurs séquences codantes des transcriptomes s'alignent avec une même protéine de Conus, cela peut être dû au fait que ce sont des isoformes ou des séquences incomplètes, elles sont alors conservées dans le jeu de données. Les transcriptomes à notre disposition sont souvent incomplets, voire très incomplets. Si plusieurs séquences d'un transcriptome s'alignent à un même locus de Conus, ce sont probablement des artefacts, et nous avons fait l'hypothèse qu'il n'y a qu'un seul locus par transcrit (mais voir étape 5 pour une discussion complémentaire).

4. Nous avons conservé les protéines de *Conus* qui sont retrouvées dans au moins 8 des 12 meilleurs transcriptomes (en excluant les 3 transcriptomes avec le score BUSCO le plus faible). Nous avons autorisé 33% de données manquantes car il est possible qu'une protéine donnée soit absente de certains transcriptomes car elle n'a pas été séquencée (ou c'est un gène non exprimé dans le tissu séquencé) plutôt que parce qu'elle est totalement absente de la lignée, étant donné qu'elle est présente dans la majorité de nos transcriptomes. À cette étape, nous conservons un total de 1675 transcrits, qui correspondent à 7675 exons d'au moins 120 paires de bases (pb) (taille suffisante pour que les sondes puissent s'accrocher lors de la capture), ce qui représente un total de 1978743 pb.
5. Pour chaque exon de *Conus*, nous avons récupéré la séquence correspondante dans les 15 transcriptomes, avec la possibilité d'avoir plusieurs séquences par transcriptome pour un exon donné (cas C de la figure 6). Nous avons cependant retiré les protéines de *Conus* qui correspondent à plus de 100 transcrits chacune, car autant de séquences s'explique difficilement par la présence d'isoformes ou de séquences codantes incomplètes, mais plutôt par la présence de paralogues spécifiques de certaines lignées. Nous réduisons à nouveau le jeu de données à 1346 protéines, qui correspondent à 5206 exons d'au moins 120 pb qui représentent 1262591 pb.

6. Nous alignons ensuite séparément la séquence de chaque exon avec tous les transcrits correspondants à l'aide du logiciel MAFFT v7.520 (Katoh & Standley, 2013). Nous avons à cette étape un total de 5206 alignements.
7. Ensuite, tous les transcrits avec une proportion d'indels supérieure à 35% dans la région de l'exon sont retirés des alignements, ainsi que les transcrits avec une distance génétique supérieure à 50%. Cela a pour objectif de retirer tous les transcrits qui ne se sont pas alignés avec l'exon, et qui ont donc peu de chance de correspondre à une séquence homologue de l'exon de *Conus*. Si dans certains cas il reste encore plusieurs transcrits pour un transcriptome donné (cas C de la figure 6), c'est le transcrit avec la distance génétique la plus faible par rapport à l'exon de *Conus* qui est conservé. Dans le cas d'une égalité entre deux transcrits, c'est le plus long qui est conservé. À cette étape nous avons éliminé 462 exons pour lesquels, après l'application de ces filtres, aucun des transcrits avec lesquels ils se sont alignés ne sont conservés. Nous conservons à ce stade 4744 exons.
8. Un second alignement est alors effectué entre chaque exon et les transcrits restants (un maximum de 15 pour chaque exon – 1 par transcriptome), afin d'améliorer la qualité des alignements pour les étapes suivantes. Les alignements sont alors trimmés avant et après la séquence de l'exon pour garder seulement la zone des transcrits qui correspond à la séquence d'exon cible.

À cette étape, il est important de rappeler que les étapes 2 ; 5 et 7 (suppression des exons de cas B lors de l'étape de BLAST, suppression des exons de cas C avec plus de 100 séquences correspondantes dans les transcriptomes, suppression des séquences non alignées) ont pour objectifs de réduire la présence de séquences non-orthologues dans notre jeu de données final, ce qui avait été identifié comme un problème majeur lors des précédents projets de l'équipe.

9. Nous avons ensuite effectué plusieurs étapes de filtrations :
    a. Les exons associés à moins de 4 transcrits sont retirés, afin d'avoir une couverture taxonomique suffisante pour chaque exon.
    b. Nous avons aussi retiré les exons qui ont une quantité d'indels dans l'alignement supérieure à 10% de la longueur de l'exon non aligné, car comme nous ciblons des séquences codantes, elles ne devraient contenir que peu d'indels.
    c. Nous retirons les exons avec des contenus en GC extrêmes, à savoir inférieurs à 30% ou supérieurs à 70%.

d. Nous retirons les exons pour lesquels au moins un transcrit aligné a une distance génétique inférieure à 2% avec l'exon de *Conus*, afin de retirer les exons trop peu variables.

Suite à ces filtres le jeu de données inclut 3269 exons qui représentent 688774 pb. En additionnant les séquences des transcrits qui sont conservées, on obtient un total de 34715 séquences, qui représentent 6812099 pb.

10. À cette étape, nous effectuons deux tests pour vérifier notre jeu de données :

a. La variabilité des exons entre les deux transcriptomes de Raphitomidae est-elle suffisante ? Un focus étant réalisé sur les Raphitomidae, avec l'inclusion de nombreuses espèces de même genre (contrairement aux autres lignées de néogastéropodes), nous souhaitions nous assurer qu'un nombre suffisant d'exons étaient variables à cette échelle taxonomique. Par comparaison, la distance génétique pour le *cox1* est de 10,5%. Les exons des Raphitomidae utilisés pour générer le jeu de données sont donc suffisamment variables (Figure 7).

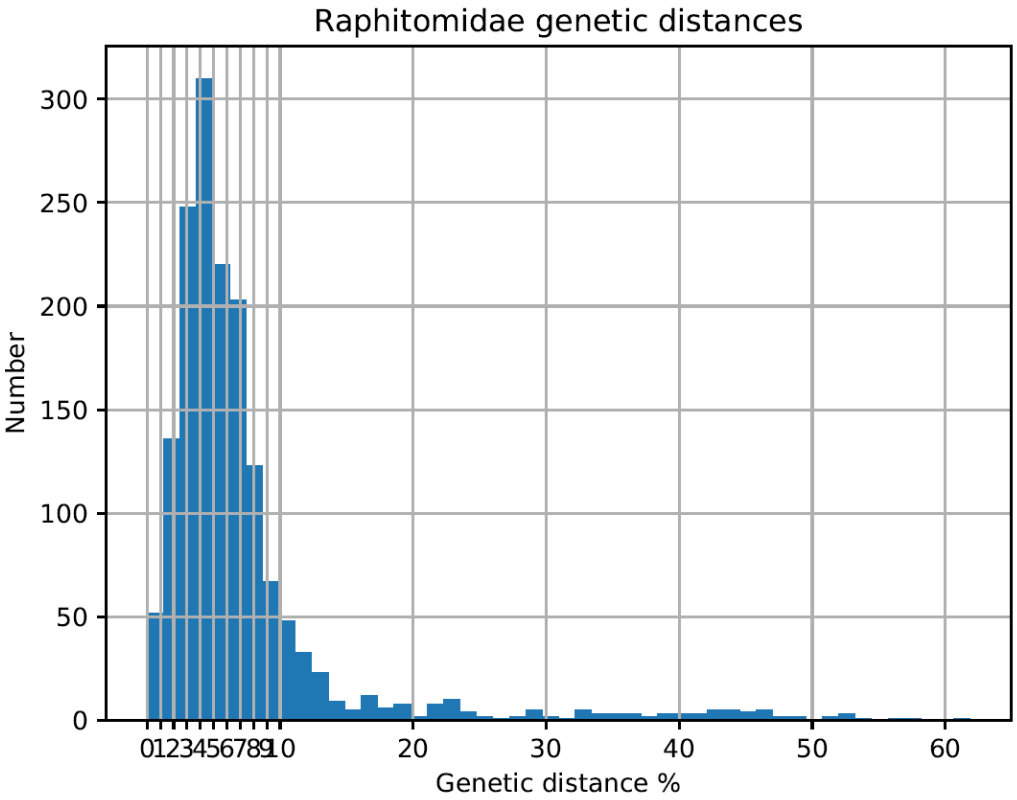


Figure 7 : Histogramme représentant le nombre d'exons en fonction de la distance génétique.-*Elle est calculée pour chaque exon entre les deux transcriptomes de Raphitomidae*.

b. Nous avons décidé de comparer les exons que nous avons identifiés avec les exons utilisés dans les autres projets de capture d'exons de l'équipe (Abdelkrim, Fedosov, Zaharias). Nous retrouvons seulement 313 exons sur les 3269 exons. Cela ne nous permettra pas de combiner les résultats de notre capture d'exons avec les résultats des captures précédentes, et nous a donc contraint à inclure dans notre jeu de données les taxons qui avaient déjà été inclus dans ces jeux de données précédents, afind e les séquencer à nouveau.

11. Ensuite, nous retirons les séquences de moins de 120 pb de longueur, qui ne comporte donc que peu d'information phylogénétique. Nous avons à ce stade 34301 séquences. Afin de réduire la quantité de sondes nécessaires pour la capture et améliorer les chances de capturer les exons d'intérêt, nous avons décidé de séparer notre jeu de données d'exons en deux sous-ensembles taxonomiques. Un premier batch comprend les super-familles des Conoidea, Buccinoidea et Mitroidea, *a priori* relativement proches phylogénétiquement (A. E. Fedosov et al., in press). Elles sont également les plus proches de notre génome de référence de *Conus*, qui est un Conoidea. Un second batch inclut les super-familles des Tonnoidea, Volutoidea, Turbinelloidea, Muricoidea et Olivoidea (et quelques spécimens de Conoidea, qui n'étaient pas disponibles au moment de la réalisation du premier batch). Autrement dit, les sondes du batch 1 sont dessinées à partir des séquences des transcriptomes des Conoidea, Buccinoidea et Mitroidea uniquement, et les sondes du batch 2 sont dessinées à partir des séquences des transcriptomes des autres groupes. Un troisième batch de séquençage regroupera les sondes des batchs 1 et 2, afin de compléter la couverture taxonomique des batchs 1 et 2, mais aussi afin de tester l'impact de la distance phylogénétique sur la capacité des sondes à capturer les exons (voir chapitre 3). Les 34301 séquences sont ainsi séparées en deux groupes, un de 20710 séquences pour le batch 1 et un de 13591 pour le batch 2.
12. Les étapes 12 à 16 vont décrire les étapes du pipeline qui ont été effectuées sur les séquences du batch 1. Du fait de la grande distance génétique qui peut parfois empêcher les sondes de se fixer sur nos séquences cibles, nous avons décidé de reconstruire des séquences ancestrales, qui sont des séquences consensus, pour trois paires de taxa, qui représentent les taxa les plus divergents dans le premier batch : *Conus* / *Gemmula*, *Conus* / *Mitra* et *Cumia* / *Hemifusus*. Ces séquences consensus ont été reconstruites avec le logiciel FastML v3.11 (Ashkenazy et al., 2012). Cela permet de réduire (diviser par deux, en théorie) la distance génétique entre les sondes définies à partir des séquences ancestrales et les exons à capturer.

7119 séquences ont été reconstruites, et parmi elles, 6293 correspondent à des paires de séquences dont la distance génétique des deux séquences d'origine était supérieure à 10%.

13. Malgré tous les filtres appliqués à nos séquences, cela représenterait à ce stade un trop grand nombre de sondes. Nous avons donc décidé de filtrer plus drastiquement notre jeu de données en modifiant le seuil de l'étape 9.b. de 10% à 5% pour la quantité d'indels autorisés, ce qui permettra par la suite de faciliter les alignements post-séquençage. Nous obtenons alors un total de 30584 séquences, qui représentent 2900 *loci* avec 18436 séquences pour le batch 1 et 12148 séquences pour le batch 2.
14. Sur les 7119 séquences ancestrales pour le batch 1, nous conservons 6340 séquences qui correspondent aux *loci* retenus dans le nouveau jeu de données de l'étape 13. Parmi ces séquences, nous retenons 5565 séquences qui correspondent aux paires de séquences pour lesquelles la distance génétique entre les deux séquences d'origine était supérieure à 10%.
15. Nous avons à ce stade un jeu de données comprenant 24001 séquences (18436 séquences d'origine + 5565 séquences ancestrales) pour le premier batch.

1.3. DESIGN DES SONDES

16. Nous avons envoyé les 24001 séquences à l'entreprise MyBaits (Daicel Arbor, Ann Arbor, MI, USA) qui s'est chargée de la synthèse des sondes. Afin d'avoir un kit de sondes final de 60000 sondes (car sinon nous aurions dû utiliser un kit bien plus gros et donc plus cher), ils ont appliqué plusieurs filtres :
    a. Masquage des répétitions simples.
    b. Comparaison par BLAST des séquences des sondes contre 3 génomes de référence, le génome de *Conus ventricosus*, celui de *Monoplex corrugatus* et celui de *Stramonita haemastoma*, pour identifier les sondes qui capturent des *loci* sur-représentés dans le génome et regrouper les sondes avec des séquences similaires avec un seuil de 95%. Les génomes de *Monoplex* et *Stramonita* ont été obtenus par l'équipe pendant ma première année de thèse et n'étaient pas encore disponibles lors de la sélection des exons cibles (Farhat et al., 2023).
    c. Sélection des *loci* les plus longs (supérieurs à 160 pb de longueur), afin d'augmenter encore la quantité d'information phylogénétique portée par chaque exon.

Nous avons ainsi un jeu de données final de 1125 exons qui représentent 9429 séquences (7208 séquences d'origine et 2221 séquences ancestrales).

17. Afin de conserver les mêmes *loci* pour les batchs 1 et 2, nous avons retiré des 2900 *loci* de l'étape 13 pour le batch 2 les *loci* qui n'étaient pas inclus dans les 1125 retenus finalement. Nous avons à cette étape 4580 séquences pour le batch 2.
18. Pour le batch 2 nous avons également reconstruit des séquences ancestrales. Nous avons fait 3 paires pour les taxa les plus divergents : *Charonia* / *Babylonia*, *Charonia* / *Alcithoe* et *Charonia* / *Rapana*. Pour les exons où *Charonia* n'était pas représenté par une séquence, nous avons reconstruit des paires avec *Alcithoe* / *Babylonia* et *Alcithoe* / *Rapana*. Cela nous a permis de reconstruire 1949 séquences ancestrales. Parmi elles, nous avons 1896 séquences qui correspondent à des paires de séquences dont la distance génétique entre les 2 séquences d'origine est supérieure à 10%.
19. Nous obtenons finalement un jeu de données avec 6476 séquences (4580 séquences d'origine et 1896 séquences ancestrales) pour le batch 2.
20. Nous avons alors envoyé les séquences à MyBaits qui a appliqué les mêmes filtres que pour le batch 1 afin d'obtenir un kit de 40000 sondes de capture.

Ce protocole de design des sondes a été présenté sur un poster lors du World Congress of Malacology en août 2022 à Munich, présenté sur la page suivante.

Exon design for large-scale phylogeny of the Neogastropoda

Thomas Lemarcis[1], Jawad Abdelkrim[2], Alessandro Derzelle[1], Paul Zaharias[1], Yuri I. Kantor[1,3], Alexander E. Fedosov[1,3], Nicolas Puillandre[1]

[1]Institut Systématique Evolution Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, Paris, France. thomas.lemarcis@mnhn.fr
[2]UAR 2700 'Acquisition et Analyse de Données pour l'Histoire naturelle', Service d'Analyse de Données, CNRS, Muséum National d'Histoire Naturelle, Sorbonne Universités, CP26, 43 rue Cuvier, 75231 PARIS CEDEX 05
[3]A.N. Severtzov Institute of Ecology and Evolution, Russian Academy of Sciences, Moscow, Russia.

CONTEXT

- The published neogastropod phylogenies are mostly unresolved and largely incomplete
- Goal: to produce an, as complete as possible, phylogeny using an exon-capture approach
- However, previously used set of baits are not able to capture all specimens/all exons (lots of missing data), and many nodes remain unresolved (Abdelkrim *et al.* 2018): design of a new set of baits.

Reference:

DATA

Distorsio, *Profundiconus*, *Alcithoe*, *Gymnobela*, *Hemifusus*, *Charonia*, *Oliva*, *Ovula*, *Vasum*, *Rapana*, *Babylonia*, *Mitra*, *Typhlosyrinx*, *Cumia*, *Gemmula*

15 transcriptomes

1 complete annotated proteome

Conus ventricosus

METHODS

1. BLASTp e-value 10^{e-10}
2. Keeping *Conus* proteins found in :
   - Unique BLAST match
   - >75% of the transcriptomes
   - with <100 transcripts
3. Alignment between *Conus* exons and contigs from transcriptomes
4. Keeping sequences with indels <35% and PID >50%
5. 2nd alignment between *Conus* exons and contigs from transcriptomes

Reconstruction of ancestral sequences

We reconstructed an ancestral sequence for 6 pairs of the most divergent taxa.
1st batch : *Conus* / *Gemmula*, *Conus* / *Mitra* and *Cumia* / *Hemifusus*
2nd batch : *Charonia* / *Babylonia*, *Charonia* / *Alcithoe* and *Charonia* / *Rapana*

1st batch : 2,221 ancestral sequences
2nd batch : 1,896 ancestral sequences

6. Trimming transcript sequences at *Conus* exon boundaries
Removing exons with :
   - <4 transcripts aligned
   - >5% indels
   - GC content <30% or >70%
   - genetic diversity <2%

Pipeline Legend: BLAST, Filtering, Alignments

7. Splitting dataset in 2 batches to reduce genetic distance

1st batch: CONOIDEA MITROIDEA BUCCINOIDEA

2nd batch: OTHER NEOGASTROPODA + TONNOIDEA/FICOIDEA

RESULTS

FINAL DATASET
1,125 exons
Probe design:
7,208 original sequences + 2,221 ancestral sequences for 1st batch
6,476 original sequences + 1,896 ancestral sequences for 2nd batch

1st batch:
- 384 samples (5 samples removed during assembly (no reads))
- 1,124 exons captured
- An average of 1,006 exons per sample
- An average of 326 samples per exon

Phylogenetic Tree:
- 371 correctly placed samples
- 8 samples with doubtful placement

CONOIDEA
MITROIDEA
BUCCINOIDEA

This action has received funding from the European Union's Horizon 2020 research and innovation program & European Research Council under grant agreement ERC HYPERDIVERSE (GRANT AGREEMENT 865101)

2. EXTRACTION ET QUANTIFICATION D'ADN

Je vais aborder dans cette sous-partie la méthodologie qui a été suivie au laboratoire pour la sélection des échantillons ainsi que les différents tests effectués pour exploiter des ressources de collection disponibles.

Pour reconstruire la phylogénie des néogastéropodes, j'ai eu accès à plusieurs types d'échantillons, plus ou moins bien conservés. Tout d'abord, la majorité des échantillons utilisés sont des échantillons récents, collectés lors de missions organisées par l'équipe 3E, depuis 2003 (date à laquelle ont été mis en place sur le terrain les premiers protocoles de conservation des échantillons pour le séquençage ADN). Pour ces échantillons, l'ADN était déjà extrait avant ma thèse ou l'a été pendant la thèse, et il a fallu quantifier la concentration en ADN afin de faire une sélection des échantillons avec les meilleures concentrations en ADN. Cependant, certains groupes taxonomiques de néogastéropodes, parfois même des familles entières, n'étaient pas représentées parmi ces échantillons récents. J'ai donc également exploré les collections plus anciennes, incluant des spécimens collectés vivants, mais qui n'avaient pas été conservées spécifiquement pour du séquençage ADN, et dont l'état de conservation est variable : fixation au formol ou dans de l'alcool faiblement concentré, éventuellement séchés par la suite, puis conservés (en fluide ou non) pendant plusieurs décennies. L'ADN sera par conséquent plus ou moins bien conservé, à savoir plus ou moins fragmenté ou dégradé. La fragmentation de l'ADN n'est en soit pas un souci majeur pour l'approche de capture d'exons, étant donné que même des petits fragments peuvent être capturés par les sondes, bien qu'il soit nécessaire que cet ADN soit présent en concentration suffisante. Ceci permet que les exons d'intérêt soient capturés de la manière la plus homogène possible entre les différents échantillons afin de pouvoir les comparer et reconstruire notre phylogénie par la suite, en minimisant la part de données manquantes. Enfin, certains taxons rares n'étaient pas non plus présents dans ces collections anciennes fixées en fluides, et j'ai exploré (sans succès – Voir partie 2.1) la possibilité d'extraire de l'ADN à partir des coquilles collectées vides et présentes dans les collections du Muséum National Histoire Naturelle de Paris (MNHN). J'ai pour cela appliqué un protocole qui ne détruit pas la coquille (que l'on pourrait qualifier de semi-invasif), et qui maximise les chances d'obtenir assez d'ADN pour que la capture d'exons fonctionne. Je détaille ce protocole d'extraction d'ADN dans la partie suivante, suivie par une présentation des méthodes qui ont été testées et appliquées pour quantifier les ADN extraits.

2.1. EXTRACTION D'ADN A PARTIR DES COQUILLES :

De nombreux protocoles sont déjà disponibles pour extraire de l'ADN à partir d'os (Dabney et al., 2013; Gamba et al., 2014, 2016; D. Y. Yang et al., 1998). L'ADN extrait par ces différentes méthodes permet de répondre à des questions diverses en biologie mais aussi en paléontologie, anthropologie et archéologie. Par conséquent, l'attrait pour l'exploitation des collections anciennes conservées dans les muséums est important. En effet, les espèces conservées en alcool (ou plus généralement pour le séquençage ADN) ne représentent souvent qu'une fraction de la diversité des spécimens collectés pour un groupe donné. Avec l'accélération de la disparition des espèces vivantes, de nombreuses espèces en danger d'extinction ou éteintes ne sont représentées dans les collections que par des spécimens qui n'ont pas été préservés correctement en vue d'effectuer du séquençage ADN. De plus, certaines espèces sont très rares et n'ont pu être collectées qu'une seule fois sur le terrain, ou vivent dans des milieux difficiles d'accès, tels que les plaines abyssales ou les forêts isolées. Des équipes travaillant sur d'autres groupes taxonomiques ont déjà utilisé des spécimens de collection pour en extraire leur ADN, notamment en entomologie (Tin et al., 2014) mais aussi en ichtyologie (Silva et al., 2019). Il existe aussi des cas particuliers, comme pour de nombreuses espèces de mollusques, pour lesquelles il n'a jamais été possible de collecter des spécimens vivants mais seulement leurs coquilles (Bouchet et al., 2002).

La richesse des collections de mollusques du MNHN nous donne accès à des coquilles de nombreuses espèces qui n'ont pas encore pu être séquencées. Cela pourrait nous permettre d'augmenter la diversité des taxons séquencés et de parfois combler des manques dans nos phylogénies, comme cela a déjà été fait avec succès sur une espèce en danger critique d'extinction qui n'a jamais été collectée vivante sur le terrain, *Levantina rechingeri* (Psonis et al., 2022). Cela a permis également le séquençage de spécimens de collection datant de plusieurs dizaines voire des centaines d'années (Walton et al., 2023). Néanmoins, obtenir de l'ADN à partir des coquilles ne doit pas se faire au détriment de la coquille, qui doit rester intact ou presque, notamment pour les identifications morphologiques. Pour mettre en place un protocole au MNHN et tenter d'extraire de l'ADN à partir des coquilles, j'ai réalisé une revue bibliographique des techniques existantes chez les mollusques mais également pour d'autres groupes taxonomiques, comme pour les extractions faites à partir d'os. On suppose que la matrice des coquilles peut contenir de l'ADN qui serait piégé lors de la croissance de la coquille du mollusque. Un récent protocole d'extraction d'ADN à partir de petits mollusques (Goulding

et al., 2021; Inäbnit et al., 2021) permettrait d'extraire l'ADN à partir des coquilles. Cependant, ce protocole ne peut pas s'appliquer à nos échantillons car il a été utilisé sur des mollusques de petites tailles (1 à 2,5mm) et avec des coquilles très fines, comparées aux coquilles de néogastéropodes, et qui ont été entièrement digérées pour en extraire le plus d'ADN possible.

Par ailleurs, plusieurs équipes ont mis au point des protocoles d'extraction en réduisant la coquille en poudre (Ferreira et al., 2020; Geist et al., 2008; Goulding et al., 2021). Dans notre cas, le souci est double : cela entraîne d'une part la destruction d'une grande partie, voire de la totalité de la coquille, ce qui n'est pas envisageable pour des spécimens de collection du MNHN, et d'autre part la dureté des coquilles des néogastéropodes ne nous permet pas d'envisager cette solution pour nos spécimens de collection. Par conséquent, une autre méthode, moins destructrice, basée également sur la récupération de poudre à partir de la coquille a été appliquée. Dans deux articles, Der Sarkissian et al. (Der Sarkissian et al., 2017, 2020) utilisent les bases du protocole mis en place par Yang et al. en 1998 (D. Y. Yang et al., 1998) puis modifié par Gamba et al. en 2014 et 2016 (Gamba et al., 2014, 2016). Ce protocole était à l'origine utilisé pour récupérer de la poudre d'os, en meulant la surface de l'os, puis de la dissoudre le plus possible avec un tampon de lyse, et ensuite passer les échantillons sur une colonne de silice issue d'un kit de purification de produits PCRs. L'avantage de ces colonnes est la possibilité de récupérer des fragments d'ADN de petite taille allant de 100 pb et jusqu'à 10 kb. Der Sarkissian et ses collaborateurs ont adapté cette méthode pour l'utiliser sur de la poudre de coquille, en utilisant une petite meuleuse afin de poncer la surface interne de la coquille. Cette méthode est moins destructrice pour la coquille, notamment pour les parties externes qui peuvent servir à une identification morphologique. En utilisant une méthode similaire, une autre équipe de Nouvelle-Zélande (Walton et al., 2023) a séquencé des mitogénomes quasi complets à partir de 12 échantillons de coquilles collectées il y a plus de 60 ans et conservées à sec. La différence majeure avec le protocole adapté par Der Sarkissian est le fait d'ajouter une phase de fixation du surnageant qui est obtenu après la lyse de la poudre de coquille (Figure 8). De plus, un autre avantage conséquent en termes de conservation de la coquille du spécimen de collection est le fait que ce protocole utilise une quantité bien plus faible de poudre de coquille par rapport à celui de Der Sarkissian. Ce protocole est aussi une adaptation d'un autre protocole à l'origine utilisé pour extraire de l'ADN à partir d'os d'un Ursidae et publié par Dabney et al. en 2013 (Dabney et al., 2013). On retrouve ici le même avantage que pour le protocole de Yang, à savoir l'utilisation de colonnes de purification de produits PCRs qui devraient permettre la récupération de fragments d'ADN de petite taille

(jusqu'à 100 pb). En effet, on s'attend à ce que l'ADN des spécimens de collection anciens soit fragmenté, et c'est donc une nécessité de pouvoir extraire et conserver les plus petits fragments lors de l'extraction. J'ai choisi d'utiliser le protocole adapté utilisé par Walton et al. car les rendements en termes de qualité et de quantité d'ADN extrait à partir des coquilles semblent plus importants par rapport aux protocoles de Der Sarkissian, en utilisant une plus faible quantité de poudre de coquille.

Dans ce protocole, environ 50 mg de poudre de coquille sont prélevés au niveau du dernier tour de la columelle (Figure 8). Cette poudre est ensuite mise dans un tampon de lyse composé de 1 mL d'EDTA à pH 8 + 1 µL de Tween 20 et 12,5 µL de Proteinase K pour 24 heures à 37°C sous agitation constante afin de dissoudre la poudre. Un volume de 1 mL du surnageant est ensuite transféré dans 13 mL de tampon de liaison composé de 5 M de Guanidine Chlorohydrate, de 90 mM de Sodium Acetate à pH 5,2 de 40% d'isopropanol et de 0,05% de Tween 2. On transfère ensuite la totalité des 14 mL obtenus dans un réservoir pour faire passer la totalité du volume dans la colonne de purification. Les échantillons sont centrifugés pendant 4 minutes à 1500 g. La colonne de purification est alors transférée dans un tube de 2 mL vide pour être séchée par centrifugation pendant 1 minute à 3300 g. Deux étapes de lavage de la colonne de purification sont ensuite effectuées avec le tampon PE qui est fourni dans le kit de purification MinElute de Qiagen. Une centrifugation à 3300 g pendant 2 minutes est effectuée à chaque étape de lavage. La colonne est ensuite séchée par centrifugation pendant 1 minute à vitesse maximale (16100 g). Puis l'ADN est élué dans un volume de 25 µL de tampon TET en deux fois afin de récupérer le maximum d'ADN possible lors de cette phase. Toutes ces expérimentations ont été réalisées et mises en place avec l'aide de Julie Vasseur, ingénieure d'étude sur la plateforme du Service de Systématique Moléculaire (SSM) du MNHN.

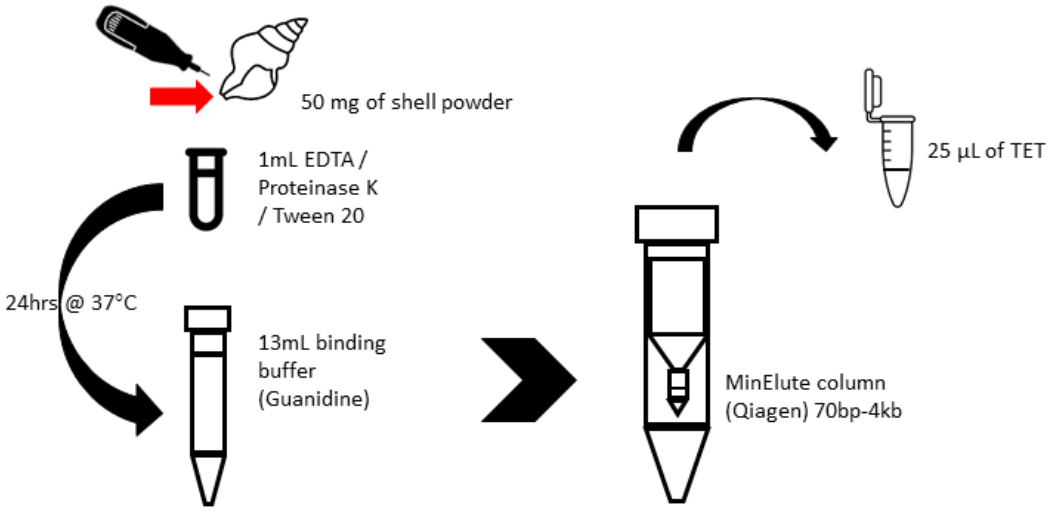


Figure 8 : Schéma représentant le protocole d'extraction d'ADN à partir de poudre de coquille.
Mis en place par Dabney et al. en 2013 puis adapté aux coquilles de mollusque par Walton et al. en 2023.

J'ai effectué les tests de ce protocole sur 5 spécimens collectés récemment (Tableau 3), appartenant à des espèces communes (et donc avec de nombreux échantillons en collection). Pour chacun, j'ai réalisé deux réplicats, pour un total de 10 échantillons. J'ai utilisé des spécimens dont les tailles de coquille étaient variables, et j'ai de plus pris des quantités de poudre différentes entre les deux réplicats afin de tester l'impact de la quantité de poudre sur le résultat de l'extraction. Les quantités vont de 10 mg à 54 mg de poudre par échantillon (Tableau 3).

Pour vérifier si l'extraction a fonctionné, des PCRs de contrôle sur les échantillons extraits ont été réalisées : une PCR de la Cytochrome Oxydase I (*cox1*) en deux fragments d'environ 350 pb chacun et une PCR d'un fragment de la protéine histone 3 (H3) d'environ 150 pb. Ces PCRs se sont toutes révélées négatives. J'ai également quantifié l'ADN de ces échantillons et un seul échantillon a pu être quantifié avec la méthode Qubit (voir partie 2.2), avec 1,24 ng/µL d'ADN (Tableau 3) ; les autres ADNs étaient tous indétectables. Cependant, j'ai pu constater que les quantifications d'ADN sur nos échantillons ne sont pas toujours fiables : en effet, j'ai pu réaliser des PCRs *cox1* de bonnes qualités sur des échantillons qui n'avaient pas pu être quantifiés en Qubit (voir partie 2.2).

Tableau 3 : Liste des échantillons utilisés afin de réaliser les tests d'extraction d'ADN à partir des coquilles de néogastéropodes.

N° temporaires	Expédition	Famille	Masse de poudre (mg)	Qubit (ng/µL)	COI (CODEX)	H3
1	1972-1	Nassariidae	50	1.24	Neg	Neg
2	1972-2	Nassariidae	35	0.354	Neg	Neg
3	1927-1	Nassariidae	10	0.152	Neg	Neg
4	1927-2	Nassariidae	24	0.24	Neg	Neg
5	Lucira-1	Muricidae	12	0.44	Neg	Neg
6	Lucira-2	Muricidae	47	0.406	Neg	Neg
7	1952-1	Fasciolaridae	14	0.22	Neg	Neg
8	1952-2	Fasciolaridae	49	0.338	Neg	Neg
9	1892-1	Nassariidae	10	0	Neg	Neg
10	1892-2	Nassariidae	54	0.14	Neg	Neg

Je n'ai pas pu poursuivre les tests sur ce protocole, mais il est possible que nous ayons réussi à extraire de l'ADN malgré tout de nos échantillons. L'équipe de Walton avait séquencé avec une approche en shotgun les échantillons extraits, et il se peut que les petites tailles de fragment ne permettent pas d'effectuer des PCRs sur nos échantillons. Il serait important de poursuivre les tests, par exemple en séquençant ces échantillons déjà extraits avec une approche NGS pour confirmer ou infirmer la présence d'ADN. En effet, la mise en place d'un protocole robuste d'extraction d'ADN à partir de coquilles de collection est un enjeu trop important pour être laissé de côté. La comparaison entre plusieurs protocoles déjà mis en place, qui ont été cités plus haut dans ce paragraphe, et leur adaptation semble possible dans des délais raisonnables. Nous pourrions alors utiliser les ressources de la plus grande collection mondiale de mollusques pour compléter les phylogénies et faciliter l'attribution des noms aux lignées, y compris aux espèces délimitées moléculairement, via le séquençage de spécimens-types.

2.2. QUANTIFICATIONS ADN : TESTS, CHOIX DU PROTOCOLE QUBIT

Suite aux extractions d'ADN, il est important de réaliser des quantifications afin de sélectionner les spécimens avec les plus fortes concentrations en ADN pour réaliser la capture d'exons. Cependant, nous avons constaté de manière empirique dans l'équipe qui travaille sur de l'ADN de mollusques depuis plus de 15 ans, qu'il est difficile de se fier aveuglément aux

quantifications d'ADN total sur les échantillons extraits au laboratoire. Il ne semble pas se dégager un pattern régulier qui pourrait expliquer les raisons pour lesquelles les quantifications peuvent varier pour un même échantillon. En effet, il est possible de passer de quantités d'ADN indétectables ou presque à une quantité bien plus importante sur le même extrait d'ADN, d'une expérience à l'autre, sans qu'aucun élément extérieur ne puisse l'expliquer. De même, différents échantillons issus de spécimens conservés de manière identique, avec des ADNs extraits avec la même méthode et conservés dans les mêmes conditions, peuvent avoir des résultats de quantification d'ADN très variables entre eux. De plus, ces résultats de quantification peuvent ne pas être corrélés avec le succès des PCRs ou le séquençage de ces échantillons. Cette étape de quantification reste cependant importante car nous souhaiterions sélectionner les échantillons pour la capture d'exons en prenant en compte ce critère pour évaluer *a priori* la qualité des ADNs.

Ce sont ces raisons qui nous ont conduit à mettre en place un protocole contrôlé de quantification d'ADN afin de sélectionner la méthode de quantification la plus robuste possible. Nous avons décidé de comparer trois méthodes de quantification différentes :

(i) une quantification au Qubit, qui est celle habituellement utilisée dans l'équipe. Cette méthode de quantification utilise la fluorescence ; les échantillons vont être mis en présence d'un fluorochrome capable de se fixer au squelette de la molécule d'ADN. L'émission du fluorochrome va changer de couleur selon qu'il est fixé à l'ADN ou non et c'est cette émission que l'on va quantifier ;

(ii) une quantification en utilisant le Fragment Analyzer, par électrophorèse à capillaire : malgré le fait que cet appareil est conçu pour qualifier la taille et l'intégrité des fragments d'ADN dans un échantillon, soit d'ADN génomique, soit de banques NGS, il peut nous permettre également de connaître la concentration en ADN pour un échantillon donné ;

(iii) une quantification au Droplet Digital PCR, un appareil conçu pour effectuer des PCR quantitatives avec des valeurs absolues de concentrations d'ADN données en nombre de copies d'un fragment amplifié par microlitre. L'intérêt de cette méthode tient au fait qu'elle fournit une valeur absolue de la concentration, et donc une meilleure précision de la quantification *a priori*. Il est alors nécessaire d'effectuer une conversion entre le nombre de copies quantifiées d'un fragment et la taille du génome du spécimen considéré. Cependant, même lorsque la taille du génome n'est pas connue, cela permet de comparer les échantillons d'un même taxon afin de choisir le meilleur échantillon pour le séquençage.

Pour effectuer ces tests, j'ai constitué un jeu de données incluant plusieurs ADNs extraits que j'ai quantifié avec les trois méthodes. Des spécimens de la famille des Raphitomidae, dont les ADNs ont été extraits avec la même méthode et dans le même laboratoire, ont été choisis. Nous avons décidé de prendre des spécimens avec des « âges » différents pour tester l'impact éventuel sur les résultats des quantifications. Au total, 12 échantillons ont été sélectionnés, ainsi que 3 standards dont on connaît la quantité d'ADN (Tableau 4). Pour chaque échantillon j'ai fait des triplicats, c'est-à-dire que l'extrait d'ADN a été séparé en 3 aliquots, afin de tester la répétabilité de l'expérience, étant donné que nous avons constaté une variabilité non négligeable lors des quantifications (notamment au Qubit) par le passé.

Les échantillons ont été transférés depuis la plaque mère de stockage d'ADN vers des tubes temporaires afin de limiter l'ouverture et la fermeture des tubes-stocks. 20 µL d'ADN de chaque échantillon est transféré dans des tubes de 0,5 mL et sont stockés à +4°C.

Tableau 4 : Échantillons utilisés pour tester les trois méthodes de quantification d'ADN.

"Age"	Préparation	Profondeur	N°MNHN	Expédition	Genre	Espèce	Qubit 1 (ng/µL)	Qubit 2 (ng/µL)	Qubit 3 (ng/µL)	Fragment Analyzer 1 (ng/µL)	Fragment Analyzer 2 (ng/µL)	Fragment Analyzer 3 (ng/µL)	ddPCR 1 (ng/µL)	ddPCR 2 (ng/µL)	ddPCR 3 (ng/µL)
récent	Micro-onde	Profond	IM_2013-48181	KANADEEP	*Spergo*	*fusiformis*	8.13	9.25	21.85	5.12	42	12.04	498	554	612
récent	Micro-onde	Profond	IM_2013-48196	KANADEEP	*Spergo*	*fusiformis*	7.84	5.79	8.02	7.72	9.18	8.76	302	344	424
récent	Micro-onde	Côtier	IM_2019-6087	CORSICABENTHOS 2019	*Raphitoma*	sp.	0.00	0.00	0.00	0.97	1.21	0.84	32.2	0	42.8
récent	Micro-onde	Côtier	IM_2019-6090	CORSICABENTHOS 2019	*Raphitoma*	sp.	23.65	4.84	5.12	5.23	11.35	3.96	390	0.42	394
moyen	Micro-onde	Profond	IM_2013-9837	PAPUA_NIUGINI	*Austrobela*	*micraulax*	2.40	9.89	14.88	8.48	14.15	9.23	376	380	388
moyen	Micro-onde	Profond	IM_2013-9842	PAPUA_NIUGINI	*Austrobela*	*micraulax*	0.21	0.00	0.00	5.28	2.71	5.54	135.6	145.6	159.4
moyen	Micro-onde	Profond	IM_2013-59256	ZhongSha_2015	*Spergo*	*fusiformis*	0.00	0.00	0.00	4.65	1.82	2.09	60.8	75.2	86.6
moyen	Micro-onde	Profond	IM_2013-59288	ZhongSha_2015	*Spergo*	*fusiformis*	5.48	0.00	0.00	0.43	4.85	0.95	54.4	59.4	62.4
ancien	Manuelle	Profond	IM_2009-18323	SANTO_2006	*Gymnobela*	sp.	0.00	5.21	8.32	3.62	4.52	4.82	109	126.2	128.2
ancien	Manuelle	Profond	IM_2009-16933	AURORA_2007	*Spergo*	*sibogae*	16.21	64.60	20.77	11.69	22.96	10.59	178	212	212
ancien	Manuelle	Profond	IM_2009-18324	PANGLAO_2005	*Buccinaria*	*urania*	0.00	0.00	0.00	0.007	0.04	0.081	0	0	1.8
ancien	Manuelle	Profond	IM_2009-18325	PANGLAO_2005	*Buccinaria*	*jonkeri*	0.00	0.32	1.20	1.35	3.39	2.94	65.8	79.2	88
			Standard 0 ng/µL				0.00	0.00	0.00	0.45	0.04	0.15	0.16	0	0
			Standard 50 ng/µL				42.05	44.10	38.27	92.14	53.88	17	650	650	650
			Standard 100 ng/µL				69.23	87.41	97.15	135.32	177.98	196.46	650	650	650

- Qubit :

Les quantifications au Qubit sont faites en plaque car c'est la méthode qui sera utilisée pour quantifier les nombreux échantillons qui seront utilisés pour la capture d'exons. Ici un protocole standard fourni par le constructeur du Kit Qubit BR (Invitrogen) a été utilisé. Pour chaque échantillon et les standards, j'utilise 1 µL de volume d'extrait. Les résultats obtenus sont présentés dans le tableau 4 : on constate qu'il existe une variabilité (relativement faible – voir ci-dessous les résultats pour le Fragment Analyser) des résultats entre les différents réplicats pour un même échantillon, notamment en ce qui concerne les standards, et en particulier pour les standards contenant 100 ng/µL d'ADN (Figure 9). En revanche, on constate que pour le standard qui ne contient pas d'ADN (0 ng/µL), la quantification est systématiquement négative. Cela est un point positif qui nous permet de considérer qu'un échantillon dont la quantification serait négative a peu de chance de contenir de l'ADN, et qu'il peut donc être éliminé, avec une faible probabilité de se tromper. Les échantillons d'âges médians et anciens, dont l'ADN est conservé depuis plus longtemps, présentent des quantités d'ADN plus faibles que les échantillons récents. Il existe néanmoins des exceptions, par exemple, l'échantillon 10 qui donne de meilleurs résultats en termes de quantité d'ADN que des échantillons plus récents (Tableau 4). Seuls 3 échantillons sur les 12 au total sont totalement indétectables sur les 3 réplicats réalisés.

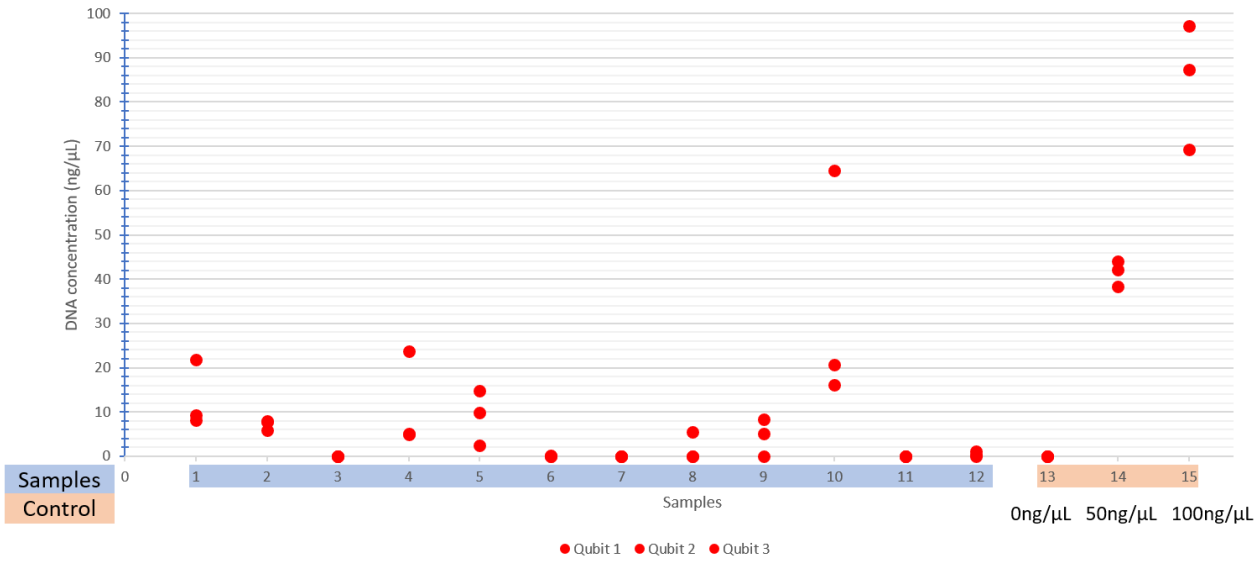


Figure 9 : Graphique représentant les résultats des quantifications réalisées avec le qubit, les différents échantillons sont représentés en abscisse et les concentrations mesurées (ng/µL) en ordonée. *Dans le cadre bleu sont représentés les 12 échantillons utilisés pour lesquels nous avons réalisés 3 réplicats pour chacun. Dans le cadre orange sont représentés les 3 échantillons contrôles que nous avons utilisés et dont nous connaissions les concentrations en ADN : 0 ng/µL, 50 ng/µL et 100ng/µL.*

- Fragment Analyzer :

Cette machine utilise l'électrophorèse à capillaires pour qualifier et quantifier les échantillons. Cette méthode permet d'analyser un grand nombre d'échantillons rapidement, jusqu'à 96 en une fois. On constate dans le tableau 4 que les résultats pour les échantillons restent variables entre chaque réplicat, et cette variabilité est plus importante que pour le Qubit (Figure 10). Il reste important de noter que les quantifications des standards sont difficiles à interpréter. En effet, on peut voir que seul le standard négatif est quantifié comme attendu. Pour les standards de 50 ng/µL et de 100 ng/µL, la quantification n'est pas suffisamment précise. Pour le standard de 50 ng/µL, les trois quantifications sont très variables allant de 17 ng/µL à 92 ng/µL. De même pour les quantifications du standard de 100 ng/µL, qui sont systématiquement supérieures à cette valeur, jusqu'à presque le double. Ces résultats pour les standards rendent les données difficilement interprétables pour les échantillons tests et, il nous est difficile de statuer sur la validité de ces résultats.

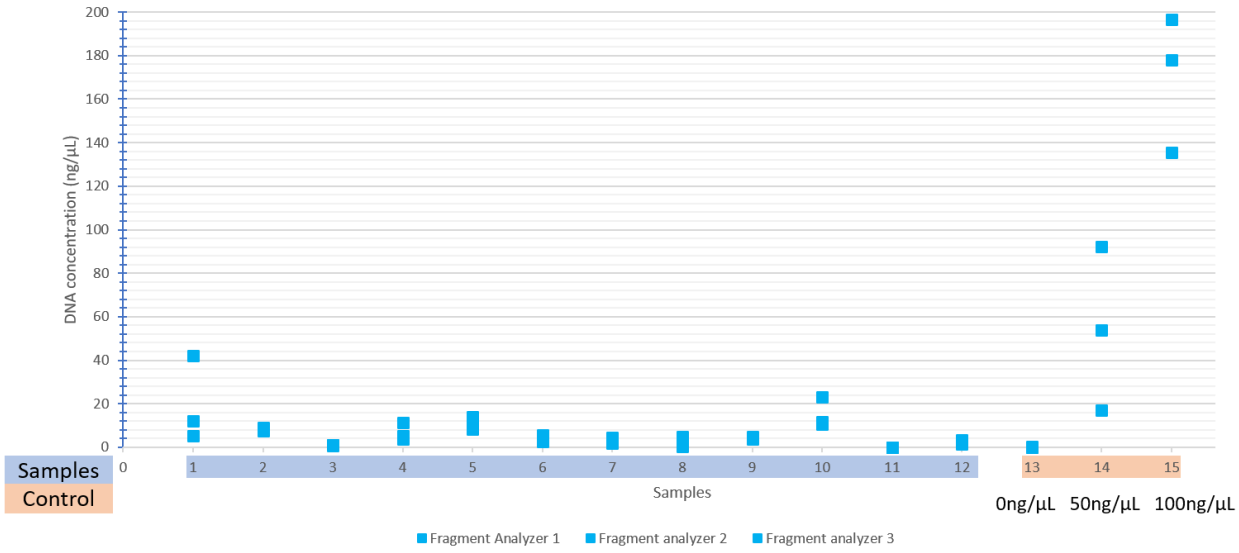


Figure 10 : Graphique représentant les résultats des quantifications réalisées avec le fragment analyzer, les différents échantillons sont représentés en abscisse et les concentrations mesurées (ng/µL) en ordonée. *Dans le cadre bleu sont représentés les 12 échantillons utilisés pour lesquels nous avons réalisés 3 réplicats pour chacun. Dans le cadre orange sont représentés les 3 échantillons contrôles que nous avons utilisés et dont nous connaissions les concentrations en ADN : 0 ng/µL, 50 ng/µL et 100ng/µL.*

- Droplet Digital PCR :

La Droplet Digital PCR (ddPCR) est une méthode de PCR quantitative (qPCR) absolue. Contrairement aux méthodes de qPCR classiques, la quantification se fait sur un échantillonnage d'environ 20000 bulles de PCRs au sein de chaque échantillon. Lors de la réaction, des bulles isolées les unes des autres vont se former et c'est à l'intérieur de chacune des billes que les PCRs vont s'effectuer. La fluorescence émise sera alors quantifiée sur l'ensemble de toutes les billes produites et, en suivant une loi de Poisson, une quantification absolue de la quantité d'ADN au cours du temps sera donnée. Ce sont ces propriétés que nous allons utiliser dans le but de quantifier de façon la plus précise possible l'ADN présent dans chacun de nos échantillons.

Cette technologie utilise une PCR pour amplifier les fragments d'ADN cibles, il est donc nécessaire d'utiliser des primers dont nous sommes sûrs qu'ils vont fonctionner sur l'ensemble de nos échantillons tests. Nous avons décidé de cibler un fragment du gène H3, aisément amplifiable chez nos mollusques. Ce fragment du gène H3 amplifié en routine dans l'équipe fait une taille d'environ 350 pb. Cependant, pour réaliser la ddPCR il est indispensable que le fragment amplifié fasse une taille comprise entre 60 et 200 pb pour être contenu dans les bulles produites lors de la réaction. Avec l'aide d'un membre de l'équipe, Dario Zuccon, nous avons donc dessiné des amorces H3 afin d'amplifier un fragment de 176 pb (-H3_ddPCR_F - 5' – ATCCGYCGTTACCAGAARAGCAC – 3' ; -H3_ddPCR_R - 5' – GGATGGCRCACAGGTTGGTGTC – 3'). Avant d'effectuer les ddPCRs sur les 12 échantillons, j'ai testé l'effet de la dilution ou non des échantillons, ainsi que la température d'hybridation des amorces, car ce sont des paramètres qui seront utilisés pour effectuer les tests avec la ddPCR. Afin de s'assurer d'avoir un ADN de qualité pour ces premiers tests, j'ai sélectionné 8 échantillons différents collectés récemment (Tableau 5), puis effectué des dilutions au 1/20ème des 8 échantillons : les échantillons dilués et non dilués ont été testés. Ces

échantillons ont été testés avec deux températures d'hybridation différentes, 53°C et 57°C, pour évaluer l'impact sur la qualité et la quantité d'ADN amplifié. Les résultats de ces PCRs montrent que les échantillons non dilués amplifiés à 53°C donnent des bandes plus intenses par rapport aux échantillons dilués au 1/20ème. En revanche, cette différence ne s'observe pas pour les PCRs faites avec une température d'hybridation de 57°C (Figure 11). On remarque également que les bandes sont plus intenses pour les PCRs réalisées avec une température d'hybridation de 53°C (Figure 11). C'est pour cette raison que j'ai décidé d'utiliser une température d'hybridation de 53°C pour effectuer les quantifications avec la ddPCR.

Tableau 5 : Liste des échantillons utilisés lors des PCRs de test des amorces H3 pour la ddPCR.

Sample	**Family**	***Genus***	***Species***
IM-2019-4002	Mangeliidae		
IM-2019-4015	Conidae	*Conus*	*ventricosus*
IM-2019-4031	Drilliidae	*Crassopleura*	*maravignae*
IM-2019-5356	Marginellidae	*Gibberula*	
IM-2019-4974	Buccinidae	*Chaevetia*	*recondita*
IM-2019-5637	Costellariidae		
IM-2019-5653	Mitridae		
IM-2019-4991	Muricidae	*Pagodula*	*echinata*

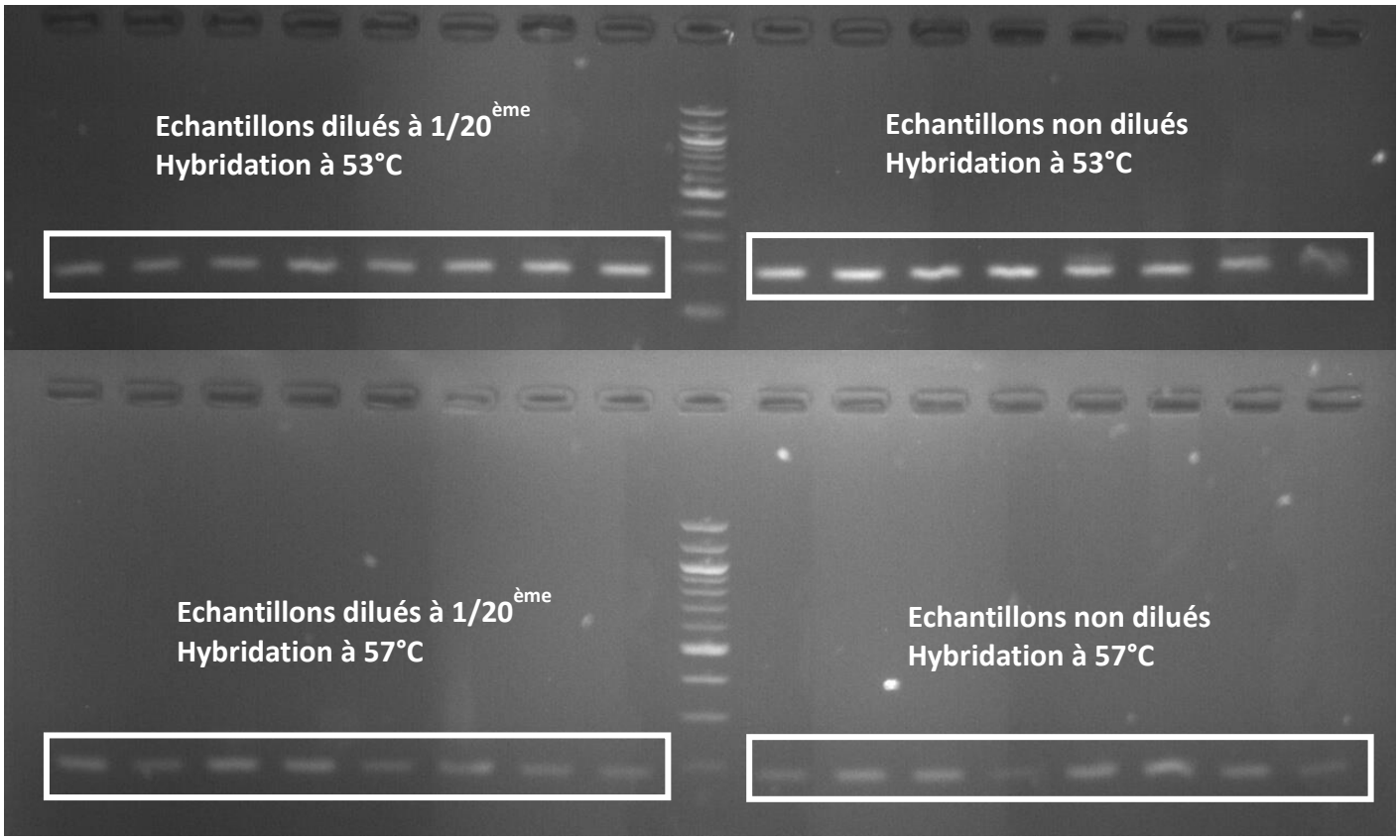


Figure 11 : Photographies des gels de PCRs.

Elles ont été réalisées avec des températures d'hybridation de 53°C (haut) et 57°C (bas). Sont encadrés en blanc à gauche les 8 échantillons qui ont été dilués au 1/20ème et à droite les échantillons non dilués.

Ensuite, j'ai vérifié la saturation de la quantification en ddPCR. Si l'ADN de l'échantillon est trop concentré avant la PCR, la limite de détection de l'appareil sera rapidement dépassée. En effet, la PCR va amplifier de manière exponentielle l'ADN. Or, la ddPCR se base sur la détection de la fluorescence émise par chaque fragment d'ADN amplifié dans l'ensemble des 20000 bulles générées. Il y aura donc une limite à la détection de cette fluorescence : la détection sera saturée et les valeurs renvoyées seront hors des seuils de détection, et donc les quantifications seront erronées. Pour tester cet effet de saturation, j'ai utilisé un échantillon que j'ai dilué 7 fois pour avoir des dilutions allant de 1/10ème à 1/10000000ème. On peut voir sur le graphique de la figure 12 que la quantification est saturée à la dilution de 1/10ème et quand l'échantillon n'est pas dilué. Ainsi, nous avons décidé pour la suite de diluer nos échantillons au 1/1000ème. J'ai donc dilué les 12 échantillons déjà utilisés pour les quantifications réalisées au Qubit et au Fragment Analyzer au 1/1000ème avant de faire la quantification en ddPCR. Le

résultat est exprimé en copies/µL, en prenant une taille de génome de 4 Gb, qui est une taille moyenne pour les génomes de néogastéropodes. Cette valeur est convertie en une concentration en ng/µL. Les standards 50 ng/µL et 100 ng/µL ont totalement saturé l'appareil de mesure (Tableau 4), dû au fait que je n'ai pas dilué ces standards comme les échantillons. Il serait nécessaire de les diluer pour s'assurer de la bonne quantification de ces standards.

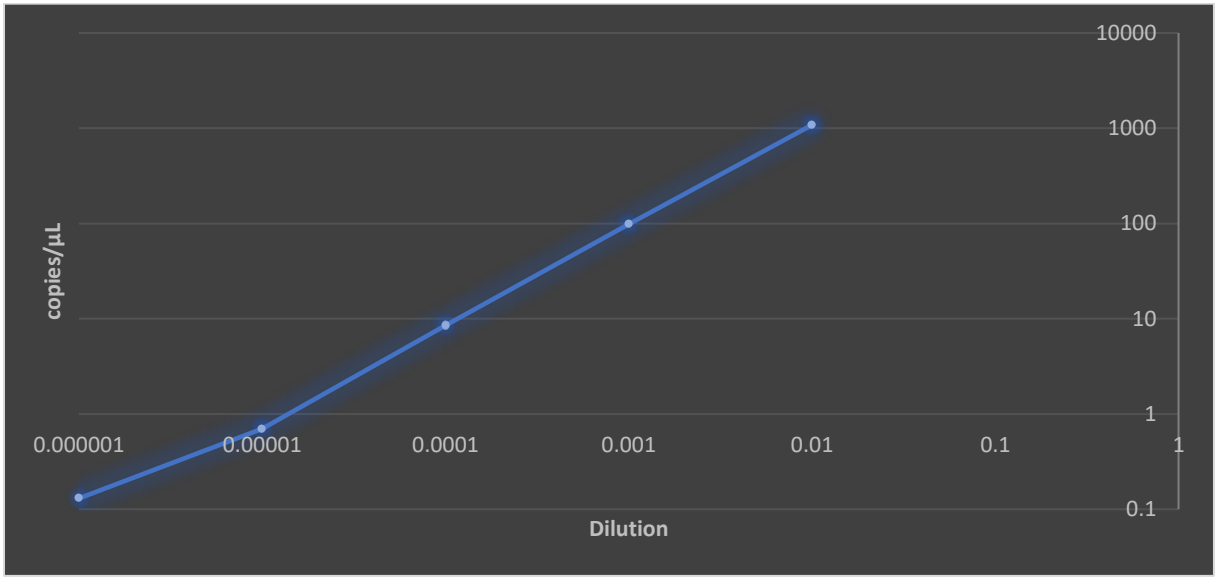


Figure 12 : **Courbe de la dilution en fonction du nombre de copies par µL**.

Par ailleurs, le standard à 0 ng/µL d'ADN est bien à zéro lorsqu'il est quantifié en ddPCR (Figure 13). De manière générale, les mesures semblent plus répétables que pour le Qubit ou le Fragment Analyzer : seuls 2 échantillons sur les 12 ont des valeurs clairement différentes entre les 3 mesures. Toutes ces quantifications restent compliquées à interpréter du fait que les standards n'ont pas toujours fonctionné, et il serait nécessaire de poursuivre les expériences pour valider l'utilisation de la ddPCR pour faire des quantifications.

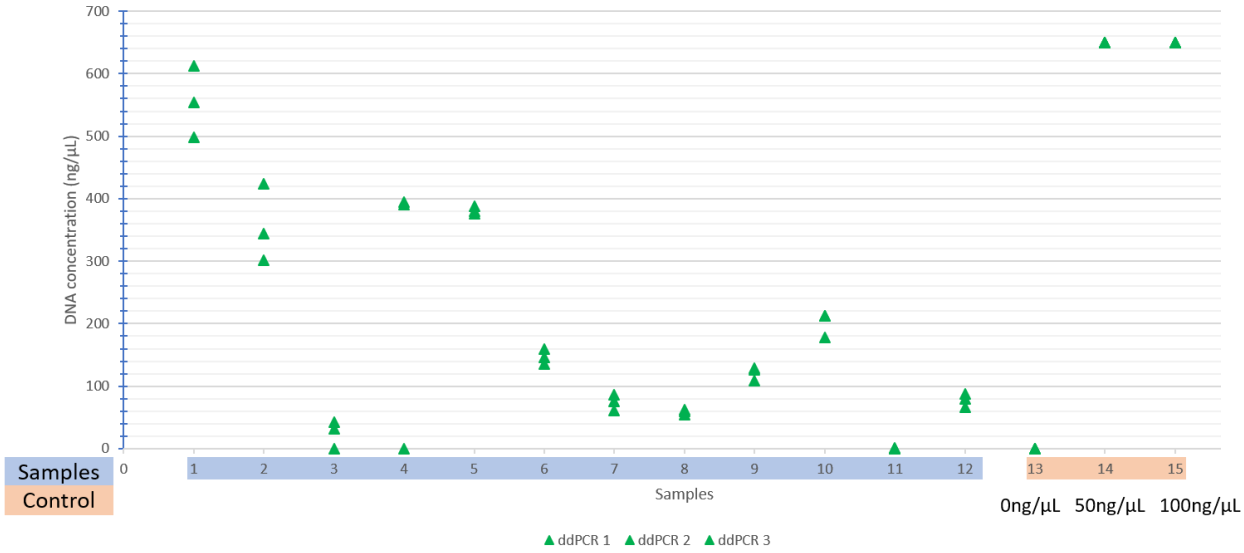


Figure 13 : Graphique représentant les résultats des quantifications réalisées avec la droplet digital PCR, les différents échantillons sont représentés en abscisse et les concentrations mesurées (ng/µL) en ordonée. *Dans le cadre bleu sont représentés les 12 échantillons utilisés pour lesquels nous avons réalisés 3 réplicats pour chacun. Dans le cadre orange sont représentés les 3 échantillons contrôles que nous avons utilisés et dont nous connaissions les concentrations en ADN : 0 ng/µL, 50 ng/µL et 100ng/µL.*

2.3. CHOIX DE LA METHODE DE QUANTIFICATION

Malgré des meilleurs résultats, mais qui restent à confirmer, avec la ddPCR, les résultats de ces quantifications ont permis de confirmer le choix fait précédemment sur l'utilisation du Qubit pour quantifier nos ADNs. En effet, c'est la seule méthode sur les trois qui nous a permis de conclure réellement sur la qualité potentielle des ADNs que nous avons quantifiés. Les quantifications erronées des standards positifs (50 ng/µL et 100 ng/µL) avec le Fragment Analyzer et la ddPCR ne nous permettent pas d'avoir la confiance nécessaire dans ces quantifications. De plus, le prix à l'échantillon pour ces différentes méthodes est inégal. Une quantification au Qubit coûte 0,13 euros par échantillon, contre 4,25 euros pour le Fragment Analyzer et 4 euros pour la ddPCR. Ce sont tous ces arguments qui justifient le choix d'utiliser le Qubit pour la quantification des ADNs pour les échantillons qui seront intégrés dans la capture d'exons (voir la suite de ce chapitre).

Ces tests nous ont par ailleurs permis d'adopter un protocole de quantification précis pour tous nos échantillons. Vu leur grand nombre et la rareté de certains ADN, il ne sera pas possible de faire des triplicats pour tous nos échantillons lors des quantifications pour la capture d'exons. En revanche, le fait que les échantillons qui n'ont pas d'ADN semblent bien quantifiés par la méthode Qubit permet de proposer un seuil de 5 ng/µL, c'est-à-dire que les échantillons qui ont une quantification en Qubit inférieure à 5 ng/µL ne seront sélectionnés que s'il n'y a aucun autre échantillon de disponible pour le taxon en question. Autrement dit, la rareté taxonomique d'un échantillon donné prime sur la concentration d'ADN quantifiée.

Au final, l'ensemble des ADNs utilisés pour la phylogénie ont été extraits à partir de tissus disponibles en collection, qu'ils aient été conservés spécifiquement pour du séquençage ADN ou non. Les échantillons d'ADN ont été extraits avant ma thèse ou dans le courant de ma thèse par Dario Zuccon en utilisant une méthode d'extraction sur colonne semi-automatisée à l'aide d'un robot d'extraction. J'ai ensuite quantifié ces ADNs extraits à l'aide de la méthode Qubit décrite précédemment afin de sélectionner les échantillons avec les meilleures concentrations en ADN possibles pour l'approche de capture d'exons.

3. STRATEGIE D'ECHANTILLONNAGE POUR LA CAPTURE D'EXONS

Afin de maximiser la diversité taxonomique échantillonnée, une stratégie globale en quatre étapes a été mise en place : (i) les spécimens sont collectés sur le terrain avec un protocole de préservation pour les étapes de biologie moléculaire ; (ii) puis le séquençage d'un fragment du gène *cox1* est effectué afin (iii) de proposer des identifications taxonomiques, qui se font avec l'aide des taxonomistes (ateliers d'identification, plate-forme « Hyperdiverse ») ; (iv) enfin, pour compléter l'échantillonnage disponible au MNHN, des spécimens d'autres muséums ont été demandés en prêt. Pour des raisons pratiques, à savoir un nombre de spécimens séquencés en capture d'exons limités pour des raisons de coûts mais aussi de temps d'analyse, nous avons décidé de nous focaliser sur le rang générique : c'est-à-dire que nous allons séquencer au maximum un spécimen par genre de néogastéropodes (avec quelques exceptions détaillées plus loin), pour établir les relations phylogénétiques entre genres et rangs taxonomiques supérieurs.

(i) Une collecte de terrain intensive grâce aux expéditions du MNHN des programmes Tropical Deep Sea Benthos et la Planète Revisitée (expeditions.mnhn.fr), (Bouchet et al., 2008). Cela représente au minimum une campagne de collecte par an (hors pandémie), avec des

échantillonnages de spécimens côtiers et profonds. Pour les collectes de spécimens côtiers, plusieurs techniques sont utilisées (brossage pour les fonds durs, aspirateur pour les fonds plus sablonneux, récolte à vue, dragage et chalutage en milieu plus profond), de manière à maximiser la diversité des habitats prospectés et la diversité des organismes collectés. Les méthodes employées permettent en particulier de collecter des spécimens de petite taille, de l'ordre de quelques millimètres (commun chez les néogastéropodes, comme les marginelles ou certaines familles de Conoidea), et qui peuvent représenter la majorité des espèces inconnues (Bouchet et al., 2009). En milieu profond, les spécimens sont collectés par dragage et chalutage.

Une équipe dédiée sur le terrain est ensuite chargée de trier le matériel, en particulier les résidus, et fixe dans de bonnes conditions les spécimens pour les extractions d'ADN : les spécimens sont extraits de leur coquille à l'aide d'un four micro-onde (Galindo et al., 2014), puis fixés en alcool 95-100°. Une documentation rigoureuse des données de chaque spécimen est également réalisée, comme la photographie des spécimens vivants, associée aux données de collectes (coordonnées, date, etc…). Toutes ces données sont sauvegardées dans les bases de données du MNHN (Basexp, INVMAR), et sont donc accessibles par la suite.

(ii) Tous les échantillons de néogastéropodes collectés sont barcodés, c'est-à-dire séquencés pour le fragment « barcode » du gène *cox1*. Cette étape est importante pour faire un tri préalable des différents spécimens et leur importance taxonomique. En effet, l'identification morphologique de nombreuses espèces de néogastéropodes est difficile sur la base seule de la forme de la coquille. Il existe ainsi de nombreux exemples de lignées cryptiques, y compris au niveau familial (Y. I. Kantor & Puillandre, 2021). Dans le cadre du projet HYPERDIVERSE, la réduction du coût du séquençage a été permise grâce à une approche par double indexage et séquençage illumina (Shokralla et al., 2014). Cela offre l'avantage d'augmenter la quantité d'échantillons séquencés en même temps et donc réduire le coût du séquençage de chaque échantillon. Avant le démarrage du projet HYPERDIVERSE, environ 18000 barcodes *cox1* issus de néogastéropodes du MNHN étaient disponibles, et environ 12000 séquences additionnelles ont été produites dans le cadre du projet par Dario Zuccon. Les séquences *cox1* ainsi obtenues nous permettent d'identifier parmi tous les spécimens collectés ceux qui potentiellement constituent des lignées profondes, de rang au moins générique, et qui pourront être intégrées dans le jeu de données de la capture d'exons.

(iii) L'identification des différents spécimens par des taxonomistes lors de sessions de travail dédiées sur plusieurs jours. Ces sessions sont essentielles afin d'avoir l'opportunité de faire

venir des spécialistes du monde entier pour favoriser les échanges entre eux et optimiser au maximum le nombre d'échantillons identifiés sur des courtes périodes de temps. De plus, un autre outil important a été mis en place avec le projet HYPERDIVERSE, la plate-forme « Hyperdiverse » (https://www.hyperdiverseproject.com/platform) : il s'agit d'un site internet qui donne accès à tous les spécimens de néogastéropodes étudiés dans le cadre du projet, et auquel les taxonomistes peuvent accéder et proposer des identifications taxonomiques. Cet outil semble être bien reçu par la communauté des taxonomistes, avec 25 utilisateurs actifs, qui ont proposé 12315 identifications. La difficulté de cette étape est de trouver les taxonomistes pour chaque groupe. En effet, la nécessité de former de nouveaux taxonomistes est un enjeu crucial car la récolte de spécimens à grande échelle sans une identification des spécimens est vaine. L'organisation d'ateliers d'identification taxonomique, couplée à l'utilisation de la plate-forme Hyperdiverse, a déjà permis d'identifier un grand nombre de néogastéropodes du MNHN. Cette identification est de plus souvent guidée par les hypothèses taxonomiques proposées sur la base des barcodes *cox1* produits pour ces mêmes spécimens.

Cette stratégie m'a permis d'inclure dans le jeu de données pour la phylogénie un grand nombre de genres décrits pour lesquels des spécimens préservés spécifiquement pour le séquençage étaient disponibles au MNHN. De plus, la stratégie de séquençage massif du gène *cox1* a permis également d'inclure dans le jeu de données des lignées non décrites, mais potentiellement de rang générique (ou supérieur). En effet, 20% des familles de néogastéropodes et 10% des genres ont été décrits au cours des 10 dernières années (WoRMS Editorial Board, 2024), ce qui laissé présager un grand nombre de taxa encore non décrits.

Cependant, un autre problème s'ajoute à celui des taxa non décrits : la majorité des taxa a été décrite sur la base de coquilles vides, dont beaucoup n'ont jamais été collectés vivants (Bouchet & Strong, 2010). Cela représente de nombreux genres sans échantillons d'ADN exploitables dans les collections du MNHN.

(iv) Au sein des collections du muséum nous avons des genres dont aucun spécimen n'a été conservé dans de bonnes conditions en vue de séquençage ADN ou aucun spécimen collecté vivant. Pour les spécimens collectés à l'état de coquille vide, l'extraction d'ADN à partir des coquilles des spécimens de collection pourrait être une solution (voir partie 2.1). Pour d'autres taxa, nous nous sommes tournés vers d'autres muséums, bien qu'il soit parfois compliqué d'obtenir des spécimens issus de collectes faites par d'autres muséums dans le monde. En effet, certains permis de collecte obtenus sous l'égide du protocole de Nagoya interdisent le partage

des spécimens collectés. Cela est un frein au partage des spécimens et donc au partage des connaissances et à l'inclusion de spécimens taxonomiquement importants pour notre échantillonnage.

Cette stratégie d'échantillonnage a été présentée sur un poster lors du World Congress of Malacology en août 2022 à Munich, représenté sur la page suivante.

HYPERDIVERSE
www.hyperdiverseproject.com

Sampling the known and unknown diversity in hyperdiverse groups for molecular phylogeny

Nicolas Puillandre[1], Thomas Lemarcis[1], Claudia Ratti[1], Alexander E. Fedosov[1,2], Yuri I. Kantor[1,2], Maria Vittoria Modica[3], Marco Oliverio[4], Philippe Bouchet[1]

[1]Institut Systématique Evolution Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, Paris, France. puillandre@mnhn.fr
[2]A.N. Severtzov Institute of Ecology and Evolution, Russian Academy of Sciences, Moscow, Russia.
[3] Dept. of Biology and Evolution of Marine Organisms, Stazione Zoologica Anton Dohrn, Roma, Italy
4 Dept. Biology and Biotechnologies "Charles Darwin" (Zoology), Sapienza Rome University, Roma, Italy

CONTEXT

With recent advances in DNA sequencing and bioinformatics, producing genomic data and reconstructing trees is no longer a limiting factor. Instead, it is to exhaustively sample the known and unknown diversity. This is particularly true in hyperdiverse and poorly known taxa such as the **neogastropods**, with more than 15,000 described species: though phylogenetic relationships at the deeper taxonomic levels are about to be resolved, only a fraction of the diversity has been sequenced so far.

OBJECTIVE

As part of the ERC "HYPERDIVERSE" project:
a) to cox-1 sequence c. **5000 species**
b) produce an exon-capture-based phylogeny of neogastropods, including c. **1000 genera**.

5-STEP STRATEGY, each associated to particular challenges, to maximize taxonomic representativity


TROPICAL DEEP SEA BENTHOS
la planète envisitée

MNHN expeditions from The Planet Reviewed and Tropical Deep Sea Benthos programs

1. INTENSIVE FIELD SAMPLING in both shallow and deep-sea waters, with a focus on small-sized molluscs, using collecting gear (trawls, dredges, brushing baskets, suction pumps…) for bulk sampling. Dedicated teams in the field sort the material (in particular the small fractions), fix it in good condition for DNA sequencing (= 'DNA' samples) and document the associated data (e.g. pictures of live animals).

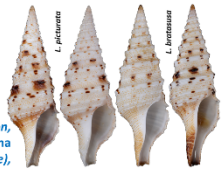
Challenge #1: increasingly difficult to conduct fieldwork, due to the need to comply with the ABS principles formalized in the Nagoya protocol

Challenge #2: efficiently sampling minute species (constituting the majority of unknown species)

2. BARCODING all MNHN neogastropod DNA samples. Most (see step 3) samples are cox-1 sequenced to avoid biased morphology-based pre-sorting: many neogastropods groups have proven to be difficult to tackle with shells only, with many cryptic lineages, even at the family level.

Challenge #3: reduce cost of sequencing with dual indexing and Illumina seq. (Shokralla et al 2015) – 40,000 samples to process!

Striking shell resemblance of Leucosyrinx *and* Sibogasyrinx, *considered as synonyms until recently*



Cryptic species are common, e.g. within Lophiotoma *(Conoidea, Turridae),*


The columbellid (Nov. 2021) and mitriform (July 2022) workshops held at the MNHN, Paris. In total, 6,000+ samples identified.

3. IDENTIFICATION OF THE VOUCHERS by taxonomists during dedicated workshops (eventually with step 2 to subsample abundant species) and using a **web platform**, to check photographed vouchers, comment and propose identifications.

Challenge #4: the taxonomic impediment, with only a few, if any, taxonomists available for some groups – we need to train and recruit taxonomists!

25,000 vouchers photographed

WEB PLATFORM
www.hyperdiverseproject.com/platform

SCAN ME

20% of the families and 10% of the genera have been described during the last 10 years (WoRMS) → **Many genera remain to be described**

Most taxa described based on empty shells, and many never collected alive → **Many genera without any DNA sample**

Consequently, genus-level taxa can be either: DESCRIBED or UNDESCRIBED SAMPLED (for DNA) or UNSAMPLED

Challenge#5: How to include UNDESCRIBED or UNSAMPLED genus-level taxa?

4. SELECTION OF DNA SAMPLES:

- one per described genus = **DESCRIBED** and **SAMPLED** genera
- one per genus-level lineage revealed by the cox-1 tree = **UNDESCRIBED** and **SAMPLED** genera

Ideally, the selected sample is well-preserved, with an intact shell and collected recently.

5. SEARCH FOR THE MISSING GENERA in museum collections, often containing rare taxa never preserved for DNA sequencing or even collected alive = **DESCRIBED** and **UNSAMPLED** genera

Taxonomists can help decide which nominal genera likely constitute synonyms and thus not worth spending effort to find and sequence.

Challenge#6: extract DNA from old, poorly-preserved samples or empty shells – the exon-capture approach can circumvent this issue.

Challenge#7: The Nagoya protocol, again… Permits issued for field sampling do not always allow sharing samples with other research teams!

CONCLUSION

With this strategy, only **UNDESCRIBED** and **UNSAMPLED** genus-level lineages would remain uncovered, still a potentially non-negligible fraction of the total diversity in such a hyperdiverse group.

→ Exome sequencing of ca 1000 genera → Tree reconstruction → Thomas Lemarcis' poster

Shokralla, S., Porter, T. M., Gibson, J. F., Dobosz, R., Janzen, D. H., Hallwachs, W., ... & Hajibabaei, M. (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific reports*, *5*(1), 1-7.

This action has received funding from the European Union's Horizon 2020 research and innovation program & European Research Council under grant agreement ERC HYPERDIVERSE (GRANT AGREEMENT 865101)

3.1. SELECTION DES ECHANTILLONS ET SEQUENÇAGE EN 3 BATCHS

Pour tous ces échantillons, le scénario idéal est d'avoir la possibilité de sélectionner des échantillons bien préservés, dont la coquille est intacte et qui ont été collectés récemment. Cela maximisera les chances d'extraire de l'ADN en quantité suffisante pour toutes les expérimentations de biologie moléculaire. De plus, la coquille intacte est nécessaire pour compléter les analyses moléculaires par des identifications morphologiques. La stratégie en quatre étapes mise en place permet de couvrir une grande partie de la diversité des néogastéropodes, et seules les lignées de niveau générique non décrites et non échantillonnées ne sont pas couvertes par notre échantillonnage. Cela nous permet d'envisager un séquençage à grande échelle des genres de néogastéropodes par la méthode de la capture d'exons.

Pour chaque genre que nous souhaitons séquencer, nous avons sélectionné entre 1 et 3 spécimens (selon le nombre de spécimens disponibles). Sont sélectionnés en priorité des spécimens d'espèces-types pour les genres décrits. Priorité est également donnée aux spécimens les plus récents et qui ont été préservés dans les meilleures conditions pour conserver l'ADN (conservation des spécimens dans l'éthanol). Ce sont ces spécimens qui ont été quantifiés et ceux avec les plus hautes concentrations en ADN (un par genre) ont été sélectionnés pour être intégrés dans le jeu de données de capture d'exons. La liste complète des échantillons sélectionnés pour la capture d'exons (1728) est disponible en Annexe 2.

Le jeu de données de capture d'exons a été séparé en trois batchs. En effet, afin de maximiser les chances de capture du plus grand nombre de spécimens et d'exons possible, il est important que la distance génétique ne soit pas trop importante entre les sondes de capture et nos échantillons, comme expliqué précédemment (voir Chapitre 2, Partie 1). Le premier batch (Batch 1) est composé des super-familles des Conoidea, Mitroidea et Buccinoidea, pour lesquelles les sondes ont été définies à partir des transcriptomes issus de ces trois groupes (voir Chapitre 2, Partie 1). Ce sont trois super-familles qui sont *a priori* proches dans les phylogénies déjà réalisées chez les néogastéropodes (Osca et al., 2015). Le batch 1 est composé de 384 échantillons (voir Annexe 2). Dans ce batch 1, nous avons 24 réplicats intra-spécimen : pour 8 échantillons, l'ADN a été séparé en trois aliquots pour tester la répétabilité de l'expérimentation de préparation des banques de séquençage et de capture.

Le second batch de séquençage (Batch 2) est composé de toutes les autres super-familles de néogastéropodes : les Muricoidea, les Olivoidea, les Turbinelloidea, les Volutoidea, ainsi que

les Tonnoidea et Ficoidea et quelques groupes externes. Nous avons également intégré à ce batch des échantillons de taxa inclus dans le batch 1, qui n'avaient soit pas fonctionné lors du premier séquençage, soit avaient été collectés et/ou extraits entre les batchs 1 et 2. Le batch 2 est également composé de 384 échantillons (Annexe 2). Dans ce batch, nous avons 11 réplicats intra-spécimen ; comme pour le batch 1, nous avons séparé l'ADN de 3 échantillons en 3 ou 4 aliquots, toujours pour tester la répétabilité des expérimentations pré-séquençage. Dans ce batch 2, nous avons intégré 18 échantillons de groupes externes (Annexe 2). Ces spécimens ont été choisis parmi les groupes les plus proches phylogénétiquement des néogastéropodes (Cunha et al., 2009; Osca et al., 2015; Q. Wang et al., 2021; Zou et al., 2011).

Enfin, le dernier batch de séquençage (Batch 3) regroupe les super-familles des batchs 1 et 2, ainsi qu'un nombre plus important de groupes externes, représentatifs de la diversité des Caenogastropoda non-Neogastropoda. Nous avons également inclus des échantillons qui n'avaient pas fonctionné lors des séquençages des batchs 1 et 2, ainsi que les échantillons qui ont été obtenus après le séquençage des deux premiers batchs et qui appartiennent aux groupes taxonomiques dont les spécimens ont été séquencés dans les deux premiers batchs. Le batch 3 inclut en particulier un nombre plus important de spécimens « anciens », c'est-à-dire qui n'avaient pas été spécifiquement conservés pour des analyses d'ADN, de manière à compléter l'échantillonnage taxonomique des batchs 1 et 2. Le batch 3 est composé de 960 échantillons. Cela inclut un grand nombre de spécimens sélectionnés pour reconstruire la phylogénie des néogastéropodes, mais également des échantillons pour reconstruire une phylogénie plus détaillée des Raphitomidae (Chapitre 4) : ainsi, un grand nombre de Raphitomidae, avec plusieurs spécimens par genres quand cela était possible, ont été inclus dans le batch 3. L'objectif était notamment de tester le pouvoir résolutif de nos exons à différentes échelles taxonomiques (Neogastropoda et Raphitomidae). De plus, le batch 3 inclut également de nombreux spécimens pour des projets annexes. En effet, les résultats concluants des deux premiers batchs de capture d'exons, à savoir un taux de capture par échantillon élevé, nous ont permis d'inclure dans le dernier batch plus de spécimens correspondants à des groupes d'intérêts spécifiques pour nos collaborateurs. Cependant, la stratégie d'échantillonnage, de sélection d'un spécimen par genre, que j'ai présentée plus haut dans ce chapitre, n'a pas été appliquée pour les échantillons qui ont été ajoutés dans le dernier batch de séquençage par les différents collaborateurs de l'équipe. En raison de questions de recherche différentes de la nôtre, ils avaient besoin d'échantillonner parfois plusieurs fois le même genre pour déterminer des relations phylogénétiques au sein de certains genres. C'est le cas de la super-famille des

Velutinoidea, groupe sur lequel Giulia Fassio (Université de Rome) travaille, et de la famille des Ovulidae sur laquelle Elisa Nocella (Université de Rome) effectue sa thèse. Dans ce dernier batch nous avons également ajouté plus de représentants des familles des Cancellariidae et des Colubrariidae, dont notre collaboratrice Maria Vittoria Modica (Stazione Zoologica Anton Dohrn) est la spécialiste. Nous avons fait de même pour les Mitridae et les Costellariidae, deux familles d'intérêt pour Alexander Fedosov (Swedish Museum of Natural History). Enfin, nous avons mis en place un protocole afin d'évaluer l'impact de la distance phylogénétique sur le succès de la capture d'exons (chapitre 3). Ainsi, dans le batch 3, 162 échantillons sur les 960 échantillons totaux sont des réplicats intraspécifiques de spécimens différents : pour 30 espèces, 5 spécimens différents ont été séquencés (et pour 4 spécimens, des réplicats intra-spécimens ont été réalisés).

Sur les 1728 échantillons des 3 batchs de séquençage, nous avons un total de 16 échantillons qui ont été séquencés deux fois car ils n'avaient pas fonctionné lors du séquençage des batchs 1 ou 2.

Notre échantillonnage couvre ainsi une grande partie de la diversité des néogastéropodes. Notre jeu de données est composé de 9 super-familles (soit toutes les superfamilles actuellement considérées comme valides), 70 familles (soit toutes les familles actuellement considérées comme valides) et 774 genres (soit 62% des 1255 genres actuellement considérés comme valides). Nous avons également 110 lignées de rangs génériques qui, à l'heure actuelle, ne sont pas décrites et constituent potentiellement des genres nouveaux, en particulier pour les Raphitomidae (cf tableau global pour les détails). L'échantillonnage inclut également 18 super-familles, incluant 33 familles, de groupes externes proches des Neogastropoda parmi les Caenogastropoda.

3.2. QUANTIFICATION DES ADN ET PREPARATION DES BANQUES

3.2.1. QUANTIFICATION DES ADN

Pour le batch 1 j'ai quantifié dans un premier temps 384 échantillons qui étaient les premiers choix pour chaque genre d'intérêt défini dans notre stratégie d'échantillonnage. Si le résultat de la quantification d'ADN n'était pas satisfaisant, à savoir des concentrations de moins de 5

ng/µL, j'utilisais le deuxième choix s'il y en avait un. De même si ce deuxième choix n'avait pas une concentration suffisante, je prenais le troisième choix, si disponible. J'ai quantifié au total 469 échantillons afin de choisir les 384 meilleurs échantillons pour la capture d'exons. Dans le cas où les trois choix n'étaient pas satisfaisants, j'ai sélectionné l'échantillon avec la plus forte concentration en ADN. Dans les cas où un seul choix était disponible pour un genre donné j'ai sélectionné l'échantillon même si sa concentration en ADN n'était pas satisfaisante, étant donné son importance taxonomique.

J'ai utilisé le même protocole pour le deuxième batch de capture d'exons. J'ai quantifié au total 539 échantillons. J'ai également sélectionné les 384 meilleurs échantillons pour la capture d'exons.

Enfin pour le troisième batch de séquençage, les 232 échantillons de nos collaborateurs étaient déjà quantifiés. En utilisant le même protocole que pour les deux premiers batchs, j'ai quantifié 804 échantillons au total. Parmi ces échantillons, j'en ai sélectionné 728 pour le séquençage.

Les concentrations d'ADN quantifiées étaient hétérogènes avec des concentrations allant de 0 ng/µL à plusieurs centaines de ng/µL. Afin d'obtenir un nombre de reads de séquençage le plus homogène possible, j'ai dû diluer tous mes échantillons pour avoir des concentrations équimolaires. Le protocole de préparation de banque utilisé que je décrirai ci-dessous indique une quantité spécifique d'ADN au début du protocole : il est nécessaire d'avoir 100 ng d'ADN dans un volume final de 26 µL.

J'ai fait la dilution des 1496 échantillons (les 232 échantillons de nos collaborateurs étant déjà dilués) avant l'envoi pour la préparation de banques, la capture d'exons et le séquençage.

Au cours de la première année de ma thèse, nous avons dû faire un choix pour réaliser les expérimentations de préparation de banques et de capture d'exons. Pour le séquençage, nous savions que nous devrions sous-traiter cette partie car aucun séquenceur n'était disponible au sein du MNHN pour séquencer un si grand nombre d'échantillons, avec une profondeur de séquençage suffisante pour chacun de nos échantillons et un coût par échantillon raisonnable. De plus, nous avons dû rapidement faire face à des problèmes logistiques importants. En effet le laboratoire du Service de Systématique Moléculaire (SSM) du muséum était indisponible pour causes de travaux puis de déménagement pendant les 6 premiers mois de ma thèse, alors qu'à l'origine il était prévu que je réalise toutes les expérimentations de la quantification des ADNs jusqu'à la préparation des banques et à la capture d'exons dans ce laboratoire. Les

premiers mois de ma thèse, qui a débuté en Janvier 2021, ont également coïncidé avec des restrictions d'accès au lieu de travail en raison de la crise covid. Cela a ajouté un élément supplémentaire qui nous a fait changer notre approche pour la partie du travail de laboratoire pour la préparation des batchs de séquençage de la thèse. Nous nous sommes tournés vers l'Institut de Cerveau et de la Moelle (ICM) dans lequel une plateforme de séquençage est disponible. Nous avons pris contact avec Yannick Marie, Delphine Bouteiller et Agnès Rastetter pour discuter de la possibilité de réaliser l'intégralité des expérimentations : la préparation des banques, la capture des exons et le séquençage. Le protocole qu'ils nous ont proposé offre un avantage conséquent dans sa capacité de permettre de préparer des banques (Protocole NEBNext Ultra II FS DNA Library Prep Kit for Illumina) à partir de quantités d'ADN très faibles, entre 500 pg et 1 µg. De plus, ils nous ont proposé de séquencer plus d'échantillons à chaque batch afin de réduire le coût à l'échantillon.

J'ai donc envoyé les échantillons quantifiés et dilués à l'ICM afin qu'ils effectuent les préparations de banque, la capture et le séquençage des 3 batchs.

3.2.2. Fragmentation des ADN

Pour effectuer la préparation des banques nous avons convenu avec l'ICM d'utiliser le protocole NEBNext Ultra II FS DNA library Prep Kit for Illumina. C'est un protocole de préparation de banques avec fragmentation enzymatique pour des séquençages Illumina dont les adaptateurs sont compatibles pour faire de la capture d'exons avec le kit fourni par MyBaits. Ce protocole présente plusieurs avantages : (i) il permet de travailler avec des faibles quantités d'ADN en entrée, (ii) il est rapide à réaliser pour préparer les banques, entre 1 à 2 jours pour une personne expérimentée contre presque 15 jours pour le protocole de Meyer & Kirscher (Meyer & Kircher, 2010) qui était utilisé dans l'équipe pour les précédentes captures d'exons (Abdelkrim et al., 2018; Zaharias et al., in press). Une différence importante entre le protocole de NEB et celui de Meyer & Kirscher est aussi le type de fragmentation d'ADN utilisé. Pour le protocole de Meyer & Kirscher, c'est une fragmentation mécanique qui est faite par sonication. Pour le protocole NEB, c'est une fragmentation enzymatique. L'avantage de la fragmentation enzymatique par rapport à la fragmentation mécanique par sonication est le gain de temps ainsi que la réduction des risques de non fragmentation de nos échantillons. En effet, j'ai pu personnellement expérimenter la fragmentation mécanique lors d'un précédent contrat

d'ingénieur d'étude et j'ai pu constater qu'il était compliqué de réaliser la fragmentation de façon homogène sur tous les échantillons. Aucun élément préalable ne permet de savoir comment les ADNs totaux des échantillons vont se fragmenter, ce qui rend cette approche particulièrement compliquée et longue, avec potentiellement plusieurs allers-retours entre la machine de fragmentation et les gels de vérification pour visualiser le degré de fragmentation de l'ADN.

Sur les conseils d'Agnès Rastetter, nous avons effectué un test sur un pool de quelques échantillons. Nous avons sélectionné des échantillons qui ont été collectés à des périodes différentes et donc que l'on peut considérer comme ayant des « âges » différents, et potentiellement de qualité différente (Tableau 6). Pour effectuer ces tests de fragmentations enzymatiques, Agnès a utilisé le protocole KAPA Hypercap, Roche, différent du protocole NEB retenu au final (voir ci-dessous). La différence avec le protocole NEB est le fait que le protocole KAPA demande plus de volume de chaque échantillon en entrée, 35 µL contre 26 µL. Les quantités d'ADN obtenues après la fragmentation sont comprises entre 21,4 ng et 100 ng. Pour 6 des 10 échantillons utilisés nous avons 100 ng d'ADN après la fragmentation, qui est la quantité demandée par le protocole de préparation de banques pour maximiser les chances de produire une banque de séquençage viable. Deux échantillons ont des quantités de 80,8 ng et 59,2 ng d'ADN, qui sont des quantités qui permettent malgré tout de faire des banques suffisamment concentrées pour la capture d'exons qui suivra. Enfin, deux échantillons ont des quantités d'ADN trop faibles (21,4 ng et 33,1 ng d'ADN). *A priori* ce sont des quantités qui seraient insuffisantes pour préparer des banques de séquençage NGS. Cependant, si l'échantillon en question fait partie d'un taxon important pour compléter l'échantillonnage, alors nous l'utiliserons malgré tout pour la préparation des banques. Ces résultats de quantifications sont cohérents avec les gels de visualisation de l'ADN total fragmenté (Figure 14). Ici, la migration sur gel est directement faite après un seul tour de fragmentation, sans besoin de repasser plusieurs fois les échantillons (contrairement à l'approche mécanique). D'après ce gel, nous pouvons voir que 8 des 10 échantillons utilisés (numéros temporaires 1 à 8) se sont fragmentés correctement avec des quantités d'ADN suffisantes pour réaliser les préparations de banque. L'échantillon 9 sur la figure 14 semble avoir des quantités d'ADN trop faibles, ce qui confirme les résultats de quantification. L'échantillon 10 est le seul qui est indétectable, ce qui, comme pour l'échantillon 9, confirme la quantification d'ADN qui a été faite. Les échantillons 9 et 10 font partie des plus anciens échantillons collectés parmi tous ceux que nous avons utilisés pour ces tests. Néanmoins, l'échantillon 8, qui a également été collecté

lors de la même campagne de l'échantillon 9, a un front de migration de bonne qualité et une quantité d'ADN de 100 ng. Il semble donc exister un lien entre l'âge des échantillons et la capacité à récupérer de l'ADN en quantité suffisante, même si des exceptions existent, ce qui confirme l'intérêt d'inclure des spécimens même anciens dans le jeu de données.

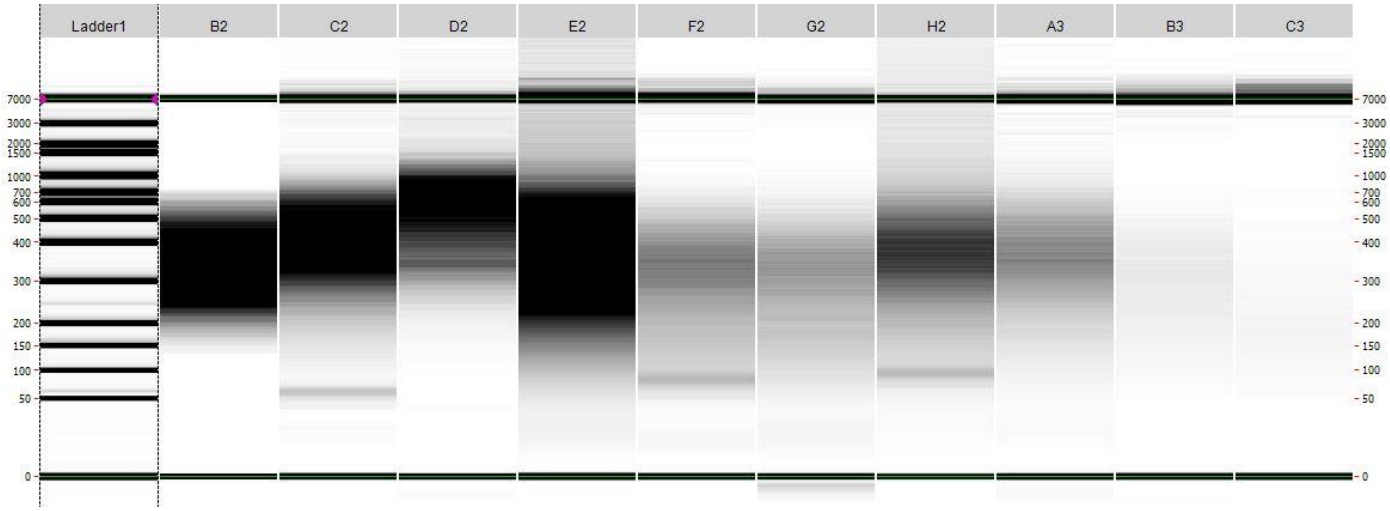


Figure 14 : Photographie du gel de migration des ADNs fragmentés.

À gauche de l'image l'échelle des tailles de fragments en paires de bases. Les intensités des bandes sont variables entre les échantillons. Les échantillons B2, C2, D2, E2 ont des bandes intenses qui sont comprises entre 200 et 700 pb. L'échantillon H2 a une bande intense mais réduite avec des tailles comprises entre 300 et 400 pb environ. Les échantillons F2, G2, A3 ont des bandes d'intensité moyenne. Les échantillons B3 et C3 ont des bandes d'intensité faible à quasiment indétectable.

Suite à ces résultats et aux discussions avec l'ICM, nous avons choisi d'utiliser le kit de préparation de banques NEBNext Ultra II FS. Bien que le kit KAPA de Roche a été utilisé pour effectuer les tests de fragmentation enzymatique, l'ICM ne possédait que 96 adaptateurs différents pour ce kit de préparation de banque. Or nous avions besoin de 196 adaptateurs différents afin de préparer des batchs de 384 échantillons au total, ce que permet le kit de préparation NEB. De plus, le volume nécessaire pour chaque échantillon étant plus faible par rapport au kit KAPA, à savoir 26 µL contre 35µL, nous avons décidé d'utiliser le kit NEB pour effectuer la préparation des banques.

Tableau 6 : Liste des échantillons utilisés pour réaliser les tests de fragmentation enzymatique.

N° temporaires	"Age"	Préparation	Profondeur	N°MNHN	Expédition	Genre	Espèce	Concentration ng/µL	260/280	260/230	DeNovix HS ng/µL	Volume ADN µL	Volume H2O µL	Quantité ng
1	récent	Micro-onde	Profond	IM_2013-48201	KANADEEP	*Spergo*	*fusiformis*	29.463	1.87	1.62	17.182	5.8	29.2	100
2	récent	Micro-onde	Profond	IM_2013-48234	KANADEEP	*Daphnella*	sp.	6.582	1.98	1.08	2.309	35.0	0.0	80.8
3	récent	Micro-onde	Côtier	IM_2019-6096	CORSICABENTHOS 2019	*Raphitoma*	sp.	8.467	2	1.41	1.692	35.0	0.0	59.2
4	moyen	Micro-onde	Profond	IM_2013-18924	PAPUA_NIUGINI	*Buccinaria*	sp.	10.51	1.93	1.6	4.405	22.7	12.3	100
5	moyen	Micro-onde	Profond	IM_2013-19168	PAPUA_NIUGINI	*Buccinaria*	sp.	22.403	1.86	1.79	12.109	8.3	26.7	100
6	moyen	Micro-onde	Profond	IM_2013-59614	ZhongSha_2015	*Spergo*	*sibogae*	21.153	1.92	1.87	11.008	9.1	25.9	100
7	moyen	Micro-onde	Profond	IM_2013-59620	ZhongSha_2015	*Spergo*	*sibogae*	12.042	1.92	1.85	5.641	17.7	17.3	100
8	ancien	Manuelle	Profond	IM_2013-57580	AURORA_2007	*Thatcheria*	*mirabilis*	9.601	1.81	1.21	7.62	13.1	21.9	100
9	ancien	Manuelle	Profond	IM_2013-57581	AURORA_2007	*Thatcheria*	*mirabilis*	5.89	1.64	0.69	0.945	35.0	0.0	33.1
10	ancien	Manuelle	Profond	IM_2013-57589	PANGLAO_2005	*Buccinaria*	*loochooensis*	4.048	2.18	1.1	0.61	35.0	0.0	21.4

3.2.3. PREPARATION DES BANQUES, CAPTURE D'EXONS ET SEQUENÇAGE

Dans un premier temps, les extrémités des fragments d'ADN sont réparées, puis les adaptateurs sont attachés aux extrémités de chaque fragment d'ADN. On effectue ensuite un nettoyage à l'aide de billes magnétiques des adaptateurs ainsi qu'une sélection de taille de nos fragments afin de conserver seulement les fragments aux alentours de 500 pb. Ensuite la PCR d'enrichissement de nos ADNs avec les adaptateurs est effectuée. La dernière étape est le nettoyage de nos ADNs amplifiés par PCR pour éliminer les petits fragments qui n'auraient pu être éliminés lors du nettoyage précédent. Les banques ainsi produites sont vérifiées pour s'assurer de la taille la plus homogène possible de chacune des banques. Le gel de la figure 15 représente les banques de 96 échantillons du batch 1 à la fin de la préparation des banques et avant la phase de capture d'exons. Nous pouvons voir sur ce gel que la majorité des échantillons ont des tailles de fragments comprises entre 350 et 700 pb. Il est important que les banques préparées soient dans cet intervalle car nous ciblons un séquençage de deux fois 150 pb. Ce que nous souhaitons avoir lors du séquençage, ce sont des fragments qui se recouvrent mutuellement sur de petites portions. En effet, si deux fragments sont superposés, il sera difficile pour un logiciel d'assemblage de différencier ces deux fragments. Les banques de certains échantillons ont des quantités très faibles d'ADNs, comme pour les échantillons D2, D3, G4, H5, C7 par exemple (Figure 15). Nous avons décidé de conserver ces échantillons malgré tout pour effectuer les étapes de capture d'exons, car ce sont potentiellement des échantillons de taxa rares dont nous avons besoin pour notre arbre phylogénétique final.

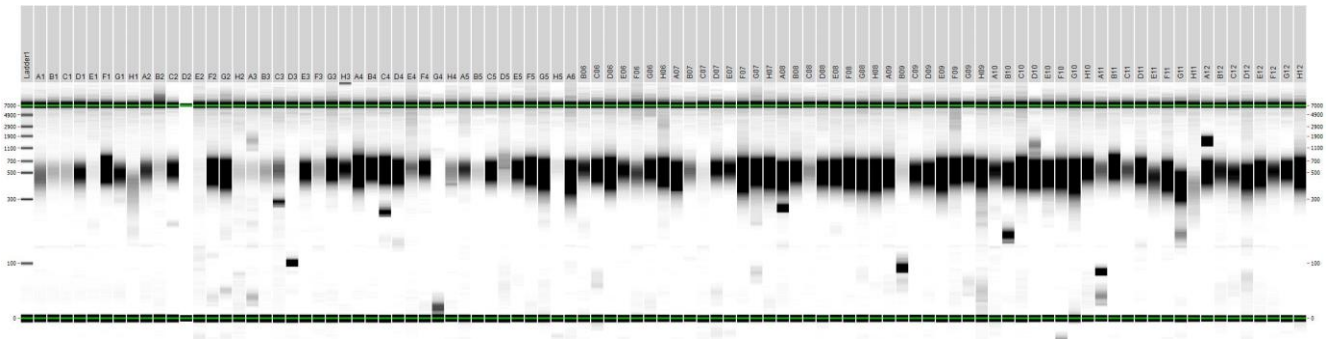


Figure 15 : Résultats de la préparation des banques de 96 échantillons du batch 1 avant la capture d'exons.

Le gel de la figure 16 représente les banques de 96 échantillons du batch 3 à la fin de la préparation des banques et avant la phase de capture d'exons. Nous pouvons voir sur cette figure que nous obtenons des résultats similaires à ceux du batch 1, c'est-à-dire des banques qui ont des fragments entre 350 et 700 pb. Nous avons également des échantillons dont les banques n'ont pas pu être amplifiées suffisamment comme les échantillons B7, A8, A9, D11, G11 et A12 par exemple. Nous avons également la présence de smears de grande taille sans avoir une concentration de fragments importante autour des valeurs de taille ciblée entre 350 et 700 pb. C'est le cas pour les échantillons D5 et E5 par exemple. Comme pour le batch 1, nous avons décidé de conserver tous ces échantillons pour effectuer la capture d'exons.

Le protocole de capture d'exons est fourni par l'entreprise MyBaits, Arbor Biosciences (MyBaits Hybridization Capture for targeted NGS v5.0), entreprise avec laquelle nous avons effectué les dernières étapes de synthèse de nos sondes de capture (Voir Chapitre 2, Partie 1). Les banques amplifiées avec leurs adaptateurs sont dénaturées et mises en présence de bloqueur d'adaptateurs pour empêcher toutes perturbations liées à l'interaction des sondes avec nos banques cibles. Les sondes biotinylées sont alors introduites et vont s'hybrider sur les zones cibles de nos banques (les exons cibles) pendant plusieurs heures d'incubation. Des billes magnétiques sont ensuite utilisées pour venir capturer les sondes et par conséquent, les fragments cibles sur lesquelles elles se sont hybridées. Ces fragments sont alors récupérés à l'aide d'un aimant et les banques non spécifiques sont pour la plupart éliminées. Les banques cibles sont alors amplifiées par PCR puis purifiées pour le séquençage.

Toutes les étapes que je viens de décrire (préparation de banques et capture d'exons) ont été réalisées à l'ICM par Delphine Bouteiller.

Le séquençage effectué à l'ICM est différent entre les deux premiers batchs et le troisième batch en raison de la quantité des échantillons que nous avons mis dans ce dernier batch. Pour les batchs 1 et 2 le séquençage a été réalisé sur NovaSeq, Ilumina, 6000 SP de 150 pb en paired-end, avec une couverture attendue de 2600X. Pour le batch 3, le séquençage s'est fait sur deux machines différentes, un premier avec le NovaSeq 6000 S1 de 150 b en paired-end comme pour les batchs 1 et 2 pour 771 échantillons, et un second avec le NexSeq 2000 P2 de 150 pb en paired-end également pour 189 échantillons.

Nous avons ensuite récupéré les reads de séquençage ainsi que les fichiers FastQC (Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) qui ont été

générés par l'ICM. Ces reads seront nettoyés et assemblés à l'aide du pipeline bioinformatique que je vais vous décrire dans la partie suivante.

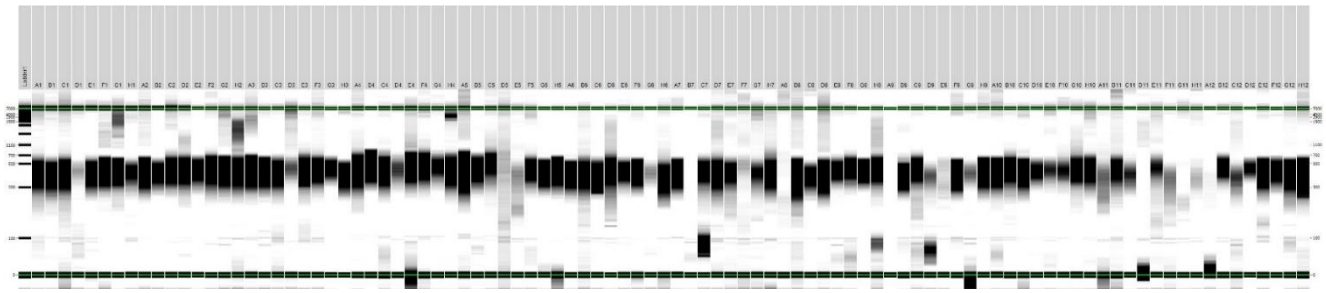


Figure 16 : Résultats de la préparation des banques de 96 échantillons du batch 3 avant la capture d'exons.

4. POST-SEQUENÇAGE : DES DONNEES BRUTES VERS LES EXONS.

4.1. DESCRIPTION DU PIPELINE

Pour les différents projets de captures d'exons déjà réalisés au sein de l'équipe, un pipeline bioinformatique avait été mis en place avant ma thèse. Il ne s'agit pas d'un pipeline en tant que tel car il ne consiste pas à lancer un seul script sur toutes les données en entrée avec un arbre phylogénétique en sortie. Néanmoins, les scripts rédigés se lancent les uns à la suite des autres, c'est pour cela que je vais faire référence à l'ensemble de ces différents scripts sous l'appellation pipeline dans la suite du manuscrit. Le traitement et l'analyse des données de séquençage se font en 14 étapes, du nettoyage des reads de séquençage jusqu'à la reconstruction des arbres phylogénétiques (Figure 17).

Pour s'assurer de la qualité de nos données de séquençage, et en particulier identifier les potentielles contaminations au sein de nos échantillons et estimer le nombre d'exons qui avaient été capturés pour chaque échantillon, le pipeline a été lancé de façon à obtenir rapidement un premier arbre phylogénétique, même préliminaire. Nous avons effectué cela pour chacun des trois batchs de séquençage qui ont été faits à des moments différents au cours de la thèse. Je vais dans un premier temps décrire les différents scripts qui constituent ce pipeline. Je

développerai ensuite les différentes améliorations que nous avons apportées au pipeline et les étapes que nous avons intégrées pour constituer un pipeline plus complet, afin de construire notre jeu de données final et réaliser les analyses phylogénétiques.

(i) Nettoyage préliminaire des séquences : La première étape de nettoyage des deux fragments séquencés (nommés ci-après R1 et R2) consiste à retirer les séquences des adaptateurs et enlever les extrémités des séquences de mauvaise qualité. Nous utilisons le logiciel Trimmomatic v0.39 (Bolger et al., 2014) pour effectuer plusieurs nettoyages qualitatifs sur nos données de séquençage. Avec l'option « ILLUMINACLIP », nous retirons les adaptateurs utilisés lors du séquençage. Avec l'option « SLIDINGWINDOW », nous allons parcourir chaque read de séquençage avec une fenêtre glissante : sur une série de 4 bases consécutives, si la qualité moyenne des bases considérées passe sous un seuil de score phred de 20, ces bases sont supprimées. Le score phred est un score de qualité attribué à chaque base séquencée. Pour un score phred de 20, nous avons 1 chance sur 100 que la base attribuée soit fausse : c'est le score minimum que nous avons choisi pour garantir une qualité suffisante à nos données de séquençage.

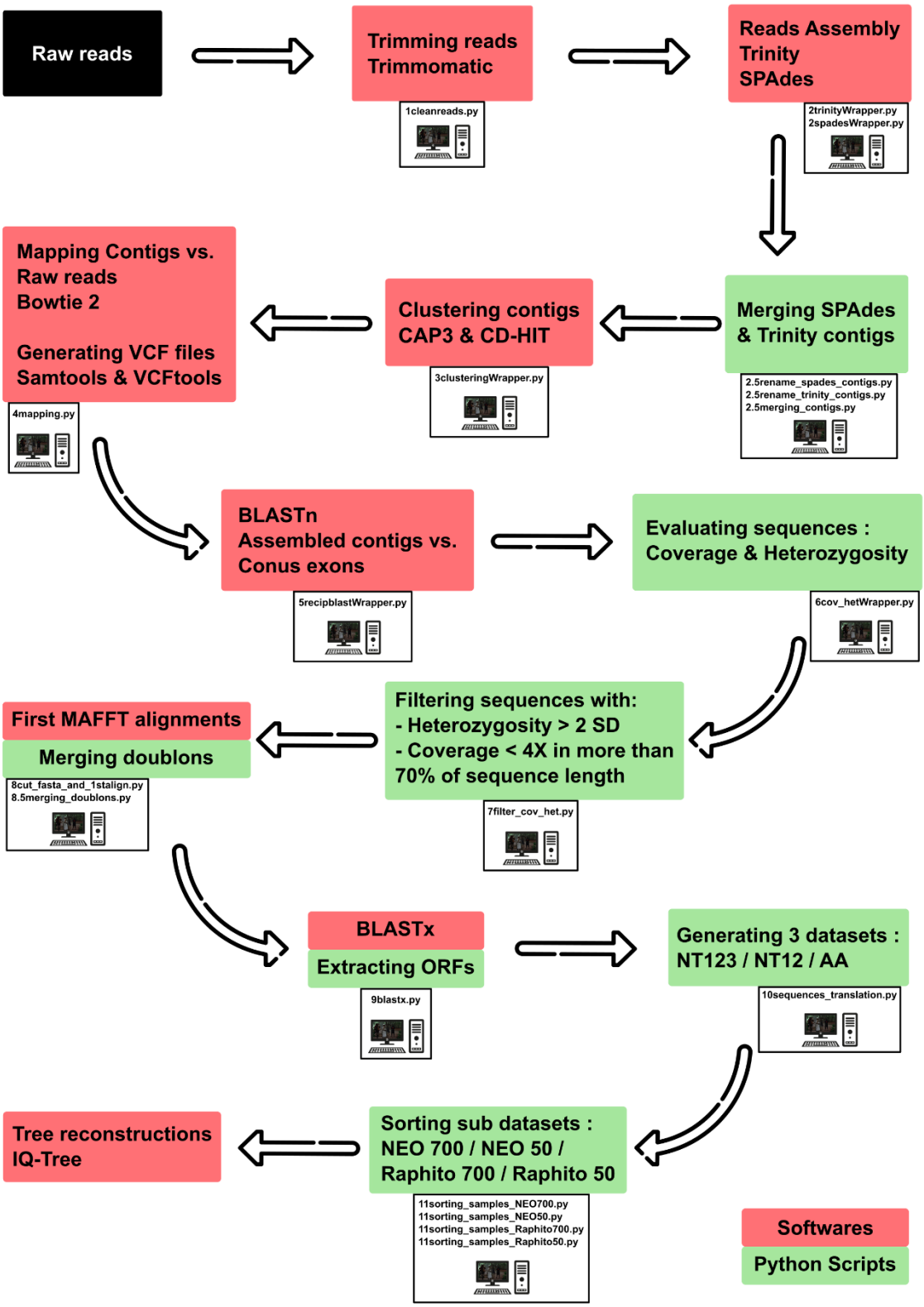

Raw reads
Trimming reads
Trimmomatic
1cleanreads.py
Reads Assembly
Trinity
SPAdes
2trinityWrapper.py
2spadesWrapper.py
Merging SPAdes
& Trinity contigs
2.5rename_spades_contigs.py
2.5rename_trinity_contigs.py
2.5merging_contigs.py
Clustering contigs
CAP3 & CD-HIT
3clusteringWrapper.py
Mapping Contigs vs.
Raw reads
Bowtie 2
Generating VCF files
Samtools & VCFtools
4mapping.py
BLASTn
Assembled contigs vs.
Conus exons
5recipblastWrapper.py
Evaluating sequences :
Coverage & Heterozygosity
6cov_hetWrapper.py
Filtering sequences with:
- Heterozygosity > 2 SD
- Coverage < 4X in more than
70% of sequence length
7filter_cov_het.py
First MAFFT alignments
Merging doublons
8cut_fasta_and_1stalign.py
8.5merging_doublons.py
BLASTx
Extracting ORFs
9blastx.py
Generating 3 datasets :
NT123 / NT12 / AA
10sequences_translation.py
Sorting sub datasets :
NEO 700 / NEO 50 /
Raphito 700 / Raphito 50
11sorting_samples_NEO700.py
11sorting_samples_NEO50.py
11sorting_samples_Raphito700.py
11sorting_samples_Raphito50.py
Tree reconstructions
IQ-Tree
Softwares
Python Scripts

Figure 17 : Pipeline d'analyses bioinformatiques post-séquençage.
Les données brutes sont représentées par le cadre bleu. Les cadres verts représentent les étapes pour lesquelles nous avons utilisé des scripts Python et les cadres rouges représentent les étapes pour lesquelles nous avons utilisé des logiciels.

Avec l'option « LEADING », nous supprimons les bases au début de chaque read qui descendent en dessous d'un score de qualité de 15. Nous faisons la même chose pour la fin de chaque read avec l'option « TRAILING » et un score de qualité de 15 également. Ces choix s'expliquent par le fait que, lors du séquençage en double lecture, la fin de la lecture du premier brin est généralement de moins bonne qualité. Cette perte de qualité s'explique par la diminution de l'efficacité de la polymérase une fois la fin du brin atteinte. Le brin est alors séquencé dans l'autre sens, et on constate que la qualité du début de la lecture est également moins bonne qualité avant d'augmenter de nouveau pour atteindre des valeurs suffisantes. Avec l'option « MINLEN », nous supprimons également les reads qui font une taille inférieure à 36 pb. Nous effectuons un séquençage de deux fois 150 pb, donc la lecture de brins de taille insuffisante peut être le résultat d'un mauvais séquençage ou de l'absence de matériel de qualité suffisante dans la préparation de banques pour un échantillon donné. Les scores phred de chacun de nos reads sont convertis en score phred+33 qui est communément utilisé pour les séquençages Illumina ainsi que pour le séquençage Sanger. Nous obtenons alors des reads qui ont été trimmés, certains étant appariés, d'autres non. Les reads appariés ainsi trimmés sont ensuite traités par le logiciel Flash v1.2.11 (Magoc & Salzberg, 2011). Il va permettre de combiner des reads qui sont plus courts que la moitié de la longueur totale des reads normalement séquencés, ici 150 pb. Nous avons utilisé l'option « minOverlap », qui est la longueur minimale chevauchante de deux reads pour les combiner, avec une longueur minimale de 5 pb, et l'option « maxOverlap », qui est la longueur maximale de chevauchement entre deux reads pour les combiner, qui est de 90% de la longueur des reads. Nous avons utilisé une longueur de 100 pb car nous avons des reads de 150 pb pour des tailles de fragments dans nos banques de séquençage d'environ 300 pb. Nous avons aussi utilisé l'option « mismatchRatio », qui est le ratio maximal autorisé du nombre d'erreurs (« mismatchs ») en fonction de la longueur du chevauchement entre deux reads. Si un chevauchement entre deux reads a plus d'erreurs que le seuil de 5% que nous avons choisi, alors ils ne seront pas combinés. Nous avons utilisé l'option « -f » pour intégrer la longueur moyenne des fragments, ici 300 pb. Une fois les reads appariés traités par le logiciel Flash, nous utilisons le logiciel Prinseq Lite v0.20.4 (Cantu

et al., 2019) pour vérifier les séquences qui ont été combinées par Flash en les filtrant sur la complexité des séquences (éléments répétés au sein des séquences par exemple). Cela nous permet au final d'avoir des séquences R1 et R2 combinées indépendamment et de retirer les séquences de mauvaise qualité du jeu de données. Les séquences non appariées sont traitées à part afin de générer un fichier comprenant toutes les séquences de bonne qualité qui n'ont pas pu être appariées.

(ii) Assemblage : Les séquences nettoyées sont alors assemblées en utilisant le logiciel SPAdes v3.8.1 (Bankevich et al., 2012). Ce logiciel utilise les k-mers pour construire un graphique de De Bruijn. Le logiciel va utiliser les structures des graphiques générés, la couverture et la longueur des séquences pour effectuer l'assemblage des séquences.

(iii) Clustering : Les séquences alors assemblées sont ensuite traitées avec le logiciel CAP3 (Huang & Madan, 1999). Ce logiciel va permettre de retirer les régions de mauvaise qualité en 3' et 5' des contigs assemblés avec SPAdes. Les séquences qui ne sont pas réellement chevauchantes sont identifiées et retirées. Ensuite, les contigs chevauchants sont regroupés des plus chevauchants vers les moins chevauchants. Enfin, des alignements multiples sont effectués entre les différents contigs afin de produire une séquence consensus. Nous utilisons alors les contigs générés par CAP3 pour les traiter avec le logiciel CD-HIT v4.8.1 (W. Li & Godzik, 2006). Nous utilisons la fonction CD-HIT-EST pour combiner des contigs qui ont une ressemblance supérieure à 99%. Cela va nous permettre de réduire la redondance dans les contigs assemblés par le logiciel SPAdes.

(iv) Mapping : Les contigs ainsi assemblés et combinés par similarité sont alors alignés avec le logiciel bowtie2 v2.5.1 (Langmead & Salzberg, 2012). Lors de cette étape nous allons aligner les séquences nettoyées lors de l'étape (i) qui ont été appariées (R1 et R2) sur nos contigs assemblés qui seront utilisés comme référence. Nous utilisons l'option « -local » pour effectuer un alignement local entre notre référence et les séquences R1 et R2 que l'on souhaite aligner. L'alignement local signifie que des portions des séquences à aligner peuvent être ignorées afin de ne pas générer de gaps dans la référence. Nous utilisons l'option « no-discordant », qui empêche bowtie2 de chercher des alignements discordants si aucun alignement concordant n'est trouvé. Un alignement discordant consiste en l'alignement de deux séquences mais qui ne satisfont pas les contraintes du séquençage en deux brins. Nous ajoutons également l'option « very-sensitive-local » qui va intégrer plusieurs options :

- Une option « -D 20 », qui autorise bowtie2 à générer plusieurs extensions de graines qui vont pouvoir échouer. Cela signifie que bowtie2 n'a pas été capable de trouver un nouvel alignement de meilleure qualité que le précédent.

- Une option « -R 3 », qui représente le nombre maximum de fois où bowtie2 va choisir de nouvelles séquences pour produire des graines répétées. Une séquence est considérée pour avoir des graines répétées lorsque le nombre total de graines trouvées pour cette séquence divisé par le nombre de graines qui vont s'aligner au moins une fois est supérieur à 300.

- Une option « -N 0 », qui est le nombre d'erreurs autorisées dans un alignement de plusieurs sous échantillons de chaque séquence.

- Une option « -L 20 », qui est la longueur des sous-échantillons de chaque séquence qui vont être alignés.

- Une option « -i S, 1, 0,50 », qui correspond à l'intervalle entre chaque sous-échantillonnage de graines qui vont être utilisées dans les alignements multiples.

Les séquences non appariées sont également alignées sur la référence à l'aide de bowtie2. Nous utilisons les mêmes options que pour les séquences R1 et R2, en retirant l'option « no-discordant ».

Le logiciel samtools v1.10 (H. Li et al., 2009) est ensuite utilisé pour effectuer plusieurs manipulations sur les fichiers de sortie produits par bowtie2. Les fichiers des séquences appariées et non appariées sont combinés puis triés et indexés. Puis nous retirons les séquences dupliquées à l'aide du logiciel Picard (ref) et de la fonction MarkDuplicates. Les fichiers sont ensuite indexés pour permettre le base calling et la création des fichiers de variant calling (vcf), qui vont nous servir pour effectuer des étapes de filtration des données dans les étapes suivantes du pipeline. Ces fichiers vcf vont nous permettre de connaître à chaque position d'un contig quelle base est présente ainsi que la qualité de cette base et les variants possibles qui ont été trouvés.

(v) BLAST : Nous effectuons ensuite un BLASTn entre les contigs assemblés (qui sont utilisés comme référence) et les 1125 exons de *Conus* que nous avons dessinés (voir chapitre 2, partie 1). Les résultats issus du BLASTn sont ensuite utilisés pour effectuer une première étape de filtration à l'aide de scripts Python spécifiques

(https://github.com/Hyperdiverseproject/Exon_capture). Un premier script (makeVCFcov.py) va être utilisé pour produire un fichier de couverture pour chaque base de chaque contig qui a eu un match avec un exon de *Conus*. Ces informations de couverture par bases sont récupérées à l'aide du fichier vcf qui a été produit lors de l'étape iv (mapping). Un second script (contig_filter.v3.py) va filtrer les contigs qui ont eu un hit en BLASTn contre au moins un exon de *Conus*. Si plusieurs séquences, d'un même échantillon, ont un match sur la même portion d'un exon de *Conus*, et si ces séquences se chevauchent sur plus de 20% de leur longueur totale, alors le contig avec la profondeur de séquençage la plus grande est conservé. Si deux séquences qui se sont alignées sur des portions différentes du même exon de *Conus* et par conséquent ne se chevauchent pas ou sur moins de 20% de leur longueur totale alors les deux séquences sont conservées pour les analyses suivantes. Les identifiants des contigs et leurs séquences qui sont supprimés à cette étape sont mis de côté, si des vérifications ultérieures seraient nécessaires.

(vi) Couverture et Hétérozygotie : Une seconde étape de filtration est réalisée sur les contigs qui ont été gardés à la suite du filtre réalisé après le BLASTn. C'est un ensemble de différents scripts python qui vont être utilisés pour réaliser ces différents filtres. Pour chaque échantillon, nous allons calculer le taux d'hétérozygotie et la profondeur de couverture. Pour chaque séquence qui a eu un match en BLASTn contre les séquences des exons de *Conus*, c'est-à-dire les séquences qui nous ont servi à dessiner les sondes de capture, nous allons calculer le nombre de différences entre chaque base d'un contig et la référence chez *Conus* sur laquelle il s'est aligné. Nous allons également récupérer la profondeur de séquençage à chaque base (get_cov_het.py). Les séquences chevauchantes vont également être récupérées et stockées dans un fichier qui sera consulté ensuite lors de la phase de filtre stricto sensu (removed_overlappers.py). La liste des positions alternatives dans nos fichiers vcf pour chaque échantillon sont conservées dans un fichier qui sera également consulté pour la phase de filtration ultérieure.

(vii) Filtration : Toutes les informations récupérées pour chaque échantillon vont nous permettre d'appliquer plusieurs filtres pour retirer du jeu de données les séquences qui ont des allèles trop divergents ou une couverture trop faible. On considère que ces séquences vont introduire des erreurs potentielles qui ne seraient pas liées à une divergence « réelle » de la séquence mais plutôt à des erreurs lors du séquençage ou à un ADN en entrée de trop faible qualité. Avec ce script Python de filtration des séquences, nous allons retirer les séquences qui ont une couverture inférieure à 4X sur plus de 70% de la longueur de la séquence. Nous retirons

également les séquences qui ont un taux d'hétérozygotie, c'est-à-dire une divergence entre les deux allèles qui est supérieure à 2 écart-type de la moyenne.

(viii) Alignements : Nous effectuons ensuite les alignements entre les séquences disponibles ; à savoir, les séquences des exons de *Conus* qui nous ont servi à produire les sondes de capture et les séquences des échantillons qui se sont alignées contre chaque exon. Pour effectuer les alignements, nous utilisons le logiciel MAFFT. Afin de minimiser l'insertion de gaps dans la séquence de référence au cours de l'alignement, nous utilisons la fonction « addfragments » de MAFFT. Avec cette fonction, le logiciel aligne les unes après les autres les séquences de nos échantillons sur notre référence en conservant la séquence de *Conus* comme référence. Cela signifie que nous ne faisons pas un alignement multiple entre toutes nos séquences mais que nous considérons spécifiquement la séquence de *Conus* pour chaque exon identifié comme la référence dans un alignement donné. Avant de lancer l'alignement, nous utilisons un script Python pour délimiter toutes les séquences à la longueur de l'exon ; c'est-à-dire que nous utilisons seulement la portion de chaque séquence pour chaque échantillon qui s'aligne sur la séquence de référence de *Conus*. Cela dans le but de minimiser l'insertion potentielle de gaps dans l'alignement. Les 1125 alignements sont concaténés avec le logiciel AMAS (Borowiec, 2016), afin de créer une supermatrice pour la reconstruction des arbres phylogénétiques.

(ix) Reconstruction phylogénétique : pour repérer les contaminations évidentes, c'est-à-dire les échantillons identifiés dans un taxon qui seraient groupés avec des échantillons de taxons que l'on sait phylogénétiquement éloignés, nous construisons un premier arbre phylogénétique en utilisant le logiciel IQTree v2.2.2.7 (Nguyen et al., 2015) avec le modèle GTR+G ainsi que 1000 UltraFastBootstraps. Cet arbre va également nous permettre de repérer les échantillons avec des longues branches, qui peuvent être associés soit à une contamination soit à un manque de données pour cet échantillon car trop peu d'exons ont été capturés.

4.2. IDENTIFICATION DES DIFFERENTS POINTS A AMELIORER DANS LE PIPELINE.

C'est avec le pipeline décrit précédemment, correspondant donc au pipeline mis au point et utilisé dans l'équipe avant ma thèse, que nous avons analysé les échantillons des 3 batchs de séquençage afin de vérifier dans un premier temps la qualité de nos données. Cependant, ces premières analyses ont également permis de mettre en évidence des améliorations que nous

pouvions mettre en place. L'objectif était de maximiser la qualité et la quantité de données conservées au final, en minimisant ainsi la proportion de données manquantes de notre jeu de données.

Un collaborateur de l'équipe, Alexander Fedosov, nous a rapporté qu'il avait amélioré la qualité de son jeu de données en utilisant deux logiciels d'assemblage et en combinant les contigs ainsi assemblés ((A. E. Fedosov et al., in press), Annexe 1). Il a noté que, en plus d'assembler les contigs avec le logiciel SPAdes, l'ajout d'un assemblage obtenu avec le logiciel Trinity v2.11 (Grabherr et al., 2011) permettait d'obtenir plus de contigs assemblés. En effet, chaque logiciel assemble certains contigs différemment, ce qui aura pour résultat d'augmenter le nombre de contigs totaux que nous pourrons utiliser pour la suite du pipeline. Le logiciel Trinity a été conçu pour l'assemblage de données transcriptomiques. En utilisant un sous-jeu de données de test sur 32 échantillons (Tableau 7), j'ai pu confirmer l'intérêt d'ajouter le logiciel Trinity au logiciel SPAdes lors de la phase d'assemblage. On obtient quasi systématiquement plus de contigs avec la combinaison des deux logiciels d'assemblage, par comparaison à l'utilisation de SPAdes seul (Tableau 7). Le temps d'assemblage avec Trinity est en revanche beaucoup plus grand que pour SPAdes. Malgré cela, l'amélioration du jeu de données est suffisante pour justifier l'utilisation des deux logiciels d'assemblage pour l'ensemble de nos données. C'est pour cela que nous avons décidé d'utiliser à la fois le logiciel SPAdes et le logiciel Trinity pour notre jeu de données final de 1728 échantillons.

Tableau 7 : Liste des 32 échantillons utilisés pour tester plusieurs points d'amélioration du pipeline bioinformatique.

Spécimen	#reads	SPAdes	Trinity	Trinity+ SPAdes	SF	Famille	Genre	Identification (* = espèce-type)
Buc172	4206534	1045	1077	1076	Buccinoidea	Buccinidae	*Metajapelion*	*Metajapelion adelphicus*
Buc242	4324310	575	923	933	Buccinoidea	Prosiphonidae	*Proneptunea*	*Proneptunea* sp.
C571611	5504772	836	963	964	Conoidea	Raphitomidae	*Trochodaphne*	*Trochodaphne cuprosa**
C571734	6401464	845	976	977	Conoidea	Raphitomidae	*Vitjazinella*	*Vitjazinella* sp.
IM-2007-17892	5112000	768	993	991	Conoidea	Raphitomidae	*Glyphostomoides*	*Glyphostomoides* sp.
IM-2007-30011	5216426	354	375	375	Mitroidea	Mitridae	*Cancillopsis*	*Cancillopsis liliformis*
IM-2007-31778	6012378	877	942	939	Buccinoidea	Nassariidae	*Reticunassa*	*Reticunassa paupera**
IM-2007-32709	4333372	1039	1083	1085	Buccinoidea	Prodotiidae	*Caducifer*	*Caducifer truncata**
IM-2007-33523	3632694	883	923	928	Buccinoidea	Columbellidae	*Graphicomassa*	*Graphicomassa ligula**
IM-2007-39266	5877174	997	1045	1047	Olivoidea	Olividae	*Calyptoliva*	*Calyptoliva bbugeae*
IM-2009-10395	5538248	816	893	895	Volutoidea	Cancellariidae	*Euclia*	*Euclia cassidiformis**
IM-2009-13301	2727632	98	318	331	Buccinoidea	Columbellidae	*Ascalista*	*Ascalista letourneuxi*
IM-2009-23725	5524406	997	1031	1031	Buccinoidea	Nassariidae	*Cyllene*	*Cyllene lamarcki*
IM-2009-5139	4178322	425	890	898	Muricoidea	Muricidae	*Ponderia*	*Ponderia magna*
IM-2009-7449	1249740	46	56	57	Buccinoidea	Buccinidae	*Buccinum*	*Buccinum undatum**
IM-2009-8140	4545286	142	225	187	Buccinoidea	Prosiphonidae	*Antarctodomus*	*Antarctodomus thielei**
IM-2009-8165	4112050	283	288	286	Conoidea	Borsoniidae	*Belaturricula*	*Belaturricula ergata*
IM-2013-47082	5146694	447	676	682	Volutoidea	Marginellidae	*Dentimargo*	*Dentimargo* sp.
IM-2013-49963	1868874	479	790	803	Buccinoidea	Nassariidae	*Phrontis*	*Phrontis tiarula**
IM-2013-54563	4538152	435	686	688	Buccinoidea	Columbellidae	*Metanachis*	*Metanachis jaspidea**
IM-2013-60590	7047094	839	1053	1053	Muricoidea	Muricidae	*Actinotrophon*	*Actinotrophon actinophorus**
IM-2013-61512	5625274	914	1043	1046	Conoidea	Borsoniidae	*Darbya*	*Darbya lira*
IM-2013-61836	3823730	1048	1081	1082	Buccinoidea	Eosiphonidae	*Manaria*	*Manaria* sp.
IM-2013-63302	6352828	875	986	988	Tonnoidea	Bursidae	*Korrigania*	*Korrigania fijiensis**
IM-2013-70867	5974960	708	808	808	Buccinoidea	Columbellidae	*Decipifus*	*Decipifus sixaolus**

Spécimen	#reads	SPAdes	Trinity	Trinity+ SPAdes	SF	Famille	Genre	Identification (* = espèce-type)
IM-2013-70992	2284842	80	108	167	Buccinoidea	Columbellidae	*Steironepion*	*Steironepion pygmaeum*
IM-2013-73023	8076120	405	474	473	Truncatelloidea	Truncatellidae	?	Gen. sp.
IM-2019-1690	4065288	838	957	961	Turbinelloidea	Costellariidae	*Austromitra*	*Austromitra weldii*
IM-2019-17604	3046410	912	967	970	Tonnoidea	Cymatiidae	*Monoplex*	*Monoplex parthenopeus**
IM-2019-1786	1041962	1110	730	386	Conoidea	Borsoniidae	*Tropidoturris*	*Tropidoturris* sp.
IM-2019-1936	5307162	948	1002	1002	Buccinoidea	Fasciolariidae	*Australaria*	*Australaria australasia**
UF450469	4946280	837	1022	1025	Volutoidea	Volutidae	*Scaphella*	*Scaphella junonia**

Une autre étape du pipeline sur laquelle nous avons identifié une potentielle amélioration de nos données est l'étape de BLAST. Nous avons initialement décidé d'aligner les contigs assemblés de chaque échantillon contre les séquences de référence de *Conus* pour les 1125 exons. Nous nous sommes posé la question de savoir si ajouter toutes les autres séquences qui nous ont servi au dessin des sondes de captures pouvait permettre de récupérer plus de contigs lors de cette étape de BLAST. Cela inclut toutes les séquences des transcriptomes pour chaque exon final retenu, plus les séquences ancestrales reconstituées pour réduire la distance génétique entre les séquences des sondes et les séquences des échantillons à capturer. L'ajout de ces séquences améliore parfois les résultats, c'est-à-dire que certains contigs s'alignent contre une séquence qui n'est pas celle de *Conus*. Cependant, l'ajout de ces séquences augmente significativement le temps de calcul pour chaque BLAST, et étant donné la taille du jeu de données, à savoir 1728 échantillons, nous avons donc décidé de ne pas retenir cette amélioration potentielle du pipeline.

Une autre étape sur laquelle j'ai effectué des modifications est l'étape post alignement. En effet, un nettoyage des données post-alignements est effectué afin de retirer les doublons potentiel, c'est-à-dire les séquences non chevauchantes ou qui se chevauchent sur quelques bases pour un échantillon donné, mais qui correspondent au même exon. J'ai remarqué que cela concernait une faible proportion des échantillons du jeu de données, mais la concaténation de ces séquences permet de réduire la proportion de données manquantes.

Concernant la reconstruction phylogénétique, nous avons décidé de faire des reconstructions basées sur des alignements en nucléotides et des alignements en acides aminés. Il a fallu dans un premier temps identifier le bon cadre de lecture afin que la traduction soit dans la bonne phase et ne pas introduire de codons stop dans nos alignements. J'ai alors utilisé les séquences des 1125 exons de *Conus*. Ces séquences nucléotidiques ont été alignées avec un BLASTx contre les protéines de référence qui ont été publiées par Pardos-Blas (Pardos-Blas et al., 2021) avec le génome complet de *Conus ventricosus*. Les résultats m'ont permis d'identifier la phase ainsi que le point de départ pour effectuer la traduction. À partir de ces résultats, j'ai écrit, avec l'aide de Sarah Farhat, un script en Python, nous permettant d'effectuer plusieurs opérations sur nos alignements. Chaque alignement contient la séquence de référence de *Conus* ainsi que tous les échantillons qui ont eu un résultat en sortie de BLAST lors de l'étape v. J'ai décidé de me baser sur la séquence de *Conus* que l'on considère comme « vraie », pour limiter l'impact des biais qui pourraient être introduits avec des erreurs de séquençage. Ainsi, en prenant le cadre de lecture donné par le résultat de BLASTx, j'ai retiré toutes les positions qui

correspondent à un gap sur une ou deux positions consécutives (et donc qui change la phase) dans la référence *Conus* et qui sont aussi des gaps dans au moins 60% des séquences présentes dans l'alignement. Nous considérons qu'il est peu probable qu'une erreur de séquençage, ici la présence d'une base supplémentaire dans un échantillon par rapport à *Conus*, soit partagée par plus de 60% des échantillons de l'alignement. Autrement dit, on considère comme plus probable que l'erreur de séquençage ait eu lieu lors du séquençage du génome de *Conus*. De plus, nous avons décidé également de retirer les gaps même s'ils sont en triplets (et donc ne changent pas la phase) lorsqu'ils sont également partagés par plus de 60% des séquences dans chaque alignement. Ici aussi, on considère que la présence de nucléotides supplémentaires dans un faible nombre d'échantillons de l'alignement serait plutôt due à une erreur de séquençage plutôt qu'à une réalité biologique. Nous obtenons à la suite de l'utilisation de ce script les alignements en nucléotides ainsi que les alignements protéiques qui ont été traduits dans le bon cadre de lecture.

5. RESULTATS BRUTS

Nous avons séquencé un total de 1728 échantillons (Annexe 2), correspondants aux différents jeux de données qui seront présentés dans les parties suivantes (succès de la capture d'exon – chapitre 3 ; phylogénie des néogastéropodes – chapitre 4 ; phylogénie des Raphitomidae – chapitre 4), ainsi que des échantillons de projets annexes (voir ci-dessus pour plus de détails). Le nombre de reads obtenus est variable entre les échantillons, avec des valeurs allant de 17432 reads jusqu'à 28 713 588 reads. La moyenne de reads par échantillons pour le batch 1 est de 4 940 624, de 4 927 656 pour le batch 2 et enfin de 4 929 537 pour le batch 3. Il y a donc peu de différence en moyenne entre le nombre de reads obtenus par échantillon des 3 batchs de séquençage. Le nombre d'exons capturés par échantillon est aussi variable, allant de 0 jusqu'à 1080 (Annexe 2). Pour le batch 1, cela représente une moyenne de 877 exons capturés par échantillon, pour le batch 2, nous obtenons une moyenne de 816 exons capturés par échantillon et enfin pour le batch 3, la moyenne d'exons capturés par échantillon est de 817 (Figure 18).

Nous avons retiré 57 échantillons du jeu de données car nous n'avons pas pu extraire d'exons à la fin des phases de filtration du pipeline utilisé en post-séquençage. Sur ces 57 échantillons, 1 échantillon avait été séquencé dans le batch 1, 9 échantillons dans le batch 2 et 47 l'ont été dans le batch 3. Pour relever les éventuelles contaminations dans notre jeu de données, nous

avons généré des arbres phylogénétiques suivant des modèles de substitutions simples pour diminuer le temps de calcul. Cela nous a permis de retirer 53 échantillons du jeu de données final. Sur ces 53 échantillons, 4 avaient été séquencés avec le batch 1, 11 avec les batch 2 et 38 avec le batch 3. Nous avons enfin 21 échantillons qui ont été retirés du jeu de données final car ils étaient placés dans l'arbre avec une longue branche. Parmi ces 21 échantillons, 6 étaient dans le batch 1, 3 dans le batch 2 et 12 dans le batch 3.

Nous avons donc une certaine disparité entre les 3 batchs de séquençage, le batch 3 étant le batch pour lequel nous avons exclu plus d'échantillons du jeu de données. Cela peut s'expliquer par le fait que dans ce dernier batch nous avons intégré des échantillons de groupes taxonomiques plus éloignés des néogastéropodes. De plus, le batch 3 incluait un plus grand nombre d'échantillons « anciens », avec des ADN de moins bonne qualité (voir chapitre ci-dessus). Il est important de noter que les aliquots d'ADN d'un même échantillon ont fourni des résultats très similaires (voir Annexe 2 et chapitre 3 pour plus de détails).

Au final, 1597 échantillons (373 pour le batch 1, 361 pour le batch 2 et 863 pour le batch 3) ont été conservés pour les analyses. Parmi eux, 155 ont été utilisés dans la cadre du test de l'efficacité de la capture d'exons (Chapitre 3), 856 pour la phylogénie des néogastéropodes (Chapitre 4) et 322 pour la phylogénie des Raphitomidae (Chapitre 4). Les échantillons restants seront utilisés par nos collaborateurs pour leurs phylogénies respectives.

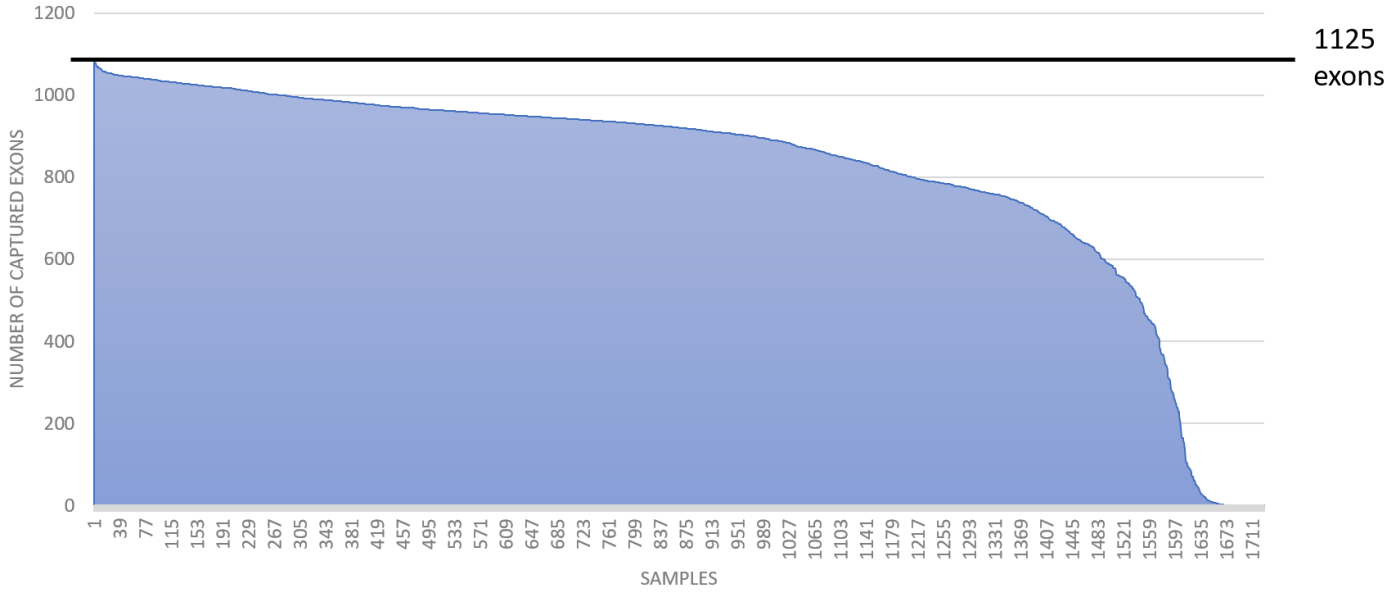


Figure 18 : Représentation graphique du nombre d'exons capturés pour chacun des 1728 échantillons séquencés au total. *La barre noire indique le nombre total d'exons que nous avons désignés pour le projet à savoir 1125.*

Chapitre 3 : Distance génétique et succès de la capture d'exon

L'effet de la distance génétique sur le succès de la capture d'exons n'a jamais été testé avec un protocole expérimental dédié. Nous avons mis en place un protocole contrôlé dans lequel nous allons utiliser des réplicats pour des spécimens différents d'une même espèce mais également pour un même échantillon, pour lequel plusieurs aliquots d'ADN ont été séquencés.

Le jeu de données se base sur des spécimens de 30 espèces différentes, 14 espèces représentatives des Neogastropoda et 16 espèces représentatives des Caenogastropoda. Ils ont été collectés lors de la même mission en Corse (CORSICABENTHOS 3) en 2021. Afin de limiter les biais liés aux expérimentations, les 150 spécimens ont été traités dans le même laboratoire et à la même période. Pour tester la variabilité des expérimentations de laboratoire, de l'extraction d'ADN jusqu'au séquençage, nous avons également fait des aliquots pour 4 spécimens. Pour chacun, l'extrait d'ADN a été séparé en 4 sous échantillons indépendants. Nous avons un total de 161 échantillons pour effectuer les tests.

Comme expliqué dans le chapitre précédent, ce jeu de données inclut un nouveau jeu d'exons dessiné afin d'effectuer la capture et le séquençage, dans le but de capturer des spécimens présents dans l'ensemble des néogastéropodes ainsi que dans les groupes de Caenogastropoda proches. Nous avons utilisé 15 transcriptomes et un génome complet afin de dessiner les sondes de capture.

Sur les 161 échantillons, 2 spécimens se sont avérés contaminés, et les 4 aliquots d'un même spécimen avaient une longue branche dans l'arbre phylogénétique que nous avons reconstruit. Ces échantillons ont été retirés du jeu de données afin d'effectuer les tests de corrélation de succès de la capture d'exons. Nous constatons un effet de la distance génétique sur le succès de la capture d'exons : plus le spécimen séquencé est génétiquement distant des transcriptomes qui nous ont servi à produire les sondes de capture, moins le nombre d'exons retrouvés est grand.

L'ajout de nouveaux transcriptomes dans les groupes ciblés pour la capture permettrait d'augmenter les chances de capturer les exons en produisant de nouvelles sondes capables de capturer les mêmes exons.

Ces résultats sont présentés dans l'article en préparation ci-dessous.

Too far from relatives? Impact of the genetic distance on the success of exon capture phylogeny.

Lemarcis Thomas[1*], Blin Amandine[2], Derzelle Alessandro[1], Farhat Sarah[1], Fedosov Alexander[1,3], Zaharias Paul[1], Zuccon Dario[1], Puillandre Nicolas[1].

1. Institut Systématique Evolution Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, CNRS, EPHE, Sorbonne Université, Université des Antilles, Paris, France, 47 rue Cuvier, CP51, 75005 Paris.

2. Service Analyse de Données, Muséum National d'Histoire Naturelle, Centre National de la Recherche Scientifique, UAR 2700 2AD, CP 51, 57 rue Cuvier, F-75231 Paris Cedex 05, France

3. A. N. Severtsov Institute of Ecology and Evolution, Russian Academy of Sciences, Moscow, Russia

Corresponding author: thomas.lemarcis@mnhn.fr

Introduction

Phylogenetic reconstruction using DNA sequences are now based on a large spectrum of dataset types, from single locus phylogeny to whole genome sequencing, including multilocus datasets based on Sanger sequencing, RAD-sequencing, reduced-genome approaches and transcriptomes. Among them, sequence capture methods, which rely on an enrichment of selected loci, have been widely used to address relationships at varying depths, from the species level to the early branches of the tree of life, and in a wide range of organisms (Andermann et al., 2020; Jones & Good, 2016; Nunes et al., 2022). For a reduced cost, these methods offer the possibility to obtain large datasets of hundreds or thousands of loci, without the need to sequence whole genomes or transcriptomes, significantly more expensive (Zaharias, Pante, et al., 2020). Furthermore, they are based on short fragment sequencing techniques, and thus appropriate for poorly preserved samples, as those available in historical collections in museums (museomics - (Andermann et al., 2020; Brewer et al., 2019; Derkarabetian et al., 2019; Forrest et al., 2019).

Among them, the exon-capture approach relies on the use of available genomic and/or transcriptomic data to identify a set of conserved exons, with an appropriate level of variability to resolve the phylogenetic relationships among the taxa of interest (Bi et al., 2012; Derkarabetian et al., 2019; Hugall et al., 2016; Nunes et al., 2022; Teasdale et al., 2016) The sequences of these loci (exons) are then used to design probes that are needed to capture the exons. The key step is thus to adequately design the probes, so they will be able to efficiently bind to the targeted DNA, maximizing the fraction of reads on targets obtained after sequencing, and thus reducing the proportion of missing data in the final dataset. Indeed, if the efficiency of the method (‘capture success’) also depends on the quantity and the quality of the DNA, themselves depending in particular of the age of the sample and the quality of its preservation (Nunes et al., 2022), the genetic distance between the probes and the DNA to which they are supposed to bind is also crucial.

Empirical tests have shown that the capture success, measured as the sequencing depth of coverage (Bragg et al., 2016), the sensitivity, the specificity and the proportion of missing data (Portik et al., 2016) or the number of recovered loci (Bartoš et al., 2023), decreases when the genetic distance between the sample’ and probe’s sequences increases. Although the distance threshold above which the capture success significantly declines is variable and not always clearly identified, Bragg et al. (Bragg et al., 2016) proposed that 10% is a reasonable estimation. However, these studies did not focus solely on testing the impact of genetic distances on capture

success, but took the opportunity of an available dataset designed to resolve phylogenetic relationships within the taxon of interest. Consequently, there is a need to control these factors, independent from the genetic distance, that may influence the capture success, i.e. by including replicates and by controlling the quality of the DNA, in order to more clearly quantify the part of the variability in the capture success linked to the genetic distance.

Here, we designed a specific protocol, including multiple DNA aliquots from the same specimen (called "intraspecimen" replicates) and multiple specimens from the same species (called "intraspecies" replicates), to test both for the reproducibility of the library preparation and sequencing and the repeatability among samples of the same species, i.e. with the same genetic distance with the probes. Furthermore, all the samples were processed with exactly the same protocol, from the sampling on the field to the sequencing, and collected during the same expedition, to limit the variability of DNA quality and quantity and its impact on capture success. We selected specimens from 30 species among the Neogastropoda (Mollusca, Gastropoda, Caenogastropoda), the taxa from which the genomes and transcriptomes used to design the probes were obtained (and thereafter referred to as the ingroup) and other Caenogastropoda non-Neogastropoda, from closely related to more distant (and thereafter referred to as the outgroup), to test for a potential correlation between the capture success (estimated as the number of reconstructed loci) and the genetic distance between the reconstructed exons and the closest probe.

Material and Methods

1. Sampling

All the specimens were collected during a single expedition, organized by the Muséum National d'Histoire Naturelle, Paris (MNHN), in Corsica (CORSICABENTHOS 3) in 2021, and all were processed with the same method, from preservation on the field to DNA extraction and sequencing, to minimize the impact of DNA quality on the sequencing. A total of 30 species, including representatives of Neogastropoda (14 species) and Caenogastropoda non-Neogastropoda (16 species) were selected, with 5 specimens per species (150 specimens in total – Supp. Mat. 1). Among non-Neogastropoda, 5 species belong to lineages closely related to the neogastropods (A. E. Fedosov et al., in press; Lemarcis et al., 2022) and to the Ovulidae family, a lineage from which was obtained one of the transcriptomes used as a reference to design the probes (see below) – these are thereafter referred to as "outgroup 1"; the 11 other non-neogastropods belong to more distantly related caenogastropods, and are thereafter referred to

as “outgroup 2” (Fig. 1). On the field, specimens were microwaved and fixed in 96% ethanol (Galindo et al., 2014). All the DNAs were extracted using the Epmotion 5075 robot (Eppendorf), following the manufacturer’s recommendations, by the same operator (DZ).

2. Identification of the targeted exons

Given that our previously published exon-capture based datasets had resulted in a large proportion of missing data (Abdelkrim et al., 2018; A. E. Fedosov et al., in press; Phuong et al., 2019; Zaharias et al., in press), partly attributed to the use of a genome distantly related (*Lottia gigantea* (Simakov et al., 2013) and the low number of transcriptomes, we choose to design a new set of exons using a more closely related genome and a greater number of transcriptomes. The only chromosome-level and annotated genome available (at the time of the exon design in 2021) for neogastropods, was from *Conus ventricosus* (Conidae, (Pardos-Blas et al., 2021)). This genome was combined with 15 transcriptomes sourced from GenBank or from other unpublished projects of our team. We aimed at designing exons covering the whole Neogastropoda diversity. Consequently, when selecting the transcriptomes, we favored the taxonomic diversity among neogastropods, even at the cost of including poor quality sequences (with low BUSCO-scores – Table 1), in order to maximize the probability to design probes able to capture to capture targets across entire neogastropod diversity. One transcriptome was obtained from an Ovulidae, a family closely related to the neogastropods (A. E. Fedosov et al., in press).

Table 1: List of genome and transcriptomes used for exon selection, with Busco scores.

Family	Species	Reference	BUSCO score
Charoniidae	*Charonia tritonis*	(Bose et al., 2017)	76,62
Muricidae	*Rapana venosa*	(Song et al., 2016)	72,43
Conidae	*Profundiconus* cf. *vaubani*	(Fassio, Modica, et al., 2019)	68,76
Raphitomidae	*Typhlosyrinx* sp.	[new transcriptome]	68,34
Turridae	*Gemmula* sp.	(Zaharias, Kantor, et al., 2020)	58,70
Ovulidae	*Ovula ovum*	(Lemarcis et al., 2022)	47,90
Colubrariidae	*Cumia reticulata*	(M. V. Modica et al., 2015)	47,06
Raphitomidae	*Gymnobela pacifica*	[new transcriptome]	44,44
Personidae	*Distorsio anus*	(Lemarcis et al., 2022)	44,13
Volutidae	*Alcithoe aillaudorum*	[new transcriptome]	39,10
Melongenidae	*Hemifusus tuba*	(R. Li et al., 2019)	38,89

Mitridae	*Mitra mitra*	(Lemarcis et al., 2022)	38,16
Babyloniidae	*Babylonia areolata*	[new transcriptome]	36,37
Vasidae	*Vasum turbinellum*	(Lemarcis et al., 2022)	31,24
Olividae	*Oliva sericea*	(Lemarcis et al., 2022)	14,78

The methodology we applied to identify the exons and design the probes is generally based on the reciprocal BLAST approach (Phuong & Mahardika, 2018; Teasdale et al., 2016), with several modifications detailed below.

First, the 15 transcriptomes sequences were translated using Transdecoder v5.5.0 (Haas, BJ. https://github.com/TransDecoder/TransDecoder) to identify the longest ORFs. Then, the published proteome of *Conus* was aligned using BLASTp (Camacho et al., 2009) against the translated transcriptomes with an e-value lower or equal to 10^{-10}. Only unique matches were kept, i.e. transcripts that matched a single protein of *Conus*, in order to ignore paralogs. Expectedly, the number of unique matches is positively correlated with the BUSCO (Simão et al., 2015) score of each transcriptome, with numbers of hits ranging from 731 and 20,249 (Table 1). Multiple sequences from a given transcriptome that were matching the same *Conus* protein were assumed to correspond to isoforms of the same gene or incomplete coding sequences, and were retained in the dataset. Because of the low BUSCO scores of some of the transcriptomes, and assuming that the absence of a protein from these lower-quality transcriptomes was more likely due to incomplete sequencing or the lack of expression in the corresponding tissue rather than an absence from the genome, we decided to retain the proteins of *Conus* that were present in at least 8 out of the 12 best transcriptomes, ignoring at this step the 3 transcriptomes with the lowest BUSCO scores. We retrieved 1,675 proteins, corresponding to 7,675 exons longer than 120 bp and representing 1,978,743 bp.

For each predicted exon in *Conus*, we extracted the corresponding sequence from the 15 transcriptomes. Then, we discarded 329 proteins of *Conus* (= 2,469 exons) that had matched with more than 100 transcripts among the 15 transcriptomes, considering that so many sequences are more likely to correspond to duplicated gene sequences. Each *Conus* exon was aligned separately against all the corresponding transcripts with the MAFFT v7.520 software (Katoh & Standley, 2013). Filters were applied to remove the transcripts that had more than 35% of indels in the exon region and/or more than 50% of genetic distance with the reference sequence (i.e. the exon sequence in *Conus*). If several transcripts from a single transcriptome matching a given *Conus* exon were still present in an alignment, the transcript with the lowest genetic distance with the *Conus* exon was retained. If distances were equal, the longest retained

transcript was kept. At this step, the dataset included 4,744 exons. A second alignment was performed (thus including maximum 16 sequences – one for the genome and maximum 15 for the transcriptomes) with MAFFT, and alignments were trimmed at the exon boundaries. A second round of filtering was then performed: exons with less than 4 transcriptomes, with more than 5% of indels, with extreme GC contents (less than 30% or higher than 70%) or with at least one transcript with a genetic distance to the *Conus* exon below 2% were discarded. This filtering process resulted in 2,900 exons remaining in the dataset.

The specimens analyzed here have been processed as part of a larger project aiming at reconstructing the phylogeny of the Neogastropoda. Our goal was thus to maximize the capture efficiency, which is supposed to decrease when the genetic distances between the baits and the targeted exons exceed 10% (Bi et al., 2012; Bragg et al., 2016; Hedtke et al., 2013; Ilves & López-Fernández, 2014; Portik et al., 2016). To capture targets with genetic divergence greater than 10% from existing references, we reconstructed, with the software FastML v3.11 (Ashkenazy et al., 2012), consensus sequences from pairs of taxa corresponding to the most divergent ones within each of the two taxonomic groups, in order to obtain sequences with shorter distance to target: *Conus* / *Gemmula*, *Conus* / *Mitra*, *Cumia* / *Hemifusus*, *Charonia* / *Babylonia*, *Charonia* / *Alcithoe* and *Charonia* / *Rapana*. The reconstructed consensus sequences were added to the list of sequences issued from the genome and transcriptomes, to design the probes.

The bait production was subcontracted to MyBaits (Daicel Arbor, Ann Arbor, MI, USA), who further applied several steps of data cleaning and filtering, to end up with a number of exons and corresponding baits compatible with the chosen bait kit size: (i) the single repeats were marked; (ii) the bait sequences were compared by BLAST with three high quality genomes, the one used to identify the exons (*Conus ventricosus*), and two neogastropod genomes published in the meantime (*Monoplex corrugatus* and *Stramonita haemastoma* – (Farhat et al., 2023)) to remove overrepresented loci and merge identical baits; and (iii) the loci shorter than 160 bp were removed. The final dataset comprised 1,125 exons, represented by 15,905 sequences (including 4,117 consensus sequences), corresponding to 99,265 baits (80 nucleotides long, with a 3x tiling).

3. Library preparation, sequencing and assembly

Library preparation and sequencing were then performed on the 150 specimens. For 4 specimens, intra-specimen replicates (i.e. the DNA extract was split into 4 aliquots) were sequenced independently, in order to test the reproducibility of library preparation and

sequencing steps, leading to a total of 162 samples. The library preparations and exon captures were subcontracted to the ICM (Institut du Cerveau et de la Moelle épinière, Paris, France). The NEBNext Ultra II FS DNA Library Prep Kit for Illumina and the MyBaits Hybridization Capture kit for targeted NGS v5.0 (Arbor Biosciences, Ann Arbor, Michigan, USA) were used, following the manufacturer instructions. Together with other samples from the larger project, samples were sequenced on the NovaSeq Illumina 6,000 platform, with a S1 Reagent Kit (300 cycles) and on the NextSeq Illumina 2,000 platform, with a P2 Reagent Kit (300 cycles).

As for the exon design step, the methodology applied generally follows approaches previously published (e.g. (Abdelkrim et al., 2018)), with several modifications. To assemble the reads and maximize the number of contigs reconstructed, both SPAdes v3.8.1 (Bankevich et al., 2012) and Trinity v2.11 (Cabau et al., 2017) were used independently. The two datasets of contigs were merged. The contigs in the combined dataset were then clustered using cap3 (Huang & Madan, 1999) followed by cd-hit v4.6.2.5 (W. Li & Godzik, 2006). Reads were mapped on the contigs using bowtie2 v2.5.1 (Langmead & Salzberg, 2012).

A BLASTn alignment was then performed between the assembled contigs of each sample and the 1,125 sequences of exons from *Conus*. BLAST hits were then filtered using an in-house python script to remove sequences with less than 4X of coverage on more than 70% of the exon sequence length and/or with an estimated heterozygosity >2 SDs. At this step, when there are several contigs for one sample and one exon, the contig with the best coverage or, when the coverage is identical, the longest, is retained.

4. Phylogenetic analysis

An initial alignment using MAFFT with auto strategy, with the option « addfragments » was performed for each exon. When present, non-overlapping contigs for a given sample and a given alignment were concatenated with an in-house python script.

To translate the nucleotide alignments into proteins on the right frame, BLASTx was performed between each exon of *Conus* and the annotated proteins from the genome (Pardos-Blas et al., 2021) to identify the starting position of the corresponding ORF. An in-house python script was then used to translate all the alignments in the correct ORF, remove the sequences shorter than 60% of the length of the *Conus* target exon and trim the positions that contained more than 40% of missing data. Alignments which generated stop codons were then checked manually. ModelFinder, implemented in IQ-TREE v2.2.2.7 (Nguyen et al., 2015) was used to determine the best-fitting substitution model for each exon alignment, following the BIC criterion (Wit et

al., 2012), and reconstruct a phylogenetic tree on the concatenated nucleotide alignments with 100 bootstraps.

5. Correlation tests

In the correlation tests, only one intra-specimen replicate (the one with the highest number of reconstructed exons) per specimen, is retained. For each exon of each sample, we identified the closest bait using a BLAST between each exon and the 100K baits. The baits that have matched were then aligned using MAFFT with the option « addfragments », with the corresponding exon, and the bait with the lowest genetic distance (p-distance calculated with Goalign v0.3.7 – (Lemoine & Gascuel, 2021)) was retained. We then calculated the average of all the lowest genetic distances (of all the exons) for each sample, and the average of all the lowest genetic distances (of all the samples) for each exon, with in-house scripts. The distances calculated are p-distances because the binding of the probes to the DNA to be captured is directly dependent on the similarity between the sequences, and not on the potential hidden substitutions that would have been considered with a more complex substitution model (Hedtke et al., 2013).

A first plot (Fig. 2) was reconstructed using the ggstatsplot function from the ggplot package in R (R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/) to visualized the number of reads per sample, as a function of the average genetic distance for each sample, and we calculated a Kendall rank correlation coefficient (Kendall, 1938) to test association between the number of reads and the average genetic distance. Then, we plotted the distribution of the number of captured exons per specimen (Fig. 3), for each of the three groups of specimens (neogastropods, outgroup 1 and outgroup 2), in order to i) identify potential outliers and ii) compare capture success among the three groups. We also plotted the number of reconstructed exons per sample, as a function of the average genetic distance for each sample, and the number of specimens sequenced, as a function of the average genetic distance for each exon, and performed a Kendall rank correlation test for both pairs of variables. Given that more variable exons will probably be less efficiently captured for distantly related taxa, more conserved exon might be over-represented for these taxa, and thus potentially bias the results. To tentatively overcome this issue, we randomly selected 100 exons (over all the sequenced exons) and tested again the association; this procedure was repeated 1000 times using the tidyverse package in R.

Results & Discussion

1. Sequencing results and phylogenetic tree

Sequencing failure occurred for one sample, a replicate of IM-2019-17972 (*Erato voluta*); consequently, this sample is not included in the analyses. Additionally, two samples exhibited contamination, as they clustered with samples of different families: one replicate of IM-2019-17538 (*Coralliophila meyendorffii*) clustered with *Turritella turbona* and one replicate of IM-2019-17838 (*Caecum subannulatum*) clustered with *Alvania lineata* (Fig. 1). Moreover, the four intra-specimen replicates of the specimen IM-2019-16936 (*Dizoniopsis concatenata*) are isolated with a very long branch and do not cluster with the four other replicates of the same species within the Triphoroidea (Fig. 1). While these six samples were included in the tree, they were excluded from the correlation tests.

Among the remaining 161 samples, intra-specimen replicates clustered together in the tree with very short branches. Similarly, specimens from the same species also clustered together with very short branches. These results suggest that the impact of the sample preparation, including preservation on the field, DNA extraction, library preparation, exon capture and sequencing, is limited, as long as these steps are realized with the same protocol for all the samples. Consequently, the variability in terms of reconstructed exons should be mostly linked to the capacity of the probes to bind to the targeted DNA, and not to the quantity and quality of the targeted DNA.

Overall, the relationships among species generally follows the current classification. The Neogastropoda (except Cystiscidae and Granulinidae) appear monophyletic, and within Neogastropoda, the superfamilies and families represented by several species (Conoidea, Buccinoidea, Muricidae) are monophyletic as well. The recovered topology is mostly congruent with the latest published phylogeny of neogastropods based on mitogenomes (Lemarcis et al., 2022).

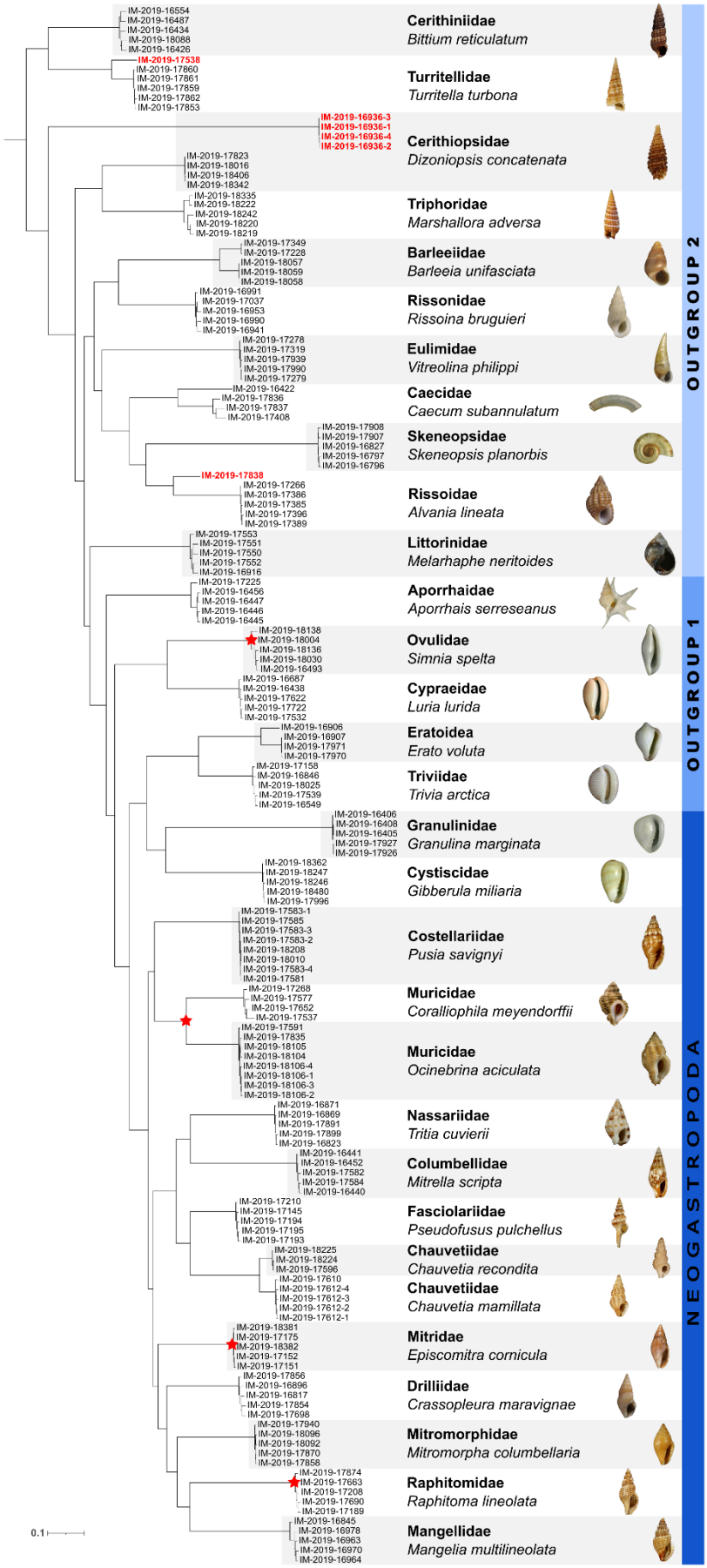

IM-2019-16554
IM-2019-16487
IM-2019-16434
IM-2019-18088
IM-2019-16426
Cerithiniidae
Bittium reticulatum
IM-2019-17538
IM-2019-17860
IM-2019-17861
IM-2019-17859
IM-2019-17862
IM-2019-17853
Turritellidae
Turritella turbona
IM-2019-16936-3
IM-2019-16936-1
IM-2019-16936-4
IM-2019-16936-2
Cerithiopsidae
Dizoniopsis concatenata
IM-2019-17823
IM-2019-18016
IM-2019-18406
IM-2019-18342
IM-2019-18335
IM-2019-18222
IM-2019-18242
IM-2019-18220
IM-2019-18219
Triphoridae
Marshallora adversa
IM-2019-17349
IM-2019-17228
IM-2019-18057
IM-2019-18059
IM-2019-18058
Barleeiidae
Barleeia unifasciata
IM-2019-16991
IM-2019-17037
IM-2019-16953
IM-2019-16990
IM-2019-16941
Rissonidae
Rissoina bruguieri
IM-2019-17278
IM-2019-17319
IM-2019-17939
IM-2019-17990
IM-2019-17279
Eulimidae
Vitreolina philippi
IM-2019-16422
IM-2019-17836
IM-2019-17837
IM-2019-17408
Caecidae
Caecum subannulatum
IM-2019-17908
IM-2019-17907
IM-2019-16827
IM-2019-16797
IM-2019-16796
Skeneopsidae
Skeneopsis planorbis
IM-2019-17838
IM-2019-17266
IM-2019-17386
IM-2019-17385
IM-2019-17396
IM-2019-17389
Rissoidae
Alvania lineata
IM-2019-17553
IM-2019-17551
IM-2019-17550
IM-2019-17552
IM-2019-16916
Littorinidae
Melarhaphe neritoides
IM-2019-17225
IM-2019-16456
IM-2019-16447
IM-2019-16446
IM-2019-16445
Aporrhaidae
Aporrhais serreseanus
IM-2019-18138
IM-2019-18004
IM-2019-18136
IM-2019-18030
IM-2019-16493
Ovulidae
Simnia spelta
IM-2019-16687
IM-2019-16438
IM-2019-17622
IM-2019-17722
IM-2019-17532
Cypraeidae
Luria lurida
IM-2019-16906
IM-2019-16907
IM-2019-17971
IM-2019-17970
Eratoidea
Erato voluta
IM-2019-17158
IM-2019-16846
IM-2019-18025
IM-2019-17539
IM-2019-16549
Triviidae
Trivia arctica
IM-2019-16406
IM-2019-16408
IM-2019-16405
IM-2019-17927
IM-2019-17926
Granulinidae
Granulina marginata
IM-2019-18362
IM-2019-18247
IM-2019-18246
IM-2019-18480
IM-2019-17996
Cystiscidae
Gibberula miliaria
IM-2019-17583-1
IM-2019-17585
IM-2019-17583-3
IM-2019-17583-2
IM-2019-18208
IM-2019-18010
IM-2019-17583-4
IM-2019-17581
Costellariidae
Pusia savignyi
IM-2019-17268
IM-2019-17577
IM-2019-17652
IM-2019-17537
Muricidae
Coralliophila meyendorffii
IM-2019-17591
IM-2019-17835
IM-2019-18105
IM-2019-18104
IM-2019-18106-4
IM-2019-18106-1
IM-2019-18106-3
IM-2019-18106-2
Muricidae
Ocinebrina aciculata
IM-2019-16871
IM-2019-16869
IM-2019-17891
IM-2019-17899
IM-2019-16823
Nassariidae
Tritia cuvierii
IM-2019-16441
IM-2019-16452
IM-2019-17582
IM-2019-17584
IM-2019-16440
Columbellidae
Mitrella scripta
IM-2019-17210
IM-2019-17145
IM-2019-17194
IM-2019-17195
IM-2019-17193
Fasciolariidae
Pseudofusus pulchellus
IM-2019-18225
IM-2019-18224
IM-2019-17596
Chauvetiidae
Chauvetia recondita
IM-2019-17610
IM-2019-17612-4
IM-2019-17612-3
IM-2019-17612-2
IM-2019-17612-1
Chauvetiidae
Chauvetia mamillata
IM-2019-18381
IM-2019-17175
IM-2019-18382
IM-2019-17152
IM-2019-17151
Mitridae
Episcomitra cornicula
IM-2019-17856
IM-2019-16896
IM-2019-16817
IM-2019-17854
IM-2019-17698
Drilliidae
Crassopleura maravignae
IM-2019-17940
IM-2019-18096
IM-2019-18092
IM-2019-17870
IM-2019-17858
Mitromorphidae
Mitromorpha columbellaria
IM-2019-17874
IM-2019-17663
IM-2019-17208
IM-2019-17690
IM-2019-17189
Raphitomidae
Raphitoma lineolata
IM-2019-16845
IM-2019-16978
IM-2019-16963
IM-2019-16970
IM-2019-16964
Mangellidae
Mangelia multilineolata
0.1
OUTGROUP 2
OUTGROUP 1
NEOGASTROPODA

Fig. 1: Maximum Likelihood tree obtained with all the samples. Red: samples ignored in the correlation tests (contaminated, wrongly placed with a long branch); Stars at nodes: lineages for which one or several transcriptomes were used for the probe design.

2. Correlations tests

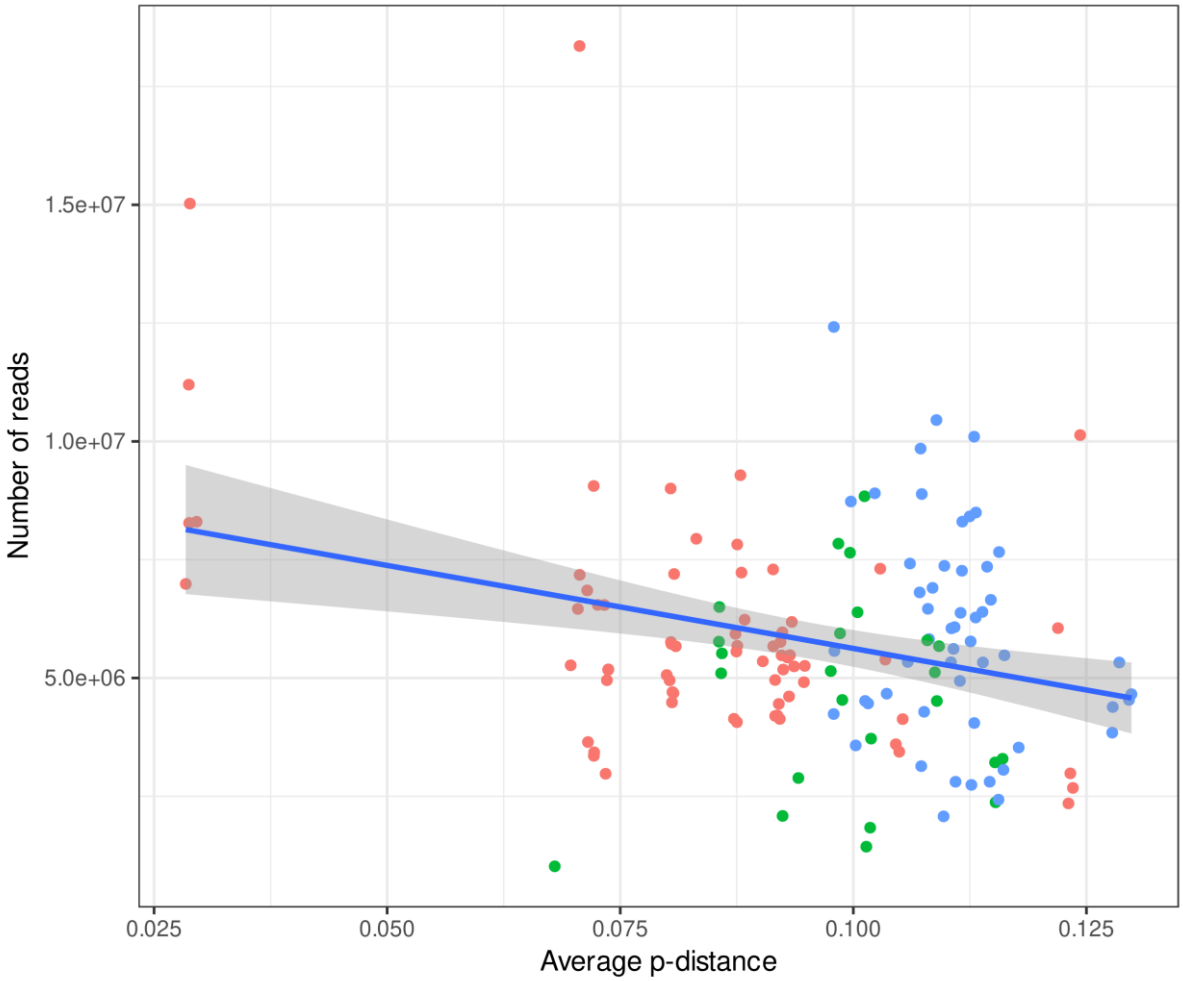


Fig. 2: Number of reads per specimen as a function of the average p-distance. Red dots: specimens of neogastropods; green dots: specimens of the outgroup 1; blue dots: specimens of the outgroup 2. Blue line: fitted linear regression line (Tau = -0.112; p-value = 0.044); Grey bands: 95% confidence interval bands.

The correlation between the number of reads and the average p-distance between the reconstructed exons and the corresponding baits for each species is weak (Tau = -0.112; p-value = 0.044; Fig. 2), with a slight decrease in the number of reads when the genetic distance increases. Indeed, the number of reads per specimen is relatively homogenous among the three

groups (neogastropods, outgroup 1 and outgroup 2). Thus, the number of reads obtained for each sample does not seem to be influenced by the genetic distance, probably because if the number of reads on target can fluctuate among the samples, the total numbers of reads is generally similar among ingroups and outgroups.

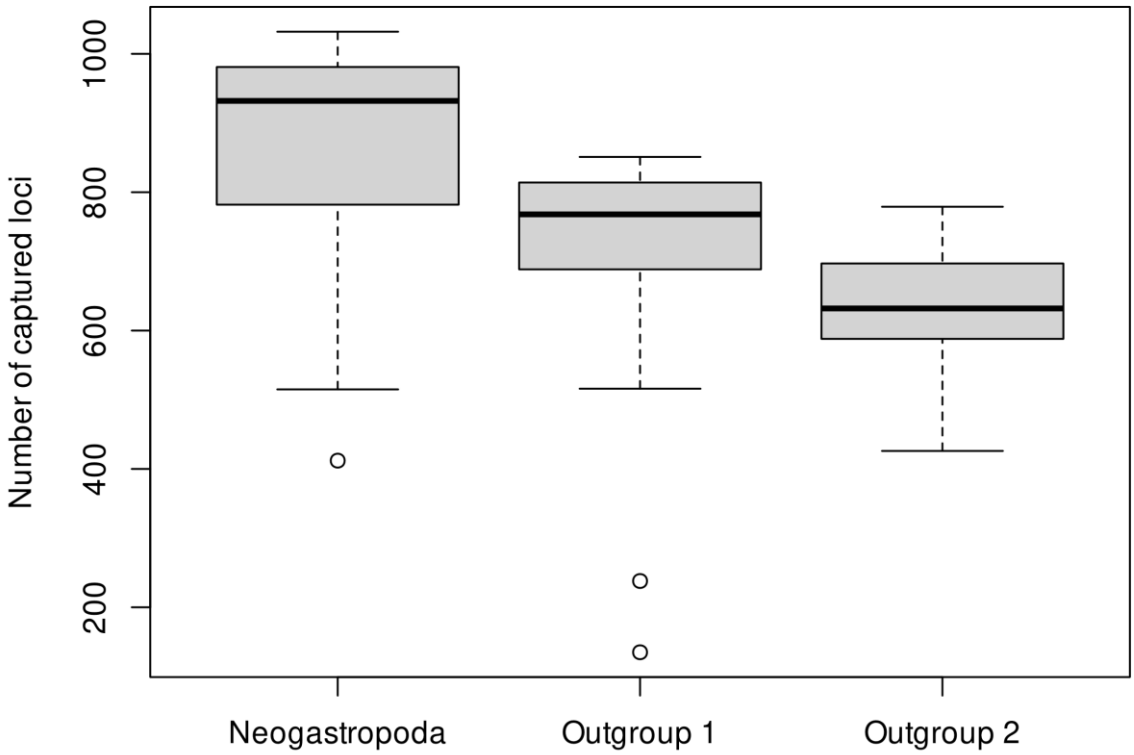


Fig. 3: Boxplot of the number of exons captured for each sample for the three taxonomic groups.

Among the three groups of specimens, the number of reconstructed exons is higher for the neogastropods, then followed by the outgroup 1 and by the outgroup 2 (Fig. 3). Only one sample among the neogastropods and two samples among the outgroup 1 are detected as outliers, with a low number of reconstructed exons. This result is also illustrated by the highly significant correlation between the number of reconstructed exons and the average p-distance between the reconstructed exons and the corresponding baits for each species (Tau = -0.593; p-value = 2.2e-16; Fig. 4): the higher the genetic distance, the lower the number of reconstructed exons. The results are similar after 1,000 randomizations: the value of Tau obtained with the complete dataset is included in the distribution obtained after randomization, indicating a weak impact of

the less variable loci on the correlation tests, and thus confirming the correlation between the number of reconstructed exons and the average p-distances (Fig. 5). Again, the exons of neogastropods, with generally lower genetic distances, are better captured than exons of outgroup 1 and outgroup 2 specimens (Fig. 4).

Our results suggest that capture efficiency decreases regularly, without a sudden decrease at 10%, as proposed in the literature (Bragg et al., 2016). However, our ingroup samples mostly have genetic distance below 10%, which seem to indicate that the use of "ancestral" sequences was efficient in reducing the genetic distances between the probes and the captured sequences, i.e. to increase the capture success.

Notably, there are a few exceptions to the general trend, with notably a neogastropod species (*Granulina marginata*, Granulinidae, with five specimens: IM-2019-16405, IM-2019-16406, IM-2019-16408, IM-2019-17926, IM-2019-17927), with higher genetic distances and a lower number of reconstructed exons. Together with the specimens of *Gibberula miliaria* (Cystiscidae), these two species represent the superfamily Volutoidea, for which one transcriptome was available (Alcithoe, Volutidae) for the bait design (although not from the same family). Thus, two lineages phylogenetically both with the same phylogenetic distance to the most closely related transcriptome used for the bait design did not end with a similar number of exons reconstructed. In another word, the phylogenetic distance is not necessarily correlated to the genetic distance between the targeted DNA and the probes, and thus belonging to a phylogenetic lineage that includes a transcriptome used for the bait design does not guarantee a high success rate in exon capture.

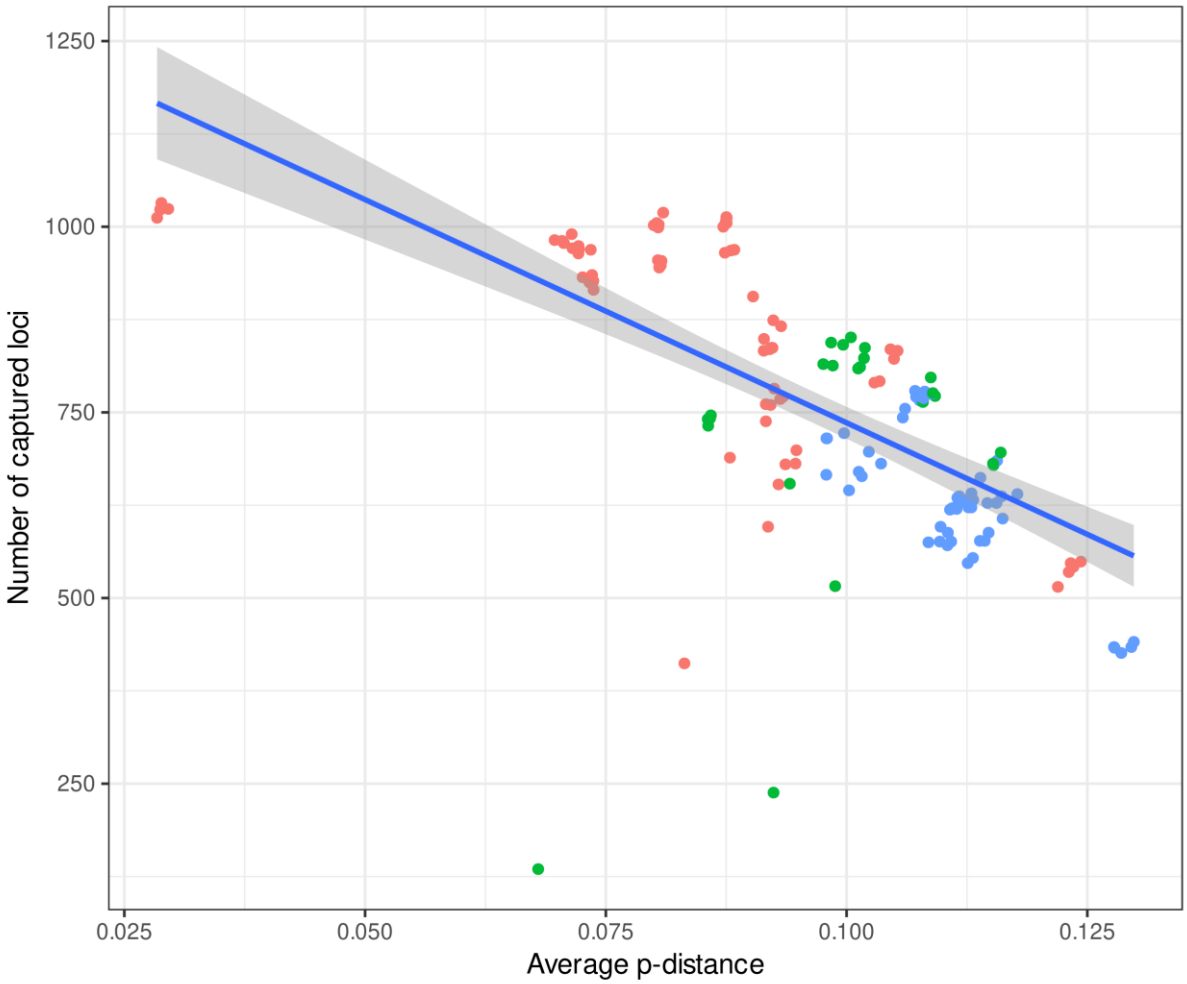


Fig. 4: Number of reconstructed exons per specimen as a function of the average p-distance. Red dots: specimens of neogastropods; green dots: specimens of the outgroup 1; blue dots: specimens of the outgroup 2. Blue line: fitted linear regression line (Tau = -0.593; p-value = 2.2e-16); Grey bands: 95% confidence interval bands.

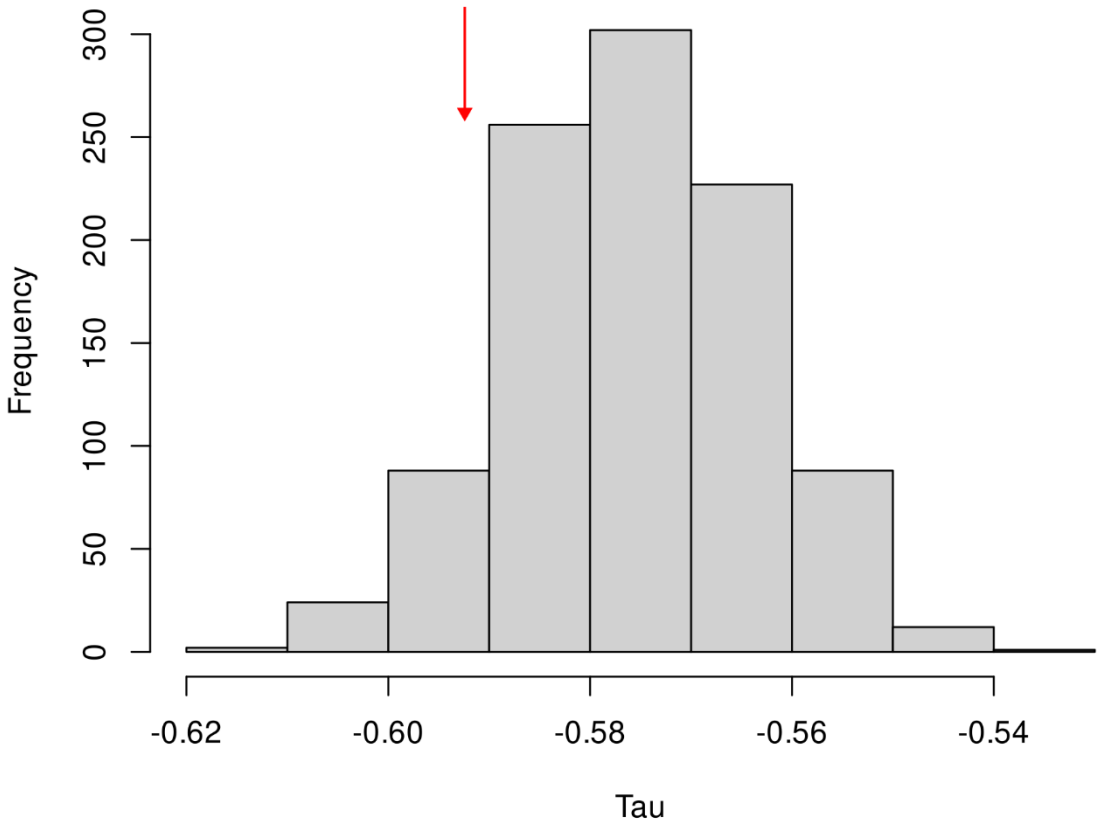


Fig.5: Distribution of the Tau values among the 1,000 replicates of randomization (random selection of 100 loci). Red arrow: Tau value of the complete dataset.

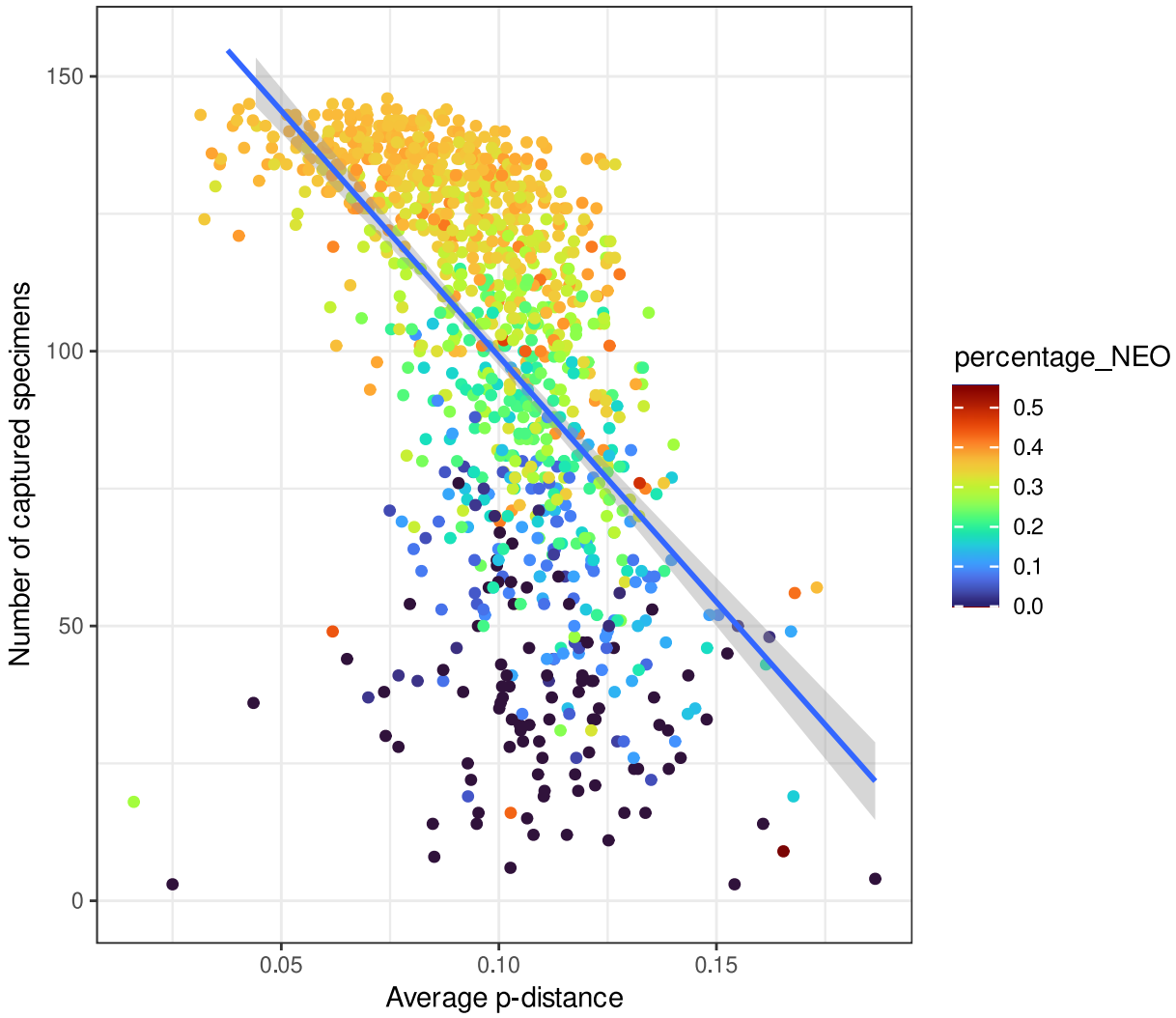


Fig. 6: Number of captured specimens (i.e. reconstructed exons) per exon as a function of the average p-distance between the captured sequence and the target *Conus* exon. Dots are colored according to the percentage of neogastropod sequences among all the reconstructed exons, for each exon, from red (all the sequences for the corresponding locus are from neogastropod specimens) to blue (all the sequences for the corresponding locus are from outgroup 1 and outgroup 2 specimens). Blue line: fitted linear regression line (Tau = -0.463; p-value = 2.2e-16); Grey bands: 95% confidence interval bands.

Again, a similar trend is observed when the number of reconstructed sequences per exon is considered, as a function of the average p-distance. The correlation is again negative, with less sequences reconstructed when the p-distance increases (Tau = -0.463; p-value = 2.2e-16; Fig. 6). The loci for which most of the samples are neogastropods are the ones with the lower number of reconstructed exons, i.e. neogastropods are more frequently captured than outgroup 1 and outgroup 2 specimens.

Conclusion

It does not seem possible to identify a clear threshold above which the exon capture efficiency significantly decreases, and thus to readily identify which lineages would necessitate the design of new baits. On the contrary, the decrease in capture efficiency is quite regular. Nonetheless, it remains obvious that designing probes using more transcriptomic and/or genomic data, from a more representative set of taxa, will help to maintain a high level of capture success. Thus, ideally, when designing a probe set for exon capture, each new lineage, significantly diverging from the previous ones, would benefit from adding new probes based on a transcriptome from this lineage.

Data availability

Newly published transcriptomes and exon sequences will be made publicly available upon acceptance of the manuscript.

Funding

The present work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 865101) to N.P.

Acknowledgments

All the specimens used for the exon capture experiments were collected during the CORSICABENTHOS 3 expedition. The CORSICABENTHOS expeditions (PIs Philippe Bouchet, Line Le Gall) are part of the MNHN "Our Planet Reviewed" programme, funded by Office Français de la Biodiversité and Collectivité Territoriale de Corse, and conducted in partnership with Université de Corse Pasquale Paoli and Office de l'Environnement de la Corse. The CORSICABENTHOS 3 expedition took place on 9-31 May 2021 and received in-kind support from Parc Naturel Régional de Corse, Réserve Naturelle des Bouches de Bonifacio, and the municipality of Porto-Ota. Sampling operated under the regulations then in force in the countries in question and satisfy the conditions set by the Nagoya Protocol for access to genetic resources. All biocomputing were done on the MNHN cluster (Plateforme de Calcul Intensif et Algorithmique PCIA, Muséum national d'histoire naturelle, Centre national de la recherche scientifique, UAR 2700 2AD, CP 26, 57 rue Cuvier, F-75231 Paris Cedex 05, France) and we are grateful to the genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12) for providing help and/or computing and/or storage resources

References

Abalde, S., Tenorio, M. J., Afonso, C. M. L., & Zardoya, R. (2017). Mitogenomic phylogeny of cone snails endemic to Senegal. *Molecular Phylogenetics and Evolution*, *112*, 79–87. https://doi.org/10.1016/j.ympev.2017.04.020

Abdelkrim, J., Aznar-Cormano, L., Fedosov, A. E., Kantor, Y. I., Lozouet, P., Phuong, M. A., Zaharias, P., & Puillandre, N. (2018). Exon-Capture-Based Phylogeny and Diversification of the Venomous Gastropods (Neogastropoda, Conoidea). *Molecular Biology and Evolution*, *35*(10), 2355–2374. hal-02002406v1. https://doi.org/10.1093/molbev/msy144

Andermann, T., Torres Jiménez, M. F., Matos-Maraví, P., Batista, R., Blanco-Pastor, J. L., Gustafsson, A. L. S., Kistler, L., Liberal, I. M., Oxelman, B., Bacon, C. D., & Antonelli, A. (2020). A Guide to Carrying Out a Phylogenomic Target Sequence Capture Project. *Frontiers in Genetics*, *10*, 1407. https://doi.org/10.3389/fgene.2019.01407

Andrews, E. B. (1991). THE FINE STRUCTURE AND FUNCTION OF THE SALIVARY GLANDS OF *NUCELLA LAPILLUS* (GASTROPODA: MURICIDAE). *Journal of Molluscan Studies*, *57*(1), 111–126. https://doi.org/10.1093/mollus/57.1.111

Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., & Pupko, T. (2012). FastML: A web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research*, *40*(W1), W580–W584. https://doi.org/10.1093/nar/gks498

Bandel, K. (1984). *THE RADULAE OF CARIBBEAN AND OTHER MESOGASTROPODA AND NEOGASTROPODA*. 199.

Bandyopadhyay, P. K., Stevenson, B. J., Cady, M. T., Olivera, B. M., & Wolstenholme, D. R. (2006). Complete mitochondrial DNA sequence of a Conoidean gastropod, Lophiotoma (Xenuroturris) cerithiformis: Gene order and gastropod phylogeny. *Toxicon*, *48*(1), 29–43. https://doi.org/10.1016/j.toxicon.2006.04.013

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Barco, A., Claremont, M., Reid, D. G., Houart, R., Bouchet, P., Williams, S. T., Cruaud, C., Couloux, A., & Oliverio, M. (2010). A molecular phylogenetic framework for the Muricidae, a diverse family of carnivorous

gastropods. *Molecular Phylogenetics and Evolution*, *56*(3), 1025–1039. 152. https://doi.org/10.1016/j.ympev.2010.03.008

Bartoš, O., Bohlen, J., Šlechtová, V. B., Kočí, J., Röslein, J., & Janko, K. (2023). Sequence capture: Obsolete or irreplaceable? A thorough validation across phylogenetic distances and its applicability to hybrids and allopolyploids. *Molecular Ecology Resources*, *23*(6), 1348–1360. https://doi.org/10.1111/1755-0998.13806

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., & Good, J. M. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, *13*(1), 403. https://doi.org/10.1186/1471-2164-13-403

Blaimer, B. B., Brady, S. G., Schultz, T. R., Lloyd, M. W., Fisher, B. L., & Ward, P. S. (2015). Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: A case study of formicine ants. *BMC Evolutionary Biology*, *15*(1), 271. https://doi.org/10.1186/s12862-015-0552-5

Blaimer, B. B., Lloyd, M. W., Guillory, W. X., & Brady, S. G. (2016). Sequence Capture and Phylogenetic Utility of Genomic Ultraconserved Elements Obtained from Pinned Insect Specimens. *PLOS ONE*, *11*(8), e0161531. https://doi.org/10.1371/journal.pone.0161531

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bondarev I. (2001). Description of a new cone species (Conus evansi) from the Red Sea, Dahlak (Gastropoda, Conidae). *La Conchiglia*, *33*(299), 25–26.

Borowiec, M. L. (2016). AMAS: A fast tool for alignment manipulation and computing of summary statistics. *PeerJ*, *4*, e1660. https://doi.org/10.7717/peerj.1660

Bose, U., Suwansa-ard, S., Maikaeo, L., Motti, C. A., Hall, M. R., & Cummins, S. F. (2017). Neuropeptides encoded within a neural transcriptome of the giant triton snail Charonia tritonis, a Crown-of-Thorns Starfish predator. *Peptides*, *98*, 3–14. https://doi.org/10.1016/j.peptides.2017.01.004

Bossert, S., Murray, E. A., Almeida, E. A. B., Brady, S. G., Blaimer, B. B., & Danforth, B. N. (2019). Combining transcriptomes and ultraconserved elements to illuminate the phylogeny of Apidae. *Molecular Phylogenetics and Evolution*, *130*, 121–131. https://doi.org/10.1016/j.ympev.2018.10.012

Bouchet, P., Héros, V., Lozouet, P., & Maestrati, P. (2008). A quarter-century of deep-sea malacological exploration in the South and West Pacific: Where do we stand? How far to go? In V. Héros, R. H. Cowie,

& P. Bouchet (Eds.), *Tropical Deep-Sea Benthos* (Vol. 25, pp. 9–40). Muséum national d'Histoire naturelle.

Bouchet, P., Kantor, Y. I., Sysoev, A. V., & Puillandre, N. (2011). A new operational classification of the Conoidea (Gastropoda). *Journal of Molluscan Studies*, *77*(3), 273–308. 160. https://doi.org/10.1093/mollus/eyr017

Bouchet, P., Lozouet, P., Maestrati, P., & Heros, V. (2002). Assessing the magnitude of species richness in tropical marine environments: Exceptionally high numbers of molluscs at a New Caledonia site. *Biological Journal of the Linnean Society*, *75*(4), 421–436. 112. https://doi.org/10.1046/j.1095-8312.2002.00052.x

Bouchet, P., Lozouet, P., & Sysoev, A. (2009). An inordinate fondness for turrids. *Deep Sea Research Part II: Topical Studies in Oceanography*, *56*(19–20), 1724–1731. 146. https://doi.org/10.1016/j.dsr2.2009.05.033

Bouchet, P., Rocroi, J.-P., Hausdorf, B., Kaim, A., Kano, Y., Nützel, A., Parkhaev, P., Schrödl, M., & Strong, E. E. (2017). Revised Classification, Nomenclator and Typification of Gastropod and Monoplacophoran Families. *Malacologia*, *61*(1–2), 1–526. hal-03929819v1. https://doi.org/10.4002/040.061.0201

Bouchet, P., & Strong, E. E. (2010). Historical name-bearing types in marine molluscs: An impediment to biodiversity studies. In A. Polaszek (Ed.), *Systema Naturae 250—The Linnaean Ark* (First Edition, pp. 63–74). CRC Press; 150. https://repository.si.edu/bitstream/handle/10088/11295/iz_Bouchet_Strong_2010_Historical_types.pdf?sequence=1&isAllowed=y

Bragg, J. G., Potter, S., Bi, K., & Moritz, C. (2016). Exon capture phylogenomics: Efficacy across scales of divergence. *Molecular Ecology Resources*, *16*(5), 1059–1068. https://doi.org/10.1111/1755-0998.12449

Branstetter, M. G., Longino, J. T., Ward, P. S., & Faircloth, B. C. (2017). Enriching the ant tree of life: Enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. *Methods in Ecology and Evolution*, *8*(6), 768–776. https://doi.org/10.1111/2041-210X.12742

Brewer, G. E., Clarkson, J. J., Maurin, O., Zuntini, A. R., Barber, V., Bellot, S., Biggs, N., Cowan, R. S., Davies, N. M. J., Dodsworth, S., Edwards, S. L., Eiserhardt, W. L., Epitawalage, N., Frisby, S., Grall, A., Kersey, P. J., Pokorny, L., Leitch, I. J., Forest, F., & Baker, W. J. (2019). Factors Affecting Targeted Sequencing of 353 Nuclear Genes From Herbarium Specimens Spanning the Diversity of Angiosperms. *Frontiers in Plant Science*, *10*, 1102. https://doi.org/10.3389/fpls.2019.01102

Cabau, C., Escudié, F., Djari, A., Guiguen, Y., Bobe, J., & Klopp, C. (2017). Compacting and correcting Trinity and Oases RNA-Seq *de novo* assemblies. *PeerJ*, *5*, e2988. https://doi.org/10.7717/peerj.2988

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Cantu, V. A., Sadural, J., & Edwards, R. (2019). *PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets* [Preprint]. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.27553v1

Choi, E. H., Choi, N. R., & Hwang, U. W. (2021). The mitochondrial genome of an Endangered freshwater snail *Koreoleptoxis nodifila* (Caenogastropoda: Semisulcospiridae) from South Korea. *Mitochondrial DNA Part B*, *6*(3), 1120–1123. https://doi.org/10.1080/23802359.2021.1901626

Choquet, M., Smolina, I., Dhanasiri, A. K. S., Blanco-Bercial, L., Kopp, M., Jueterbock, A., Sundaram, A. Y. M., & Hoarau, G. (2019). Towards population genomics in non-model species with large genomes: A case study of the marine zooplankton *Calanus finmarchicus*. *Royal Society Open Science*, *6*(2), 180608. https://doi.org/10.1098/rsos.180608

Claremont, M., Houart, R., Williams, S. T., & Reid, D. G. (2013). A molecular phylogenetic framework for the Ergalataxinae (Neogastropoda: Muricidae). *Journal of Molluscan Studies*, *79*(1), 19–29. https://doi.org/10.1093/mollus/eys028

Claremont, M., Reid, D. G., & Williams, S. T. (2011). Evolution of corallivory in the gastropod genus Drupella. *Coral Reefs*, *30*(4), 977–990. https://doi.org/10.1007/s00338-011-0788-5

Claremont, M., Vermeij, G. J., Williams, S. T., & Reid, D. G. (2013). Global phylogeny and new classification of the Rapaninae (Gastropoda: Muricidae), dominant molluscan predators on tropical rocky seashores. *Molecular Phylogenetics and Evolution*, *66*(1), 91–102. https://doi.org/10.1016/j.ympev.2012.09.014

Colgan, D. J., Ponder, W. F., Beacham, E., & Macaranas, J. (2007). Molecular phylogenetics of Caenogastropoda (Gastropoda: Mollusca). *Molecular Phylogenetics and Evolution*, *42*(3), 717–737. https://doi.org/10.1016/j.ympev.2006.10.009

Couto, D. R., Bouchet, P., Kantor, Y. I., Simone, L. R. L., & Giribet, G. (2016). A multilocus molecular phylogeny of Fasciolariidae (Neogastropoda: Buccinoidea). *Molecular Phylogenetics and Evolution*, *99*, 309–322. https://doi.org/10.1016/j.ympev.2016.03.025

Criscione, F., Hallan, A., Puillandre, N., & Fedosov, A. (2021). Where the snails have no name: A molecular phylogeny of Raphitomidae (Neogastropoda: Conoidea) uncovers vast unexplored diversity in the deep seas of temperate southern and eastern Australia. *Zoological Journal of the Linnean Society*, *191*(4), 961–1000. hal-02970382v1. https://doi.org/10.1093/zoolinnean/zlaa088

Cunha, R. L., Castilho, R., Rüber, L., & Zardoya, R. (2005). Patterns of Cladogenesis in the Venomous Marine Gastropod Genus Conus from the Cape Verde Islands. *Systematic Biology*, *54*(4), 634–650. https://doi.org/10.1080/106351591007471

Cunha, R. L., Grande, C., & Zardoya, R. (2009). Neogastropod phylogenetic relationships based on entire mitochondrial genomes. *BMC Evolutionary Biology*, *9*(1), 210. https://doi.org/10.1186/1471-2148-9-210

Cunha, R. L., Tenorio, M. J., Afonso, C., Castilho, R., & Zardoya, R. (2007). Replaying the tape: Recurring biogeographical patterns in Cape Verde Conus after 12 million years: RECURRING BIOGEOGRAPHICAL PATTERNS IN CAPE VERDE CONUS. *Molecular Ecology*, *17*(3), 885–901. https://doi.org/10.1111/j.1365-294X.2007.03618.x

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., & Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences*, *110*(39), 15758–15763. https://doi.org/10.1073/pnas.1314445110

deMaintenon, M. J. (1999). Phylogenetic Analysis of the Columbellidae (Mollusca: Neogastropoda) and the Evolution of Herbivory from Carnivory. *Invertebrate Biology*, *118*(3), 258. https://doi.org/10.2307/3226997

Der Sarkissian, C., Möller, P., Hofman, C. A., Ilsøe, P., Rick, T. C., Schiøtte, T., Sørensen, M. V., Dalén, L., & Orlando, L. (2020). Unveiling the Ecological Applications of Ancient DNA From Mollusk Shells. *Frontiers in Ecology and Evolution*, *8*, 37. https://doi.org/10.3389/fevo.2020.00037

Der Sarkissian, C., Pichereau, V., Dupont, C., Ilsøe, P. C., Perrigault, M., Butler, P., Chauvaud, L., Eiríksson, J., Scourse, J., Paillard, C., & Orlando, L. (2017). Ancient DNA analysis identifies marine mollusc shells as new metagenomic archives of the past. *Molecular Ecology Resources*, *17*(5), 835–853. https://doi.org/10.1111/1755-0998.12679

Derkarabetian, S., Benavides, L. R., & Giribet, G. (2019). Sequence capture phylogenomics of historical ethanol-preserved museum specimens: Unlocking the rest of the vault. *Molecular Ecology Resources*, *19*(6), 1531–1544. https://doi.org/10.1111/1755-0998.13072

Duda, T. F., & Kohn, A. J. (2005). Species-level phylogeography and evolutionary history of the hyperdiverse marine gastropod genus Conus. *Molecular Phylogenetics and Evolution*, *34*(2), 257–272. https://doi.org/10.1016/j.ympev.2004.09.012

Duda, T. F., Kohn, A. J., & Palumbi, S. R. (2001). Origins of diverse feeding ecologies within Conus, a genus of venomous marine gastropods. *Biological Journal of the Linnean Society*, *73*(4), 391–409. https://doi.org/10.1111/j.1095-8312.2001.tb01369.x

Duda, T. F., & Lee, T. (2009). Ecological Release and Venom Evolution of a Predatory Marine Snail at Easter Island. *PLoS ONE*, *4*(5), e5558. https://doi.org/10.1371/journal.pone.0005558

Duda, T. F., & Palumbi, S. R. (2004). Gene expression and feeding ecology: Evolution of piscivory in the venomous gastropod genus *Conus*. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *271*(1544), 1165–1174. https://doi.org/10.1098/rspb.2004.2708

Duda, T. F., & Rolán, E. (2004). Explosive radiation of Cape Verde Conus, a marine species flock: MARINE SPECIES FLOCK IN CAPE VERDE. *Molecular Ecology*, *14*(1), 267–272. https://doi.org/10.1111/j.1365-294X.2004.02397.x

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Espiritu, D. J. D., Watkins, M., Dia-Monje, V., Cartier, G. E., Cruz, L. J., & Olivera, B. M. (2001). Venomous cone snails: Molecular phylogeny and the generation of toxin diversity. *Toxicon*, *39*(12), 1899–1916. https://doi.org/10.1016/S0041-0101(01)00175-1

Faircloth, B. C., Branstetter, M. G., White, N. D., & Brady, S. G. (2015). Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among H ymenoptera. *Molecular Ecology Resources*, *15*(3), 489–501. https://doi.org/10.1111/1755-0998.12328

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. *Systematic Biology*, *61*(5), 717–726. https://doi.org/10.1093/sysbio/sys004

Farhat, S., Modica, M. V., & Puillandre, N. (2023). Whole Genome Duplication and Gene Evolution in the Hyperdiverse Venomous Gastropods. *Molecular Biology and Evolution*, *40*(8), msad171. https://doi.org/10.1093/molbev/msad171

Fassio, G., Modica, M. V., Mary, L., Zaharias, P., Fedosov, A. E., Gorson, J., Kantor, Y. I., Holford, M., & Puillandre, N. (2019). Venom Diversity and Evolution in the Most Divergent Cone Snail Genus Profundiconus. *Toxins*, *11*(11), 623. hal-02430744v1. https://doi.org/10.3390/toxins11110623

Fassio, G., Russini, V., Pusateri, F., Giannuzzi-Savelli, R., Høisæter, T., Puillandre, N., Modica, M. V., & Oliverio, M. (2019). An assessment of *Raphitoma* and allied genera (Neogastropoda: Raphitomidae). *Journal of Molluscan Studies*, *85*(4), 413–424. hal-02970454v1. https://doi.org/10.1093/mollus/eyz022

Fedesov, A. E., & Kantor, Y. I. (2008). Toxoglossan gastropods of the subfamily Crassispirinae (Turridae) lacking a radula, and a discussion of the status of the subfamily Zemaciinae. *Journal of Molluscan Studies*, *74*(1), 27–35. https://doi.org/10.1093/mollus/eym042

Fedosov, A. E., Caballer Gutierrez, M., Buge, B., Sorokin, P. V., Puillandre, N., & Bouchet, P. (2019). Mapping the missing branch on the neogastropod tree of life: Molecular phylogeny of marginelliform gastropods. *Journal of Molluscan Studies*, *85*(4), 439–451. hal-02559712v1. https://doi.org/10.1093/mollus/eyz028

Fedosov, A. E., Malcolm, G., Terryn, Y., Gorson, J., Modica, M. V., Holford, M., & Puillandre, N. (2019). Phylogenetic classification of the family Terebridae (Neogastropoda: Conoidea). *Journal of Molluscan Studies*, *85*(4), 359–388. hal-02559725v1. https://doi.org/10.1093/mollus/eyz004

Fedosov, A. E., & Puillandre, N. (2012). Phylogeny and taxonomy of the *Kermia–Pseudodaphnella* (Mollusca: Gastropoda: Raphitomidae) genus complex: a remarkable radiation via diversification of larval development. *Systematics and Biodiversity*, *10*(4), 447–477. https://doi.org/10.1080/14772000.2012.753137

Fedosov, A. E., Zaharias, P., Lemarcis, T., Modica, M. V., Holford, M., Oliverio, M., Kantor, Y. I., & Puillandre, N. (in press). *Phylogenomics of Neogastropoda: The backbone hidden in the bush.*

Fedosov, A., & Kantor, Y. (2007). Toxoglossan gastropods of the subfamily Crassispirinae (Turridae) lacking a radula, and a discussion of the status of the subfamily Zemaciinae. *Journal of Molluscan Studies*, *74*(1), 27–35. https://doi.org/10.1093/mollus/eym042

Fedosov, A., Puillandre, N., Herrmann, M., Kantor, Y., Oliverio, M., Dgebuadze, P., Modica, M. V., & Bouchet, P. (2018). The collapse of Mitra: Molecular systematics and morphology of the Mitridae (Gastropoda:

Neogastropoda). *Zoological Journal of the Linnean Society*, *183*(2), 253–337. hal-03926162. https://doi.org/10.1093/zoolinnean/zlx073

Fedosov, A., Puillandre, N., Kantor, Y., & Bouchet, P. (2015). Phylogeny and systematics of mitriform gastropods (Mollusca: Gastropoda: Neogastropoda): Phylogeny of Mitriform Gastropods. *Zoological Journal of the Linnean Society*, *175*(2), 336–359. https://doi.org/10.1111/zoj.12278

Ferreira, S., Ashby, R., Jeunen, G.-J., Rutherford, K., Collins, C., Todd, E. V., & Gemmell, N. J. (2020). DNA from mollusc shell: A valuable and underutilised substrate for genetic analyses. *PeerJ*, *8*, e9420. https://doi.org/10.7717/peerj.9420

Forrest, L. L., Hart, M. L., Hughes, M., Wilson, H. P., Chung, K.-F., Tseng, Y.-H., & Kidner, C. A. (2019). The Limits of Hyb-Seq for Herbarium Specimens: Impact of Preservation Techniques. *Frontiers in Ecology and Evolution*, *7*, 439. https://doi.org/10.3389/fevo.2019.00439

Fourdrilis, S., de Frias Martins, A. M., & Backeljau, T. (2018). Relation between mitochondrial DNA hyperdiversity, mutation rate and mitochondrial genome evolution in Melarhaphe neritoides (Gastropoda: Littorinidae) and other Caenogastropoda. *Scientific Reports*, *8*(1), 17964. https://doi.org/10.1038/s41598-018-36428-7

Galindo, L. A., Puillandre, N., Strong, E. E., & Bouchet, P. (2014). Using microwaves to prepare gastropods for DNA barcoding. *Molecular Ecology Resources*, *14*(4), 700–705. 176. https://doi.org/10.1111/1755-0998.12231

Galindo, L. A., Puillandre, N., Utge, J., Lozouet, P., & Bouchet, P. (2016). The phylogeny and systematics of the Nassariidae revisited (Gastropoda, Buccinoidea). *Molecular Phylogenetics and Evolution*, *99*, 337–353. https://doi.org/10.1016/j.ympev.2016.03.019

Gamba, C., Hanghøj, K., Gaunitz, C., Alfarhan, A. H., Alquraishi, S. A., Al-Rasheid, K. A. S., Bradley, D. G., & Orlando, L. (2016). Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. *Molecular Ecology Resources*, *16*(2), 459–469. https://doi.org/10.1111/1755-0998.12470

Gamba, C., Jones, E. R., Teasdale, M. D., McLaughlin, R. L., Gonzalez-Fortes, G., Mattiangeli, V., Domboróczki, L., Kővári, I., Pap, I., Anders, A., Whittle, A., Dani, J., Raczky, P., Higham, T. F. G., Hofreiter, M., Bradley, D. G., & Pinhasi, R. (2014). Genome flux and stasis in a five millennium transect of European prehistory. *Nature Communications*, *5*(1), 5257. https://doi.org/10.1038/ncomms6257

Geist, J., Wunderlich, H., & Kuehn, R. (2008). Use of mollusc shells for DNA-based molecular analyses. *Journal of Molluscan Studies*, *74*(4), 337–343. https://doi.org/10.1093/mollus/eyn025

Goulding, T. C., Yeung, N. W., & Hayes, K. A. (2021). Historical DNA from Museum Shell Collections: Evaluating the Suitability of Dried Micromollusks for Molecular Systematics. *American Malacological Bulletin*, *38*(2). https://doi.org/10.4003/006.038.0209

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Harasewych, M. G., Sei, M., Wirshing, H. H., & Uribe, J. E. (2019). The complete mitochondrial genome of Neptuneopsis gilchristi G.B. Sowerby III, 1898 (Neogastropoda: Volutidae: Calliotectinae). *THE NAUTILUS*, *133*, 7.

Hayashi, S. (2005). The molecular phylogeny of the Buccinidae (Caenogastropoda: Neogastropoda). *MOLLUSCAN RESEARCH*, *25*, 14.

Hedtke, S. M., Morgan, M. J., Cannatella, D. C., & Hillis, D. M. (2013). Targeted Enrichment: Maximizing Orthologous Gene Comparisons across Deep Evolutionary Time. *PLoS ONE*, *8*(7), e67908. https://doi.org/10.1371/journal.pone.0067908

Holford, M., Puillandre, N., Terryn, Y., Cruaud, C., Olivera, B., & Bouchet, P. (2009). Evolution of the Toxoglossa Venom Apparatus as Inferred by Molecular Phylogeny of the Terebridae. *Molecular Biology and Evolution*, *26*(1), 15–25. 143. https://doi.org/10.1093/molbev/msn211

Huang, X., & Madan, A. (1999). CAP3: A DNA Sequence Assembly Program. *Genome Research*, *9*(9), 868–877. https://doi.org/10.1101/gr.9.9.868

Hugall, A. F., O'Hara, T. D., Hunjan, S., Nilsen, R., & Moussalli, A. (2016). An Exon-Capture System for the Entire Class Ophiuroidea. *Molecular Biology and Evolution*, *33*(1), 281–294. https://doi.org/10.1093/molbev/msv216

Hunter, J. P. (1998). Key innovations and the ecology of macroevolution. *Trends in Ecology & Evolution*, *13*(1), 31–36. https://doi.org/10.1016/S0169-5347(97)01273-1

Ilves, K. L., & López-Fernández, H. (2014). A targeted next-generation sequencing toolkit for exon-based cichlid phylogenomics. *Molecular Ecology Resources*, *14*(4), 802–811. https://doi.org/10.1111/1755-0998.12222

Inäbnit, T., Jochum, A., Slapnik, R., & Neubert, E. (2021). New genetic data reveals a new species of Zospeum in Bosnia (Gastropoda, Ellobioidea, Carychiinae). *ZooKeys*, *1071*, 175–193. https://doi.org/10.3897/zookeys.1071.66417

Jacobs, D. K., & Lindberg, D. R. (1998). Oxygen and evolutionary patterns in the sea: Onshore/offshore trends and recent recruitment of deep-sea faunas. *Proceedings of the National Academy of Sciences*, *95*(16), 9396–9401. https://doi.org/10.1073/pnas.95.16.9396

Jiang, J., Yuan, H., Zheng, X., Wang, Q., Kuang, T., Li, J., Liu, J., Song, S., Wang, W., Cheng, F., Li, H., Huang, J., & Li, C. (2019). Gene markers for exon capture and phylogenomics in ray-finned fishes. *Ecology and Evolution*, *9*(7), 3973–3983. https://doi.org/10.1002/ece3.5026

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, *25*(1), 185–202. https://doi.org/10.1111/mec.13304

Kantor, Y., Fedosov, A. E., Puillandre, N., Bonillo, C., & Bouchet, P. (2017). Returning to the roots: Morphology, molecular phylogeny and classification of the Olivoidea (Gastropoda: Neogastropoda). *Zoological Journal of the Linnean Society*, *180*(3), 493–541. hal-03921031v1. https://doi.org/10.1093/zoolinnean/zlw003

Kantor, Y. I. (n.d.). *ANATOMICAL BASIS FOR THE ORIGIN AND EVOLUTION OF THE TOXOGLOSSAN MODE OF FEEDING*. 16.

Kantor, Y. I., & Fedosov, A. (2009). Morphology and development of the valve of Leiblein: Possible evidence for paraphyly of the Neogastropoda. *THE NAUTILUS*, *123*(3), 10.

Kantor, Y. I., Fedosov, A. E., Kosyan, A. R., Puillandre, N., Sorokin, P. A., Kano, Y., Clark, R., & Bouchet, P. (2022). Molecular phylogeny and revised classification of the Buccinoidea (Neogastropoda). *Zoological Journal of the Linnean Society*, *194*(3), 789–857. hal-03321428v1. https://doi.org/10.1093/zoolinnean/zlab031

Kantor, Y. I., Lozouet, P., Puillandre, N., & Bouchet, P. (2014). Lost and found: The Eocene family Pyramimitridae (Neogastropoda) discovered in the Recent fauna of the Indo-Pacific. *Zootaxa*, *3754*(3), 239–276. 175. https://doi.org/10.11646/zootaxa.3754.3.2

Kantor, Y. I., & Puillandre, N. (2021). Rare, deep-water and similar: Revision of Sibogasyrinx (Conoidea: Cochlespiridae). *European Journal of Taxonomy*, *773*, 19–60. hal-03360999v1. https://doi.org/10.5852/ejt.2021.773.1509

Kantor, Y. I., Puillandre, N., Fraussen, K., Fedosov, A., & Bouchet, P. (2013). Deep-water Buccinidae (Gastropoda: Neogastropoda) from sunken wood, vents and seeps: molecular phylogeny and taxonomy. *Journal of the Marine Biological Association of the United Kingdom*, *93*(8), 2177–2195. 173. https://doi.org/10.1017/S0025315413000672

Kantor, Y. I., Puillandre, N., Rivasseau, A., & Bouchet, P. (2012). Neither a buccinid nor a turrid: A new family of deep-sea snails for Belomitra P. Fischer, 1883 (Mollusca, Neogastropoda), with a review of Recent Indo-Pacific species. *Zootaxa*, *3496*(1), 1–64. 167. https://doi.org/10.11646/zootaxa.3496.1.1

Kantor, Y. I., Strong, E. E., & Puillandre, N. (2012). A new lineage of Conoidea (Gastropoda: Neogastropoda) revealed by morphological and molecular data. *Journal of Molluscan Studies*, *78*(3), 246–255. https://doi.org/10.1093/mollus/eys007

Kantor, Y. I., & Taylor, I. D. (n.d.). *Foregut anatomy and relationships of raphitomine gastropods (Gastropoda: Conoidea: Raphitominae)*. 1.

Kantor, Y. I., & Taylor, J. D. (1991). Evolution of the toxoglossan feeding mechanism: New information on the use of the radula. *Journal of Molluscan Studies*, *57*(1), 129–134. https://doi.org/10.1093/mollus/57.1.129

Kantor, Y., Sirenko, B., Zvonareva, S. S., & Fedosov, A. (2022). Taxonomic status of genera of Buccininae (Neogastropoda, Buccinidae) updated based on molecular data with description of new species and corrections of nomenclature of Buccinum. *European Journal of Taxonomy*, *817*, 11–34. hal-03949236v1. https://doi.org/10.5852/ejt.2022.817.1759

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kendall, M. G. (2024). *A New Measure of Rank Correlation*.

Kocot, K. M., Cannon, J. T., Todt, C., Citarella, M. R., Kohn, A. B., Meyer, A., Santos, S. R., Schander, C., Moroz, L. L., Lieb, B., & Halanych, K. M. (2011). Phylogenomics reveals deep molluscan relationships. *Nature*, *477*, 452–456. https://doi.org/10.1038/nature10382

Kraus, N. J., Corneli, P. S., Watkins, M., Bandyopadhyay, P. K., Seger, J., & Olivera, B. M. (2011). Against expectation: A short sequence with high signal elucidates cone snail phylogeny. *Molecular Phylogenetics and Evolution*, *58*(2), 383–389. https://doi.org/10.1016/j.ympev.2010.11.020

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lemarcis, T., Fedosov, A. E., Kantor, Y. I., Abdelkrim, J., Zaharias, P., & Puillandre, N. (2022). Neogastropod (Mollusca, Gastropoda) phylogeny: A step forward with mitogenomes. *Zoologica Scripta*, *51*(5), 550–561. hal-03709615v1. https://doi.org/10.1111/zsc.12552

Lemoine, F., & Gascuel, O. (2021). Gotree/Goalign: Toolkit and Go API to facilitate the development of phylogenetic workflows. *NAR Genomics and Bioinformatics*, *3*(3), lqab075. https://doi.org/10.1093/nargab/lqab075

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., Lin, D., Fang, H., Zhu, A., & Gao, Y. (2010). Species identification and phylogenetic analysis of genus Nassarius (Nassariidae) based on mitochondrial genes. *Chinese Journal of Oceanology and Limnology*, *28*(3), 565–572. https://doi.org/10.1007/s00343-010-9031-4

Li, R., Bekaert, M., Wu, L., Mu, C., Song, W., Migaud, H., & Wang, C. (2019). Transcriptomic Analysis of Marine Gastropod Hemifusus tuba Provides Novel Insights into Conotoxin Genes. *Marine Drugs*, *17*(8), 466. https://doi.org/10.3390/md17080466

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Lorion, J., Duperron, S., Gros, O., Cruaud, C., & Samadi, S. (2009). Several deep-sea mussels and their associated symbionts are able to live both on wood and on whale falls. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1654), 177–185. https://doi.org/10.1098/rspb.2008.1101

Machkour-M'Rabet, S., Hanes, M. M., Martínez-Noguez, J. J., Cruz-Medina, J., & García-De León, F. J. (2021). The queen conch mitogenome: Intra- and interspecific mitogenomic variability in Strombidae and phylogenetic considerations within the Hypsogastropoda. *Scientific Reports*, *11*(1), 11972. https://doi.org/10.1038/s41598-021-91224-0

Magoc, T., & Salzberg, S. L. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, *27*(21), 2957–2963. https://doi.org/10.1093/bioinformatics/btr507

McCormack, J. E., Tsai, W. L. E., & Faircloth, B. C. (2016). Sequence capture of ultraconserved elements from bird museum specimens. *Molecular Ecology Resources*, *16*(5), 1189–1203. https://doi.org/10.1111/1755-0998.12466

Meyer, M., & Kircher, M. (2010). Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harbor Protocols*, *2010*(6), pdb.prot5448. https://doi.org/10.1101/pdb.prot5448

Mirarab, S., Reaz, R., Bayzid, Md. S., Zimmermann, T., Swenson, M. S., & Warnow, T. (2014). ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics*, *30*(17), i541–i548. https://doi.org/10.1093/bioinformatics/btu462

Modica, M. V., Bouchet, P., Cruaud, C., Utge, J., & Oliverio, M. (2011). Molecular phylogeny of the nutmeg shells (Neogastropoda, Cancellariidae). *Molecular Phylogenetics and Evolution*, *59*(3), 685–697. 158. https://doi.org/10.1016/j.ympev.2011.03.022

Modica, M. V., Kosyan, A. R., & Oliverio, M. (2009). The relationships of the enigmatic gastropod Tritonoharpa (Neogastropoda): New data on early neogastropod evolution? *THE NAUTILUS*, *123*(3).

Modica, M. V., Lombardo, F., Franchini, P., & Oliverio, M. (2015). The venomous cocktail of the vampire snail Colubraria reticulata (Mollusca, Gastropoda). *BMC Genomics*, *16*(441), 1–21. https://doi.org/10.1186/s12864-015-1648-4

Modica, M. V., Reinoso Sánchez, J., Pasquadibisceglie, A., Oliverio, M., Mariottini, P., & Cervelli, M. (2018). Anti-haemostatic compounds from the vampire snail Cumia reticulata: Molecular cloning and in-silico structure-function analysis. *Computational Biology and Chemistry*, *75*, 168–177. https://doi.org/10.1016/j.compbiolchem.2018.05.014

Modica, M.-V., Verhecken, A., & Oliverio, M. (2011). The relationships of the enigmatic neogastropod Loxotaphrus (Cancellariidae). *New Zealand Journal of Geology and Geophysics*, *54*(1), 115–124. https://doi.org/10.1080/00288306.2011.537610

Moles, J., & Giribet, G. (2021). A polyvalent and universal tool for genomic studies in gastropod molluscs (Heterobranchia). *Molecular Phylogenetics and Evolution*, *155*, 106996. https://doi.org/10.1016/j.ympev.2020.106996

Morton, B. (2003). [No title found]. *Molluscan Research*, *23*(3), 239. https://doi.org/10.1071/MR03008

Morton, B., & Jones, D. S. (2003). THE DIETARY PREFERENCES OF A SUITE OF CARRION-SCAVENGING GASTROPODS(NASSARIIDAE, BUCCINIDAE) IN PRINCESS ROYAL HARBOUR, ALBANY, WESTERNAUSTRALIA. *Journal of Molluscan Studies*, *69*(2), 151–156. https://doi.org/10.1093/mollus/69.2.151

Nam, H. H., Corneli, P. S., Watkins, M., Olivera, B., & Bandyopadhyay, P. (2009). Multiple genes elucidate the evolution of venomous snail-hunting Conus species. *Molecular Phylogenetics and Evolution*, *53*(3), 645–652. https://doi.org/10.1016/j.ympev.2009.07.013

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Nunes, R., Storer, C., Doleck, T., Kawahara, A. Y., Pierce, N. E., & Lohman, D. J. (2022). Predictors of sequence capture in a large-scale anchored phylogenomics project. *Frontiers in Ecology and Evolution*, *10*, 943361. https://doi.org/10.3389/fevo.2022.943361

Olivera, B. M., Fedosov, A., Imperial, J. S., & Kantor, Y. (2017). Physiology of Envenomation by Conoidean Gastropods. In S. Saleuddin & S. Mukai (Eds.), *Physiology of Molluscs* (1st ed., Vol. 1, pp. 153–188). Apple Academic Press; hal-03943153v1. https://doi.org/10.1201/9781315207124-5

Oliverio, M. (2009). Diversity of Coralliophilinae (Mollusca, Neogastropoda, Muricidae) at Austral Islands (South Pacific). *Zoosystema*, *31*(4), 759–789.

Oliverio, M., Cervelli, M., & Mariottini, P. (2002). ITS2 rRNA evolution and its congruence with the phylogeny of muricid neogastropods (Caenogastropoda, Muricoidea). *Molecular Phylogenetics and Evolution*, *25*(1), 63–69. https://doi.org/10.1016/S1055-7903(02)00227-0

Oliverio, M., & Modica, M. V. (2010). Relationships of the haematophagous marine snail *Colubraria* (Rachiglossa: Colubrariidae), within the neogastropod phylogenetic framework. *Zoological Journal of the Linnean Society*, *158*(4), 779–800. https://doi.org/10.1111/j.1096-3642.2009.00568.x

Osca, D., Templado, J., & Zardoya, R. (2015). Caenogastropod mitogenomics. *Molecular Phylogenetics and Evolution*, *93*, 118–128. https://doi.org/10.1016/j.ympev.2015.07.011

Pardos-Blas, J. R., Irisarri, I., Abalde, S., Afonso, C. M. L., Tenorio, M. J., & Zardoya, R. (2021). The genome of the venomous snail *Lautoconus ventricosus* sheds light on the origin of conotoxin diversity. *GigaScience*, *10*(5), giab037. https://doi.org/10.1093/gigascience/giab037

Pereira, C. M., Rosado, J., Seabra, S. G., Pina-Martins, F., Paulo, O. S., & Fonseca, P. J. (2010). *Conus pennaceus*: A phylogenetic analysis of the Mozambican molluscan complex. *African Journal of Marine Science*, *32*(3), 591–599. https://doi.org/10.2989/1814232X.2010.538157

Phuong, M. A., Alfaro, M. E., Mahardika, G. N., Marwoto, R. M., Prabowo, R. E., von Rintelen, T., Vogt, P. W. H., Hendricks, J. R., & Puillandre, N. (2019). Lack of Signal for the Impact of Conotoxin Gene Diversity on Speciation Rates in Cone Snails. *Systematic Biology*, *68*(5), 781–796. hal-02343430v1. https://doi.org/10.1093/sysbio/syz016

Phuong, M. A., & Mahardika, G. N. (2018). Targeted sequencing of venom genes from cone snail genomes improves understanding of conotoxin molecular evolution. *Molecular Biology and Evolution*, *35*(5), 1210–1224.

Phuong, M. A., Mahardika, G. N., & Alfaro, M. E. (2016). Dietary breadth is positively correlated with venom complexity in cone snails. *BMC Genomics*, *17*(1), 401. https://doi.org/10.1186/s12864-016-2755-6

Ponder, W. F., Colgan, D. J., Healy, J. M., Alexander, N., Simone, L. R. L., & Mielke, E. E. (2008). Caenogastropoda. In W. Ponder (Ed.), *Phylogeny and Evolution of the Mollusca* (pp. 331–383). University of California Press. https://doi.org/10.1525/california/9780520250925.003.0013

Ponder, W. F., & Lindberg, D. R. (1997). Towards a phylogeny of gastropod molluscs: An analysis using morphological characters. *Zoological Journal of the Linnean Society*, *119*(2), 83–265. https://doi.org/10.1111/j.1096-3642.1997.tb00137.x

Portik, D. M., Smith, L. L., & Bi, K. (2016). An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). *Molecular Ecology Resources*, *16*(5), 1069–1083. https://doi.org/10.1111/1755-0998.12541

Psonis, N., Vardinoyannis, K., & Poulakakis, N. (2022). High-throughput degraded DNA sequencing of subfossil shells of a critically endangered stenoendemic land snail in the Aegean. *Molecular Phylogenetics and Evolution*, *175*, 107561. https://doi.org/10.1016/j.ympev.2022.107561

Puillandre, N., Bouchet, P., Duda, T. F., Kauferstein, S., Kohn, A. J., Olivera, B. M., Watkins, M., & Meyer, C. (2014). Molecular phylogeny and evolution of the cone snails (Gastropoda, Conoidea). *Molecular Phylogenetics and Evolution*, *78*, 290–303. https://doi.org/10.1016/j.ympev.2014.05.023

Puillandre, N., Duda, T. F., Meyer, C. P., Olivera, B. M., & Bouchet, P. (2015). One, four or 100 genera? A new classification of the cone snails. *Journal of Molluscan Studies*, *81*(1), 1–23. https://doi.org/10.1093/mollus/eyu055

Puillandre, N., Kantor, Y. I., Sysoev, A. V., Couloux, A., Meyer, C. P., Rawlings, T., Todd, J. A., & Bouchet, P. (2011). The dragon tamed? A molecular phylogeny of the Conoidea (Gastropoda). *Journal of Molluscan Studies*, *77*(3), 259–272. 159. https://doi.org/10.1093/mollus/eyr015

Russini, V., Fassio, G., Modica, M. V., deMaintenon, M. J., & Oliverio, M. (2017). An assessment of the genus *Columbella* Lamarck, 1799 (Gastropoda: Columbellidae) from eastern Atlantic. *Zoosystema*, *39*(2), 197–212. https://doi.org/10.5252/z2017n2a2

Russini, V., Fassio, G., Nocella, E., HOUART, R., Barco, A., Puillandre, N., Lozouet, P., Modica, M. V., & Oliverio, M. (2023). Whelks, rock-snails, and allied: A new phylogenetic framework for the family Muricidae (Mollusca: Gastropoda). *The European Zoological Journal*, *90(2)*, 856–868. https://doi.org/doi.org/10.1080/24750263.2023.2283517

Ryu, T., Seridi, L., & Ravasi, T. (2012). The evolution of ultraconserved elements with different phylogenetic origins. *BMC Evolutionary Biology*, *12*(1), 236. https://doi.org/10.1186/1471-2148-12-236

Shokralla, S., Gibson, J. F., Nikbakht, H., Janzen, D. H., Hallwachs, W., & Hajibabaei, M. (2014). Next-generation DNA barcoding: Using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular Ecology Resources*, *14*(5), 892–901. https://doi.org/10.1111/1755-0998.12236

Silva, P. C., Malabarba, M. C., Vari (In memoriam), R., & Malabarba, L. R. (2019). Comparison and optimization for DNA extraction of archived fish specimens. *MethodsX*, *6*, 1433–1442. https://doi.org/10.1016/j.mex.2019.06.001

Simakov, O., Marletaz, F., Cho, S.-J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo, D.-H., Larsson, T., Lv, J., Arendt, D., Savage, R., Osoegawa, K., De Jong, P., Grimwood, J., Chapman, J. A., Shapiro, H., Aerts, A., Otillar, R. P., Terry, A. Y., … Rokhsar, D. S. (2013). Insights into bilaterian evolution from three spiralian genomes. *Nature*, *493*(7433), 526–531. https://doi.org/10.1038/nature11696

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. https://doi.org/10.1093/bioinformatics/btv351

Smith, S. A., Wilson, N. G., Goetz, F. E., Feehery, C., Andrade, S. C. S., Rouse, G. W., Giribet, G., & Dunn, C. W. (2011). Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*, *480*(7377), 364–367. https://doi.org/10.1038/nature10526

Song, H., Yu, Z.-L., Sun, L.-N., Gao, Y., Zhang, T., & Wang, H.-Y. (2016). De novo transcriptome sequencing and analysis of Rapana venosa from six different developmental stages using Hi-seq 2500. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, *17*, 48–57. https://doi.org/10.1016/j.cbd.2016.01.006

Starrett, J., Derkarabetian, S., Hedin, M., Bryson, R. W., McCormack, J. E., & Faircloth, B. C. (2017). High phylogenetic utility of an ultraconserved element probe set designed for Arachnida. *Molecular Ecology Resources*, *17*(4), 812–823. https://doi.org/10.1111/1755-0998.12621

Strong, E. E. (2003). Refining molluscan characters: Morphology, character coding and a phylogeny of the Caenogastropoda. *Zoological Journal of the Linnean Society*, *137*(4), 447–554. https://doi.org/10.1046/j.1096-3642.2003.00058.x

Taylor, J. D. (1993). Dietary and anatomical specialization of mitrid gastropods (Mitridae) at Rottnest Island, Western Australia. *Proceedings of the Fifth International Marine Biological Workshop: The Marine Flora and Fauna of Rottnest Island, Western Australia*, 583–599.

Teasdale, L. C., Ko, F., Murray, K. D., O'Hara, T., & Moussalli, A. (2016). *Identification and qualification of 500 nuclear, single-copy, orthologous genes for the Eupulmonata (Gastropoda) using transcriptome sequencing and exon capture*. 17.

Tin, M. M.-Y., Economo, E. P., & Mikheyev, A. S. (2014). Sequencing Degraded DNA from Non-Destructively Sampled Museum Specimens for RAD-Tagging and Low-Coverage Shotgun Phylogenetics. *PLoS ONE*, *9*(5), e96793. https://doi.org/10.1371/journal.pone.0096793

Trevisan, B., Alcantara, D. M. C., Machado, D. J., Marques, F. P. L., & Lahr, D. J. G. (2019). Genome skimming is a low-cost and robust strategy to assemble complete mitochondrial genomes from ethanol preserved specimens in biodiversity studies. *PeerJ*, *7*, e7543. https://doi.org/10.7717/peerj.7543

Uribe, J. E., Zardoya, R., & Puillandre, N. (2018). Phylogenetic relationships of the conoidean snails (Gastropoda: Caenogastropoda) based on mitochondrial genomes. *Molecular Phylogenetics and Evolution*, *127*, 898–906. hal-02002442v1. https://doi.org/10.1016/j.ympev.2018.06.037

Vallejo, B. (2005). Inferring the mode of speciation in Indo-West Pacific Conus (Gastropoda: Conidae): Conus speciation in the IWP. *Journal of Biogeography*, *32*(8), 1429–1439. https://doi.org/10.1111/j.1365-2699.2005.01260.x

Van Valen, L. (1965). Morphological Variation and Width of Ecological Niche. *The American Naturalist*, *99*(908), 377–390. https://doi.org/10.1086/282379

Van Valen, L. (1971). ADAPTIVE ZONES AND THE ORDERS OF MAMMALS. *Evolution*, *25*(2), 420–428. https://doi.org/10.1111/j.1558-5646.1971.tb01898.x

Vaux, F., Hills, S. F. K., Marshall, B. A., Trewick, S. A., & Morgan-Richards, M. (2018). Genome statistics and phylogenetic reconstructions for Southern Hemisphere whelks (Gastropoda: Buccinulidae). *Data in Brief*, *16*, 172–181. https://doi.org/10.1016/j.dib.2017.11.021

Verhecken, A. (2007). Revision of the Cancellariidae (Mollusca, Neogastropoda, Cancellarioidea) of the eastern Atlantic (40 degrees N-40 degrees S) and the Mediterranean. *Zoosystema*, *29*(2), 281–364.

Vermeij, G. J. (2024). *Shell-based genus-level reclassification of the Family Vasidae (Mollusca: Neogastropoda)*.

Walton, K., Scarsbrook, L., Mitchell, K. J., Verry, A. J. F., Marshall, B. A., Rawlence, N. J., & Spencer, H. G. (2023). Application of palaeogenetic techniques to historic mollusc shells reveals phylogeographic structure in a New Zealand abalone. *Molecular Ecology Resources*, *23*(1), 118–130. https://doi.org/10.1111/1755-0998.13696

Wang, J.-G., Zhang, D., Jakovlić, I., & Wang, W.-M. (2017). Sequencing of the complete mitochondrial genomes of eight freshwater snail species exposes pervasive paraphyly within the Viviparidae family (Caenogastropoda). *PLOS ONE*, *12*(7), e0181699. https://doi.org/10.1371/journal.pone.0181699

Wang, Q., Liu, H., Yue, C., Xie, X., Li, D., Liang, M., & Li, Q. (2021). Characterization of the complete mitochondrial genome of *Ficus variegata* (Littorinimorpha: Ficidae) and molecular phylogeny of Caenogastropoda. *Mitochondrial DNA Part B*, *6*(3), 1126–1128. https://doi.org/10.1080/23802359.2021.1901628

Wit, E., Heuvel, E. V. D., & Romeijn, J. (2012). 'All models are wrong...': An introduction to model uncertainty. *Statistica Neerlandica*, *66*(3), 217–236. https://doi.org/10.1111/j.1467-9574.2012.00530.x

WoRMS Editorial Board. (2017). *World Register of Marine Species. Available from http://www.marinespecies.org at VLIZ. Accessed yyyy-mm-dd.* [Darwin Core Archive]. VLIZ. https://doi.org/10.14284/170

Yang, D. Y., Eng, B., Waye, J. S., Dudar, J. C., & Saunders, S. R. (1998). Improved DNA extraction from ancient bones using silica-based spin columns. *American Journal of Physical Anthropology*, *105*(4), 539–543. https://doi.org/10.1002/(SICI)1096-8644(199804)105:4<539::AID-AJPA10>3.0.CO;2-1

Yang, M., Dong, D., & Li, X. (2021). The complete mitogenome of *Phymorhynchus* sp. (Neogastropoda, Conoidea, Raphitomidae) provides insights into the deep-sea adaptive evolution of Conoidea. *Ecology and Evolution*, ece3.7582. https://doi.org/10.1002/ece3.7582

Zaharias, P., Kantor, Y. I., Fedosov, A. E., Criscione, F., Hallan, A., Kano, Y., Bardin, J., & Puillandre, N. (2020). Just the once will not hurt: DNA suggests species lumping over two oceans in deep-sea snails (Cryptogemma). *Zoological Journal of the Linnean Society*, *190*(2), 532–557. hal-02559713v1. https://doi.org/10.1093/zoolinnean/zlaa010

Zaharias, P., Kantor, Y. I., Fedosov, A. E., & Puillandre, N. (in press). *Coupling DNA barcodes and exon-capture to resolve the phylogeny of Turridae (Gastropoda, Conoidea)*.

Zaharias, P., Pante, E., Gey, D., Fedosov, A. E., & Puillandre, N. (2020). Data, time and money: Evaluating the best compromise for inferring molecular phylogenies of non-model animal taxa. *Molecular Phylogenetics and Evolution*, *142*, 106660. hal-02458233v1. https://doi.org/10.1016/j.ympev.2019.106660

Zapata, F., Wilson, N. G., Howison, M., Andrade, S. C. S., Jörger, K. M., Schrödl, M., Goetz, F. E., Giribet, G., & Dunn, C. W. (2014a). *Data from: Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda* (Version 1, p. 68327139 bytes) [dataset]. Dryad. https://doi.org/10.5061/DRYAD.5BC98

Zapata, F., Wilson, N. G., Howison, M., Andrade, S. C. S., Jörger, K. M., Schrödl, M., Goetz, F. E., Giribet, G., & Dunn, C. W. (2014b). Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1794), 20141739. https://doi.org/10.1098/rspb.2014.1739

Zou, S., Li, Q., & Kong, L. (2011). Additional gene data and increased sampling give new insights into the phylogenetic relationships of Neogastropoda, within the caenogastropod phylogenetic framework. *Molecular Phylogenetics and Evolution*, *61*(2), 425–435. https://doi.org/10.1016/j.ympev.2011.07.014

Supplementary Material

(Voir Annexe 2)

Chapitre 4 : Phylogénies

1. NEOGASTROPODA

Afin de reconstruire les arbres phylogénétiques, nous avons utilisé les 1125 exons que nous avons dessiné (Chapitre 2). Nous avons séquencé 1728 spécimens au total, mais nous avons fait une sous-sélection de 861 spécimens pour reconstruire l'arbre des Neogastropoda (voir Chapitre 2 pour les critères de sélection des spécimens). Nous avons ajouté à ces spécimens les exons du génome de référence *Conus ventricosus* qui nous ont permis de produire les sondes de capture. Nous avons donc 862 spécimens dans le jeu de données final pour les Neogastropdoa.

L'objectif est de reconstruire un arbre phylogénétique des néogastéropodes le plus complet possible, avec un spécimen par genre. Pour faciliter la présentation des résultats, je ne vais discuter que les rangs super-familial et familial, et les genres transférés d'une famille à une autre. Une discussion complète sur le rang générique nécessitera des collaborations avec les spécialistes de chaque famille, et cela n'a pas pu être effectué dans le cadre de la thèse (voir Conclusions et perspectives).

1.1. JEUX DE DONNEES ET ANALYSES PHYLOGENETIQUES

Deux jeux de données ont été utilisés pour effectuer les analyses de reconstruction phylogénétique : un jeu de données regroupant les spécimens pour lesquels au moins 700 exons avaient été capturés et un jeu de données regroupant les spécimens pour lesquels au moins 50 exons avaient été capturés. Les deux jeux de données incluent 1119 exons qui ont été capturés par au minimum 100 spécimens, sur les 1125 exons que nous avions dessinés. Le jeu de données « 700 exons » inclut 790 spécimens, tandis que le jeu de données « 50 exons » inclut 857 spécimens. Comme cité précédemment pour ces deux jeux de données, nous avons ajouté les exons du génome de référence de *Conus ventricosus*. Nous avons donc un total de 791 spécimens pour le jeu de données 700 exons et 858 spécimens pour le jeu de données 50 exons. Pour chacun des jeux de données nous avons construit trois sous-jeux de données à l'aide du script Python que j'ai décrit dans le chapitre 2. Nous avons un premier sous-jeu de données,

appelé NT123, qui correspond aux alignements en nucléotides dans lesquels toutes les positions de chaque codon sont conservées., un second sous-jeu de données, appelé NT12, dans lequel seules les deux premières bases de chaque codon sont conservées (supprimant ainsi la 3ème base, plus variable, et soumise à plus d'homoplasie), et enfin un sous-jeu de données AA qui est la traduction, dans le bon cadre de lecture, en acides aminés des alignements en nucléotides du sous-jeu de données NT123 (Figure 19).

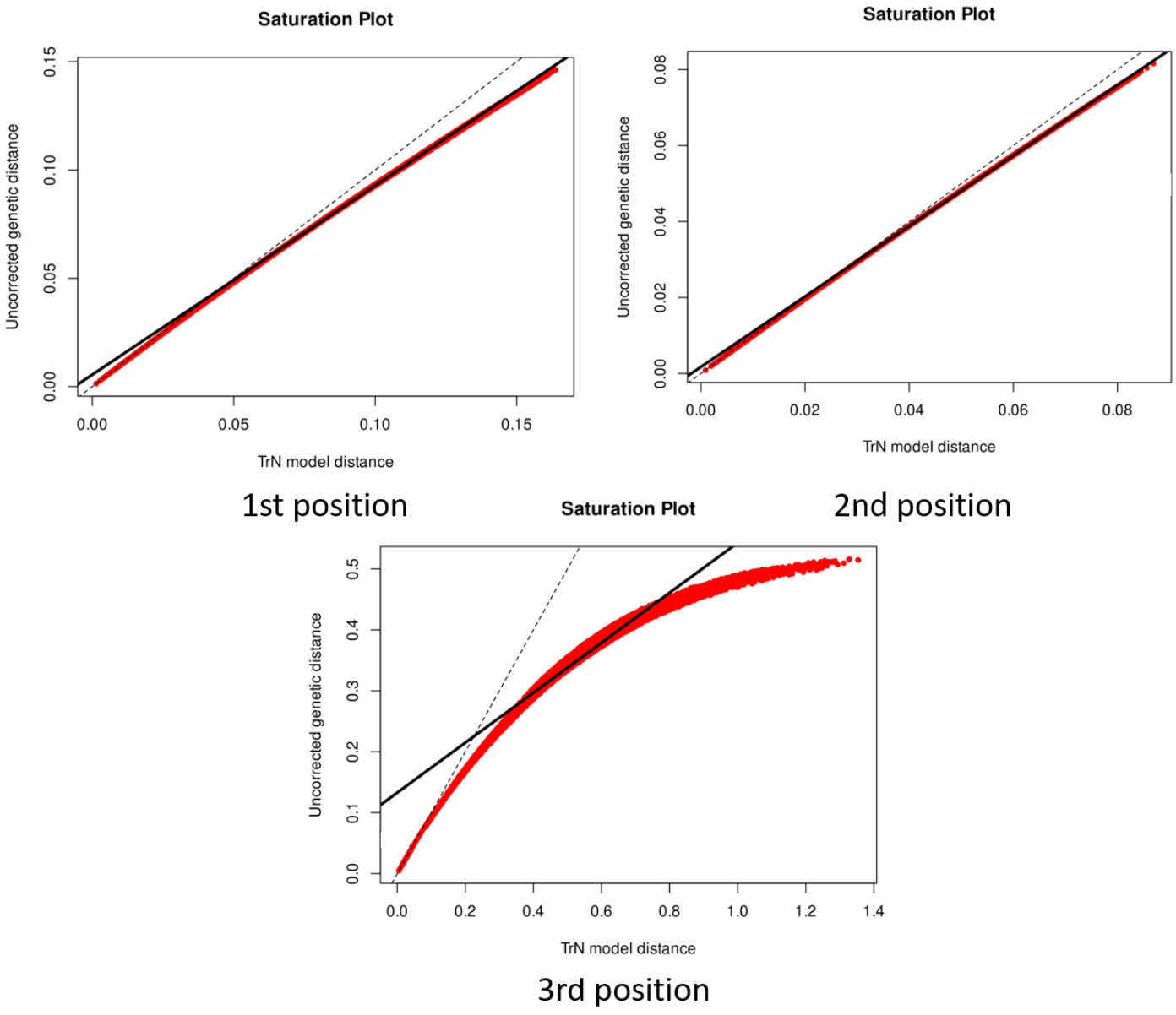


Figure 19 : Graphiques représentant la distance génétique calculée par le modèle TrN en fonction de la distance génétique non corrigée pour le jeu de données NT123-700. *Les données des trois positions de chaque codon sont représentées. On constate une saturation sur la 3ème base de chaque codon alors qu'il n'y a aucune saturation pour les deux autres positions de chaque codon.*

Toutes les analyses phylogénétiques ont été réalisées avec le logiciel IQTree, couplé avec ModelFinder pour identifier les modèles de substitution adéquats. Pour tous ces arbres, la robustesse des nœuds a été estimée avec une approche d'UltraFastBootstrap (1000 réplicats). Pour la suite de la description des arbres, je considère les nœuds avec score de bootstrap de 99 ou 100 comme supportés. Toutes les autres valeurs de bootstrap ne seront pas considérées comme supportées, et ne seront généralement pas discutées.

Enfin, pour chaque jeu de données (NT123-700, NT123-50, NT12-700, NT12-50, AA-700 et AA-50), plusieurs arbres ont été reconstruits : un premier avec un modèle non partitionné (NP), avec un modèle de substitution pour l'ensemble des exons, un second arbre utilisant un modèle partitionné (P), avec un modèle pour chaque exon, identifié de manière indépendante, ce qui résulte en 1119 partitions indépendantes, et un troisième arbre utilisant un modèle partitionné, mais pour lequel les exons suivant un modèle de substitution similaire sont regroupés et suivent un seul et même modèle (P+Merge). Ces trois analyses (NP, P et P+Merge) ont été réalisées pour les jeux de données 700 (NT123-700, NT12-700 et AA-700) ; seules les analyses NP et P ont été réalisées pour les jeux de données 50 (NT123-50, NT12-50 et AA-50). À l'heure de finaliser la rédaction de cette thèse, l'analyse AA-700 P+Merge n'est pas terminée (malgré 6 semaines d'analyse), et ne sera donc pas discutée. Cela représente un total de 14 reconstructions phylogénétiques indépendantes (Tableau 8).

Tableau 8 : Résultats des analyses phylogénétiques pour les néogastéropodes.

Les valeurs en gras des colonnes LogLikelihood, AIC, AICc et BIC sont les meilleures valeurs de chaque sous-jeu de données (NT123-700, NT12-700, AA-700, NT123-50, NT12-50 et AA-50).

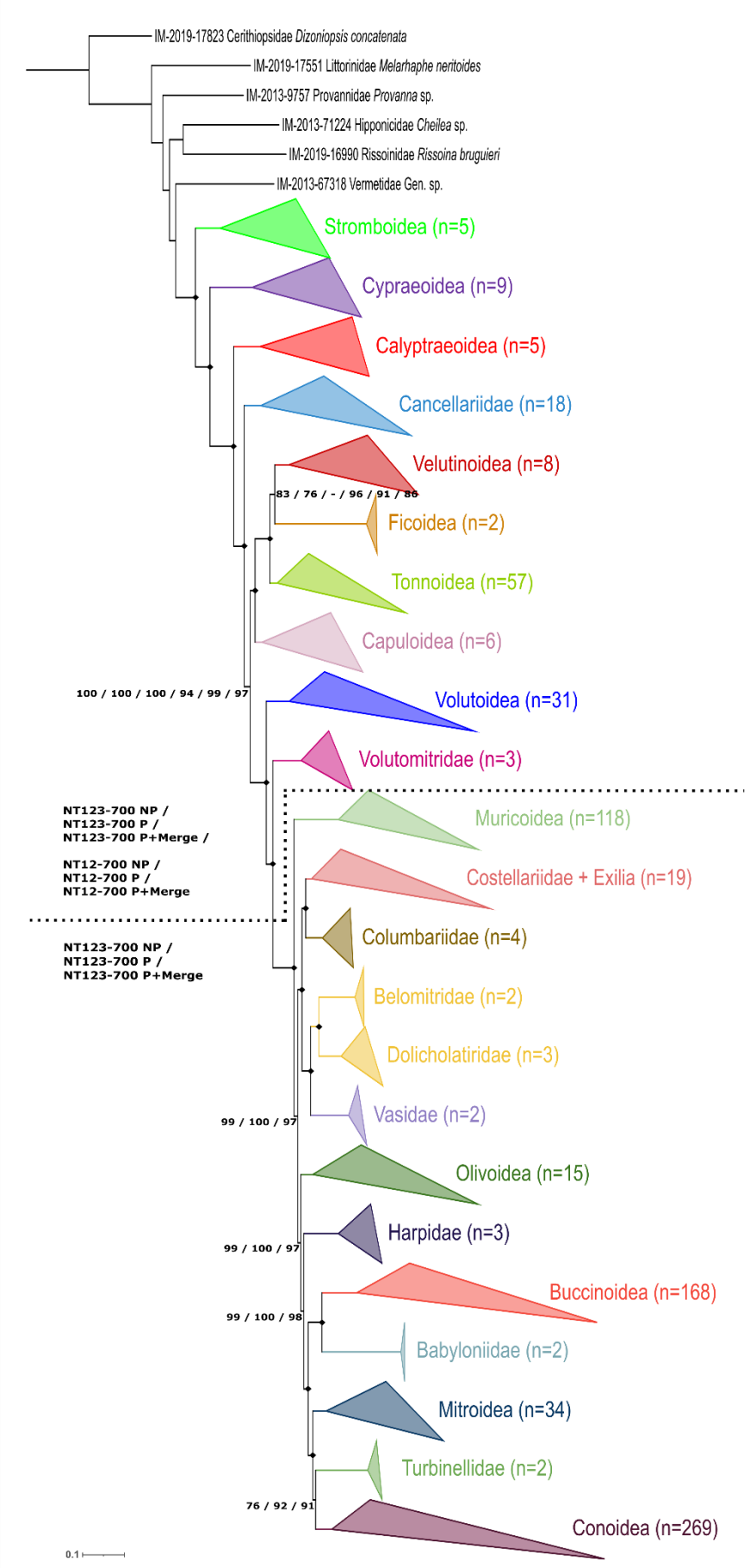
taxon	Dataset samples	NT/AA	Partitioned	software	Best Model	LogLikelihood	AIC	AICc	BIC	Free parameters
Neo	Neo > 700 exons	NT123	NP	IQ-Tree	GTR+F+I+R13	**-30027738.5541**	60058701.1081	60058723.1248	60075432.4361	1612
Neo	Neo > 700 exons	NT123	P	IQ-Tree	N.A	-29929898.9266	**59910191.8531**	**59916164.3103**	60171717.4557	25197
Neo	Neo > 700 exons	NT123	P+Merge	IQ-Tree	N.A	-29953939.9001	59924337.8002	59924927.7826	**60009748.5310**	8229
Neo	Neo > 700 exons	NT12	NP	IQ-Tree	GTR+F+I+R12	**-9407904.7514**	18819029.5028	18819062.4656	18835091.7524	1610
Neo	Neo > 700 exons	NT12	P	IQ-Tree	N.A	-9336040.5382	**18710795.0764**	**18716162.4935**	18903911.2042	19357
Neo	Neo > 700 exons	NT12	P+Merge	IQ-Tree	N.A	-9354755.0889	18720658.1777	18721063.3072	**18776267.4817**	5574
Neo	Neo > 700 exons	AA	NP	IQ-Tree	JTT+F+I+R15	**-8986489.3757**	17976232.7513	17976301.0834	**17991329.9293**	1627
Neo	Neo > 700 exons	AA	P	IQ-Tree	N.A	-8969009.8237	**17962167.6473**	**17966514.5681**	18074204.1117	12074
Neo	Neo > 700 exons	AA	P+Merge	IQ-Tree	N.A	N.A	N.A	N.A	N.A	N.A
Neo	Neo > 50 exons	NT123	NP	IQ-Tree	GTR+F+I+R19	**-31716134.6150**	63435785.2300	63435811.4303	**63454031.9264**	1758
Neo	Neo > 50 exons	NT123	P	IQ-Tree	N.A	-31616390.9655	**63284027.9310**	**63290216.4408**	63549975.0880	25623
Neo	Neo > 50 exons	NT12	NP	IQ-Tree	GTR+F+I+R12	**-9906637.1722**	19816762.3444	19816801.0536	**19834161.4521**	1744
Neo	Neo > 50 exons	NT12	P	IQ-Tree	N.A	-9832717.7891	**19705103.5783**	**19710758.0956**	19902978.5217	19834
Neo	Neo > 50 exons	AA	NP	IQ-Tree	JTT+F+I+R15	**-9430620.5080**	18864763.0161	18864843.2022	**18881103.6003**	1761
Neo	Neo > 50 exons	AA	P	IQ-Tree	N.A	-9414205.5986	**18853071.1971**	**18857621.7630**	18967483.1241	12330

Le nombre de spécimens ainsi que la quantité d'exons capturés pour chaque spécimen ne nous ont pas permis d'envisager de reconstruire des arbres à partir d'analyses bayésiennes. En effet, la complexité du jeu de données rendrait les temps de calcul bien trop longs. J'ai néanmoins reconstruit des arbres de gènes (pour chaque exon indépendamment), pour ensuite reconstruire un super-arbre avec le logiciel ASTRAL (Mirarab et al., 2014). Cependant, les arbres obtenus avaient des supports de nœuds significativement moins bons par rapport aux arbres de maximum de vraisemblance. De plus, les topologies obtenues étaient divergentes entre elles avec parfois des groupes retrouvés non monophylétiques, alors qu'ils sont systématiquement retrouvés monophylétiques dans les arbres obtenus avec IQ-tree, et dans les arbres publiés par le passé. Ces analyses ne seront donc pas discutées par la suite.

Dans la partie suivante, chaque topologie obtenue est discutée, en se focalisant sur les rangs famille et super-famille. Les transferts de genres d'une famille à une autre seront discutés dans la partie suivante, avec les autres changements proposés dans la classification des néogastéropodes. Quatre topologies sont présentées, l'ensemble des arbres étant disponible en Annexe 3.

1.2. ARBRE PHYLOGENETIQUE NT123-700

Figure 20 : Arbre phylogénétique obtenu avec le jeu de données NT123-700.
Les familles et super-familles sont représentées par des triangles de taille et de couleurs différentes. La taille des triangles est fonction du nombre de spécimens par groupe mais également de la longueur des branches. Les nœuds supportés avec des valeurs de bootstrap de 99 ou 100 et retrouvés dans tous les arbres sont représentés par des losanges noirs. Le bas de l'arbre en-dessous des traits pointillés représente les 3 arbres des jeux de données NT123-700 NP, P et P+Merge. Les topologies sont identiques et seules les valeurs de nœuds qui sont différentes sont notées. Le haut de l'arbre au-dessus des pointillés représente les 6 arbres des jeux de données NT123-700 NP, P et P+Merge et NT12-700 NP, P et P+Merge. Les topologies de ces arbres sont les mêmes et seuls les nœuds avec des supports variables sont notés. La valeur de nœud notée « - » signifie que le nœud n'est pas retrouvé et n'est pas supporté, il n'est donc pas représenté.


IM-2019-17823 Cerithiopsidae *Dizoniopsis concatenata*
IM-2019-17551 Littorinidae *Melarhaphe neritoides*
IM-2013-9757 Provannidae *Provanna* sp.
IM-2013-71224 Hipponicidae *Cheilea* sp.
IM-2019-16990 Rissoinidae *Rissoina bruguieri*
IM-2013-67318 Vermetidae Gen. sp.
Stromboidea (n=5)
Cypraeoidea (n=9)
Calyptraeoidea (n=5)
Cancellariidae (n=18)
Velutinoidea (n=8)
83 / 76 / - / 96 / 91 / 88
Ficoidea (n=2)
Tonnoidea (n=57)
Capuloidea (n=6)
100 / 100 / 100 / 94 / 99 / 97
Volutoidea (n=31)
Volutomitridae (n=3)
NT123-700 NP /
NT123-700 P /
NT123-700 P+Merge /
NT12-700 NP /
NT12-700 P /
NT12-700 P+Merge
Muricoidea (n=118)
Costellariidae + Exilia (n=19)
NT123-700 NP /
NT123-700 P /
NT123-700 P+Merge
Columbariidae (n=4)
Belomitridae (n=2)
Dolicholatiridae (n=3)
Vasidae (n=2)
99 / 100 / 97
Olivoidea (n=15)
Harpidae (n=3)
99 / 100 / 97
Buccinoidea (n=168)
99 / 100 / 98
Babyloniidae (n=2)
Mitroidea (n=34)
Turbinellidae (n=2)
76 / 92 / 91
Conoidea (n=269)
0.1

Les trois arbres partitionnés (P et P+Merge) et non partitionnés (NP) pour le jeu de données NT123-700 ont la même topologie, avec des différences de supports sur certains nœuds (Figure 20). Dans ces trois arbres phylogénétiques, tous les nœuds correspondants à des familles ou des super-familles sont monophylétiques et soutenus avec un score de bootstrap de 100, avec quelques exceptions décrites ci-après. En décrivant les arbres des nœuds les plus récents vers les nœuds les plus anciens, nous avons un premier groupe qui comprend tous les spécimens de la super-famille des Conoidea (n=269). Au sein des Conoidea, toutes les familles (telles que définies dans Abdelkrim et al 2018) sont monophylétiques (avec quelques transferts de genres qui seront discutés plus loin), à l'exception des Borsoniidae, séparés en 5 clades indépendants, et des Pseudomelatomidae, dont 3 genres (*Antiplanes*, *Abyssocomitas* et *Leucosyrinx*) sont exclus et forment un clade indépendant. Les Conoidea sont groupe-frère de la famille des Turbinellidae (n=2), la relation entre ces deux groupes n'étant pas supportée (NP : 76, P : 92 et P+Merge : 91). Les Conoidea+Turbinellidae sont groupe-frère d'un clade regroupant les trois familles monophylétiques de Mitroidea (n=34), avec un support de bootstrap de 100. Ce clade sera désigné par la suite MTC. En groupe-frère du clade MTC, la famille des Babyloniidae (n=2) se regroupe avec les Buccinoidea (n=168), avec un score de bootstrap de 100 (Figure 20). Au sein des Buccinoidea, les familles sont retrouvées monophylétiques (avec, comme pour les Conoidea, quelques transferts de genres qui seront discutés plus loin), à l'exception des Nassariidae, qui incluent les Chauvetiidae et les Columbellidae, et de certaines lignées indépendantes non assignées à des familles, comme le spécimen IM-2013-68253 identifié comme Dolicholatiridae (mais qui ne se placent pas avec les autres membres de cette famille), et les spécimens de *Liomesus* et *Macron* (Figure 20).

Les clades BB (Babyloniidae et Buccinoidea) et MTC sont groupe-frères avec une valeur de bootstrap de 100 également (Figure 20). Les Harpidae (n=3) sont groupe-frère de MTC/BB avec des supports de bootstrap élevés (NP : 99, P : 100 et P+Merge :98). Viennent ensuite les Olivoidea (n=15), dont les quatre familles constitutives sont monophylétiques, avec des supports de bootstrap également élevés (NP : 99, P : 100 et P+Merge : 97).

Les spécimens de la famille des Belomitridae (n=2) sont placés en groupe-frère des Dolicholatiridae (n=3) avec un support de bootstrap de 100. Ces deux familles se groupent avec un clade regroupant les Vasidae (n=2), récemment ré-élevés au rang familial (Vermeij, 2024), avec un support de bootstrap de 100 également, formant le clade BDV (Figure 20). Les spécimens de la famille des Columbariidae (n=4) sont groupe-frère d'un clade incluant les Costellariidae et trois spécimens des genres *Exilia* (Ptychatractidae), *Egestas* (Ptychatractidae)

et *Enigmavasum* (unassigned Neogastropoda) (n=19), avec un score de boostrap de 100 (Figure 20). Ce clade CC est groupe-frère de BDV avec un score de bootstrap de 100. L'ensemble CC/BDV est groupe-frère de tous les autres groupes que j'ai décrit précédemment (clade MTC/BB, + Harpidae et Olivoidea), avec des scores de support de nœud élevés (NP : 99, P : 100 et P+Merge : 97). Viennent ensuite se brancher les Muricoidea (n=118), puis les Volutomitridae (n=3) et enfin les Volutoidea (n=31), avec à chaque fois des supports de bootstrap de 100 (Figure 20), regroupant ainsi tous les néogastéropodes (sauf les Cancellariidae – voir ci-dessous). Au sein des Volutoidea, les Volutidae incluent le seul représentant des Marginellonidae, et les familles des Marginellidae et Cystiscidae sont mélangées, reflétant probablement des erreurs d'identification, compliquée pour ces taxa.

Les Tonnoidea, parfois placés au sein des néogastéropodes dans des phylogénies publiées (Cunha et al., 2009; Osca et al., 2015), au sein desquels les neuf familles représentées sont monophylétiques, et groupe-frère (NP : 83 et P : 76) d'un clade Ficidae (n=2) + Velutinoidea(n=8), avec dans ce dernier les Eratoidae, les Triviidae et les Velutinidae (ces dernier étant séparés en trois lignées indépendantes). Ce nœud n'est pas retrouvé dans l'analyse P+Merge (Figure 20), les Ficidae se groupant d'abord avec les Tonnoidea avant les Velutinoidea. Cependant, ce nœud n'étant pas supporté, il ne sera pas discuté. Les Tonnoidea, Ficidae et Velutinoidea sont groupe-frère des Capuloidea (n=6) avec un support de bootstrap de 100, l'ensemble étant groupe-frère des néogastéropodes (sauf Cancellariidae) avec un support de 100 (Figure 20). Les derniers représentants des néogastéropodes, la famille des Cancellariidae (n=18), se branchent ensuite avec un support de 100. Viennent après les Calyptraeoidea (n=5) puis les Cypraeoidea (n=9), puis les Stromboidea (n=5), à chaque fois avec des bootstraps de 100, et enfin les groupes externes les plus éloignés des Neogastropoda.

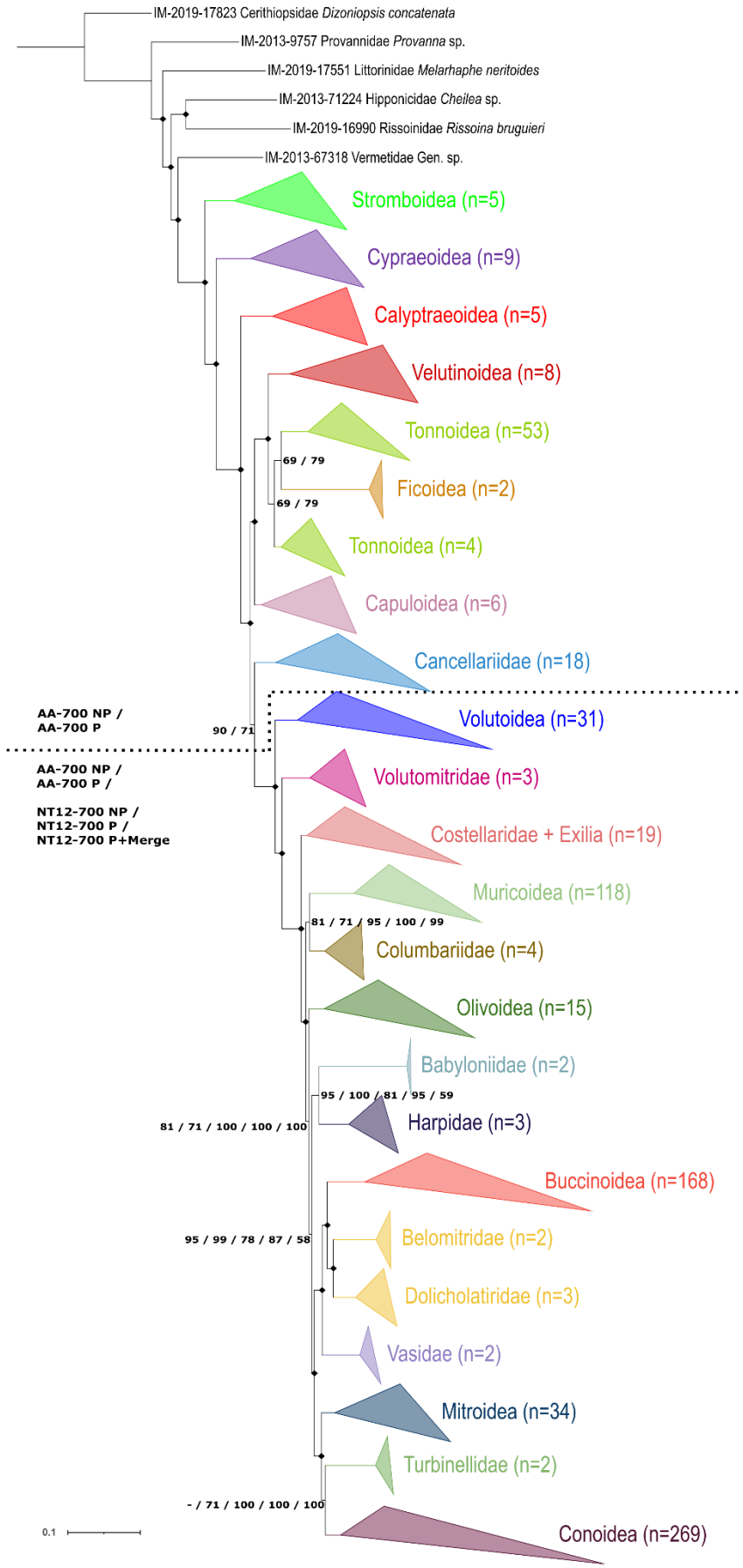
1.3. Comparaison avec les autres arbres

Dans cette partie, je vais principalement décrire les différences entre la topologie décrite ci-dessus, obtenue avec le jeu de données NT123-700, et les autres topologies obtenues avec les autres jeux de données (NT12-700, AA-700, NT123-50, NT12-50 et AA-50). Comme précisé précédemment, je discuterai surtout les nœuds supportés (les nœuds non supportés sont cependant présents sur les figures).

AA-700

Comme dans la topologie NT123-700, les Belomitridae (n=2) sont groupe-frère des Dolicholatiridae (n=3), l'ensemble se plaçant en groupe-frère des Buccinoidea (n=168) avec un nœud supporté à 100, les Vasidae (n=2) se retrouvant en groupe-frère du clade Buccinoidea/Belomitridae/Dolicholatiridae avec un support de bootstrap de 100 (Figure 21). Au sein des Buccinoidea, les Collumbelidae ne sont plus inclus dans les Nassariidae. Cet ensemble est lui-même groupe-frère du groupe MTC avec un support de bootstrap de 100 (Figure 21). Les nœuds suivants, concernant les taxons Harpidae, Babyloniidae, Muricoidea, Olivoidea et Columbariidae, sont différents de la topologie NT123-700, mais ne sont pas supportés. Viennent ensuite les spécimens de la famille des Costellariidae (plus les genres *Exilia*, *Egestas* et *Enigmavasum*) (n=19) avec un score de bootstrap de 100, suivis, comme pour la topologie NT123-700, des Volutomitridae (n=3) puis des Volutoidea (n=31). Les Cancellariidae (n=18) se placent ici en groupe-frère des autres néogastéropodes, avec un nœud non supporté (NP : 90 et P : 71). On retrouve ensuite le clade Tonnoidea/Ficoidea/Velutinoidea/Capuloidea avec un score de bootstrap de 100 (Figure 21), à la différence que les Tonnoidea ne sont pas retrouvés monophylétiques : les Ficoidae viennent se placer entre les Thallasocyonidae/Personidae d'un côté, et les autres familles de Tonnoidea de l'autre, avec des nœuds non supportés (NP : 69 et P : 79). Le reste de l'arbre est identique à l'arbre NT123-700 (Figure 21).

Figure 21 : Arbre phylogénétique obtenu avec le jeu de données AA-700.
Les familles et super-familles sont représentées par des triangles de taille et de couleurs différentes. La taille des triangles est fonction du nombre de spécimens par groupe mais également de la longueur des branches. Les nœuds supportés avec des valeurs de bootstrap de 99 ou 100 et retrouvés dans tous les arbres sont représentés par des losanges noirs. Le haut de l'arbre au-dessus des traits pointillés représente les 2 arbres des jeux de données AA-700 NP, P. Les topologies sont identiques et seules les valeurs de nœuds qui sont différentes sont notées. Le bas de l'arbre en-dessous des pointillés représente les 5 arbres des jeux de données AA-700 NP, P et NT12-700 NP, P et P+Merge. Les topologies de ces arbres sont les mêmes et seuls les nœuds avec des supports variables sont notés. La valeur de nœud notée « - » signifie que le nœud n'est pas retrouvé et n'est pas supporté, il n'est donc pas représenté.


IM-2019-17823 Cerithiopsidae *Dizoniopsis concatenata*
IM-2013-9757 Provannidae *Provanna* sp.
IM-2019-17551 Littorinidae *Melarhaphe neritoides*
IM-2013-71224 Hipponicidae *Cheilea* sp.
IM-2019-16990 Rissoinidae *Rissoina bruguieri*
IM-2013-67318 Vermetidae Gen. sp.
Stromboidea (n=5)
Cypraeoidea (n=9)
Calyptraeoidea (n=5)
Velutinoidea (n=8)
Tonnoidea (n=53)
69 / 79
Ficoidea (n=2)
69 / 79
Tonnoidea (n=4)
Capuloidea (n=6)
Cancellariidae (n=18)
AA-700 NP /
AA-700 P
90 / 71
Volutoidea (n=31)
AA-700 NP /
AA-700 P /
NT12-700 NP /
NT12-700 P /
NT12-700 P+Merge
Volutomitridae (n=3)
Costellaridae + Exilia (n=19)
Muricoidea (n=118)
81 / 71 / 95 / 100 / 99
Columbariidae (n=4)
Olivoidea (n=15)
Babyloniidae (n=2)
95 / 100 / 81 / 95 / 59
81 / 71 / 100 / 100 / 100
Harpidae (n=3)
Buccinoidea (n=168)
95 / 99 / 78 / 87 / 58
Belomitridae (n=2)
Dolicholatiridae (n=3)
Vasidae (n=2)
Mitroidea (n=34)
Turbinellidae (n=2)
- / 71 / 100 / 100 / 100
Conoidea (n=269)
0.1

NT12-700

Les trois arbres obtenus avec le jeu de données NT12-700 ont la même topologie avec des différences de support de bootstrap pour quelques nœuds. Les arbres de ce jeu de données suivent la topologie des arbres AA-700 pour les nœuds les plus récents avec quelques différences dans le support de certains nœuds (Figure 21). Pour les nœuds plus anciens, en particulier pour la position des Cancellariidae hors des néogastéropodes, les arbres NT12-700 suivent les topologies des arbres NT123-700 (Figure 20).

Le nœud regroupant les Conoidea et les Turbinellidae est toujours soutenu dans les trois analyses du jeu de données NT12-700, contrairement aux analyses NT123-700 et AA-700 (Figure 20 et 21). Le nœud regroupant les Babyloniidae, Harpidae, Buccinoidea, Belomitridae, Dolicholatiridae, Vasidae et MTC est soutenu dans les trois arbres, contrairement aux analyses AA-700 (Figure 21). Par ailleurs, le nœud qui lie les Muricoidea et les Columbariidae est également soutenu pour les analyses P et P+Merge contrairement aux analyses en AA-700 (Figure 21).

NT123-50

Pour le jeu de données NT123-50, comme cité précédemment, nous avons reconstruit deux arbres phylogénétiques différents, un avec un modèle partitionné (P) et un avec un modèle non-partitionné (NP). Les arbres reconstruits avec les jeux de données 50 exons intègrent une famille qui était absente des arbres 700 exons, les Strepsiduridae, représentée par un seul spécimen (Figure 22). Les deux arbres suivent globalement la même topologie avec quelques différences dans les relations entre quelques familles, mais ces nœuds ne sont pas supportés.

Le spécimen de Strepsiduridae vient s'intégrer dans un clade regroupant les Columbariidae, Costellariidae, Belomitridae, Dolicholatiridae et Vasidae (Figure 22). Les Volutomitridae sont monophylétiques comme pour le jeu de données NT123-700 mais ils intègrent en plus un spécimen du genre *Exilioidea*, de la famille des Ptychatractidae (Figure 22).

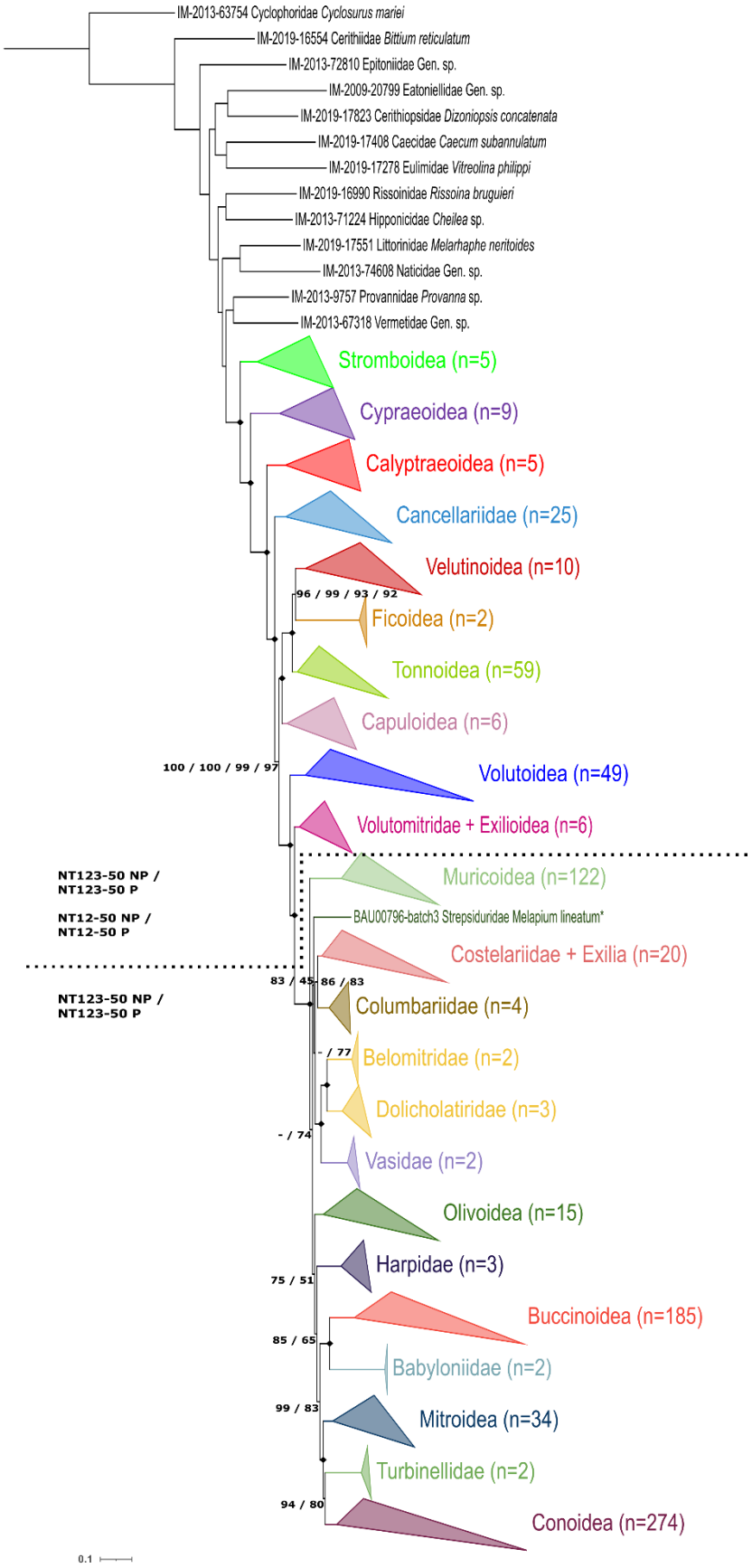

IM-2013-63754 Cyclophoridae *Cyclosurus mariei*
IM-2019-16554 Cerithiidae *Bittium reticulatum*
IM-2013-72810 Epitoniidae Gen. sp.
IM-2009-20799 Eatoniellidae Gen. sp.
IM-2019-17823 Cerithiopsidae *Dizoniopsis concatenata*
IM-2019-17408 Caecidae *Caecum subannulatum*
IM-2019-17278 Eulimidae *Vitreolina philippi*
IM-2019-16990 Rissoinidae *Rissoina bruguieri*
IM-2013-71224 Hipponicidae *Cheilea* sp.
IM-2019-17551 Littorinidae *Melarhaphe neritoides*
IM-2013-74608 Naticidae Gen. sp.
IM-2013-9757 Provannidae *Provanna* sp.
IM-2013-67318 Vermetidae Gen. sp.
Stromboidea (n=5)
Cypraeoidea (n=9)
Calyptraeoidea (n=5)
Cancellariidae (n=25)
Velutinoidea (n=10)
96 / 99 / 93 / 92
Ficoidea (n=2)
Tonnoidea (n=59)
Capuloidea (n=6)
100 / 100 / 99 / 97
Volutoidea (n=49)
Volutomitridae + Exilioidea (n=6)
NT123-50 NP /
NT123-50 P
NT12-50 NP /
NT12-50 P
Muricoidea (n=122)
BAU00796-batch3 Strepsiduridae Melapium lineatum*
Costelariidae + Exilia (n=20)
83 / 45
86 / 83
NT123-50 NP /
NT123-50 P
Columbariidae (n=4)
- / 77
Belomitridae (n=2)
Dolicholatiridae (n=3)
- / 74
Vasidae (n=2)
Olivoidea (n=15)
Harpidae (n=3)
75 / 51
Buccinoidea (n=185)
85 / 65
Babyloniidae (n=2)
99 / 83
Mitroidea (n=34)
Turbinellidae (n=2)
94 / 80
Conoidea (n=274)
0.1

Figure 22 : Arbre phylogénétique obtenu avec le jeu de données NT123-50.

Les familles et super-familles sont représentées par des triangles de taille et de couleurs différentes. La taille des triangles est fonction du nombre de spécimens par groupe mais également de la longueur des branches. Les nœuds supportés avec des valeurs de bootstrap de 99 ou 100 et retrouvés dans tous les arbres sont représentés par des losanges noirs. Le bas de l'arbre en-dessous des traits pointillés représente les 2 arbres des jeux de données NT123-50 NP et P. Les topologies sont identiques et seules les valeurs de nœuds qui sont différentes sont notées. Le haut de l'arbre au-dessus des pointillés représente les 4 arbres des jeux de données NT123-50 NP, P et NT12-50 NP, P. Les topologies de ces arbres sont les mêmes et seuls les nœuds avec des supports variables sont notés. Les valeurs des nœuds notées « - » signifient que les nœuds ne sont pas retrouvés et ne sont pas supportés, ils ne sont donc pas représentés.

NT12-50

Pour le jeu de données NT12-50, les deux arbres reconstruits ont la même topologie, sauf pour le positionnement des Harpidae entre l'analyse partitionnée (P) et non partitionnée (NP), mais ce nœud n'est jamais soutenu. Le spécimen de Strepsiduridae se place en groupe-frère des Babyloniidae avec un nœud non soutenu (NP : 83 et P : 86). Le spécimen du genre *Exilioidea* se place avec les Volutomitridae avec un nœud soutenu à 100 (Figure 23). Le nœud regroupant les Tonnoidea n'est également pas soutenu dans ces deux analyses (NP : 94 et P : 93) (Figure 22).

AA-50

Les deux arbres, partitionné (P) et non partitionné (NP) du jeu de données AA-50 ont la même topologie (Figure 23). Le spécimen de Strepsiduridae se place en groupe-frère des Babyloniidae avec un nœud non soutenu (NP : 74 et P : 89), comme pour les arbres NT12-50 (Figure 23).

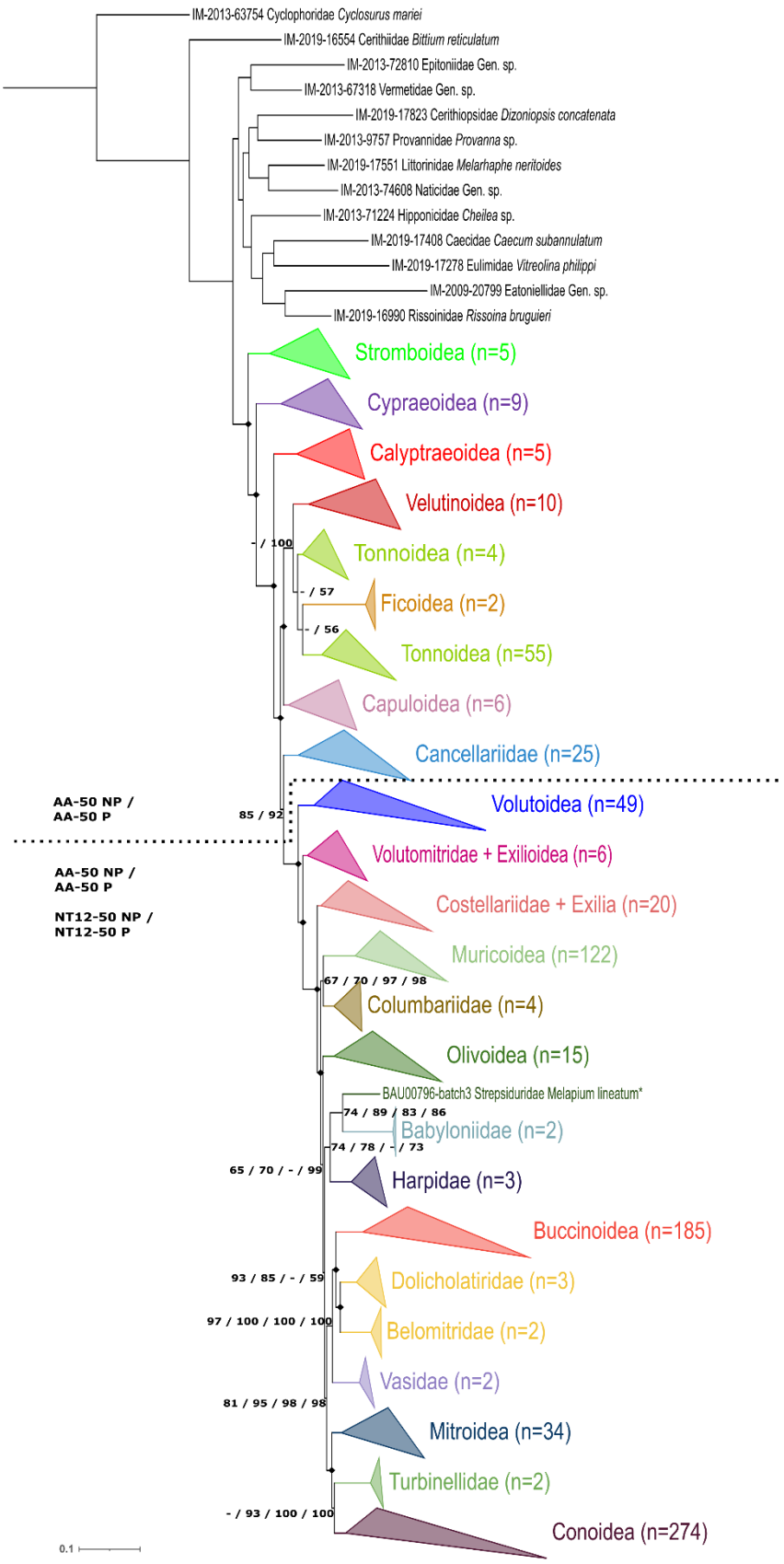

IM-2013-63754 Cyclophoridae *Cyclosurus mariei*
IM-2019-16554 Cerithiidae *Bittium reticulatum*
IM-2013-72810 Epitoniidae Gen. sp.
IM-2013-67318 Vermetidae Gen. sp.
IM-2019-17823 Cerithiopsidae *Dizoniopsis concatenata*
IM-2013-9757 Provannidae *Provanna* sp.
IM-2019-17551 Littorinidae *Melarhaphe neritoides*
IM-2013-74608 Naticidae Gen. sp.
IM-2013-71224 Hipponicidae *Cheilea* sp.
IM-2019-17408 Caecidae *Caecum subannulatum*
IM-2019-17278 Eulimidae *Vitreolina philippi*
IM-2009-20799 Eatoniellidae Gen. sp.
IM-2019-16990 Rissoinidae *Rissoina bruguieri*
Stromboidea (n=5)
Cypraeoidea (n=9)
Calyptraeoidea (n=5)
Velutinoidea (n=10)
- / 100
Tonnoidea (n=4)
- / 57
Ficoidea (n=2)
- / 56
Tonnoidea (n=55)
Capuloidea (n=6)
Cancellariidae (n=25)
AA-50 NP /
AA-50 P
85 / 92
Volutoidea (n=49)
Volutomitridae + Exilioidea (n=6)
AA-50 NP /
AA-50 P
NT12-50 NP /
NT12-50 P
Costellariidae + Exilia (n=20)
Muricoidea (n=122)
67 / 70 / 97 / 98
Columbariidae (n=4)
Olivoidea (n=15)
BAU00796-batch3 Strepsiduridae Melapium lineatum*
74 / 89 / 83 / 86
Babyloniidae (n=2)
74 / 78 / - / 73
65 / 70 / - / 99
Harpidae (n=3)
Buccinoidea (n=185)
93 / 85 / - / 59
Dolicholatiridae (n=3)
97 / 100 / 100 / 100
Belomitridae (n=2)
Vasidae (n=2)
81 / 95 / 98 / 98
Mitroidea (n=34)
Turbinellidae (n=2)
- / 93 / 100 / 100
Conoidea (n=274)
0.1

Figure 23 : Arbre phylogénétique obtenu avec le jeu de données AA-50.

Les familles et super-familles sont représentées par des triangles de taille et de couleurs différentes. La taille des triangles est fonction du nombre de spécimens par groupe mais également de la longueur des branches. Les nœuds supportés avec des valeurs de bootstrap de 99 ou 100 et retrouvés dans tous les arbres sont représentés par des losanges noirs. Le haut de l'arbre au-dessus des traits pointillés représente les 2 arbres des jeux de données AA-50 NP, P. Les topologies sont identiques et seules les valeurs de nœuds qui sont différentes sont notées. Le bas de l'arbre en-dessous des pointillés représente les 4 arbres des jeux de données AA-50 NP, P et NT12-50 NP, P. Les topologies de ces arbres sont les mêmes et seuls les nœuds avec des supports variables sont notés. Les valeurs des nœuds notées « - » signifient que les nœuds ne sont pas retrouvés et ne sont pas supportés, ils ne sont donc pas représentés.

1.4. CHANGEMENTS POTENTIELS DANS LA CLASSIFICATION

Les arbres sont largement congruents entre toutes les analyses réalisées, à l'échelle des familles et des super-familles, et également congruents avec les résultats présentés dans Fedosov et al ((A. E. Fedosov et al., in press) ; Annexe 1) avec un échantillonnage bien plus limité. Cependant, certains nœuds qui posaient problème dans Fedosov et al. (A. E. Fedosov et al., in press) sont maintenant plus stables et mieux résolus, comme en particulier l'association Belomitidae/Dolicholatiridae.

La plupart des taxons sont retrouvés monophylétiques, à quelques exceptions près, souvent liées à des nœuds non soutenus (comme la non-monophylie des Tonnoidea dans un des arbres du jeu de données AA-700).

Cependant, plusieurs relations entre les grandes lignées restent instables, en particulier pour les plus récentes. C'est le cas pour les groupes MTC, Buccinoidea, Belomitridae/Dolicholatiridae, Vasidae, Olivoidea, Babyloniidae, Strepsiduridae, Harpidae, Columbariidae et Muricidae. La position des Cancellariidae est également incertaine, parfois en groupe-frère des autres néogastéropodes, parfois positionnés en groupe-frère des clades Tonnoidea/Ficoidea/Velutinoidea/Capuloidea et des autres néogastéropodes (Figures 20-21-22 et 23).

Si on s'intéresse au rang générique, plusieurs changements de familles, listés ci-dessous, sont retrouvés très bien supportés dans l'ensemble des arbres.

Par rapport à la classification actuelle (WoRMS), les changements suggérés dans la classification des néogastéropodes tiennent compte des topologies soutenues obtenues, mais doivent être formalisés avec l'aide des taxonomistes spécialistes de chaque groupe, et feront l'objet de publications ultérieures.

Les taxons qui ne sont pas discutés ci-dessous ne nécessitent pas de modifications de leur classification.

Turbinelloidea

Dans l'ensemble des arbres phylogénétiques pour les jeux de données comprenant les spécimens pour lesquels nous avons capturé plus de 700 exons et 50 exons par spécimen, la super-famille des Turbinelloidea est retrouvée polyphylétique : cinq clades distincts sont retrouvés systématiquement (Figures 20-21-22 et 23).

1. Un premier clade qui regroupe les spécimens de la famille des Turbinellidae, incluant le genre *Turbinella* (Annexe 3). Cette famille est systématiquement monophylétique et soutenue. Les spécimens des Turbinellidae sont, dans tous les arbres, placés en groupe-frère des Conoidea. Les nœuds ne sont cependant pas toujours soutenus selon l'arbre considéré.

2. Une seconde lignée inclut les genres de Vasidae. Ce clade est soutenu dans tous les arbres reconstruits (Figures 20-21-22 et 23). La position de la famille change selon les arbres : dans les arbres NT123, les spécimens de Vasidae se placent en groupe-frère d'un clade Belomitridae + Dolicholatiridae (Figures 20 et 22), tandis que pour les arbres NT12 et AA les Vasidae sont en groupe-frère d'un clade Belomitridae + Dolicholatiridae + Buccinoidea (Figures 21 et 23).

3. La famille des Columbariidae, est toujours retrouvée monophylétique et soutenue. Le positionnement de cette famille varie entre les analyses : dans l'arbre NT123, elle est groupe-frère de la famille des Costellariidae (Figures 20 et 22), alors que dans les arbres NT12 et AA, les spécimens de Columbariidae sont placés en groupe-frère des Muricoidea (Figures 21 et 23).

4. La famille des Costellariidae est monophylétique et soutenue, à condition d'y inclure le genre *Egestas*, actuellement placé dans les Ptychatractidae (Figures 20-21-22 et 23). Les genres *Exilia*

(Ptychatractidae) et *Enigmavasum* (unassigned Neogastropoda) pourraient également être incluent dans les Costellariidae (Annexe 3), ou constituer deux nouvelles familles (mais voir ci-dessous pour les Ptychatractidae).

5. Enfin, la famille des Volutomitridae est monophylétique et soutenue dans tous les arbres (Figures 20-21-22 et 23). Leur positionnement dans l'arbre est également le même dans tous les arbres considérés. Dans les jeux de données 50, le genre *Exilioidea* (Ptychtatractidae) est groupe-frère des Volutomitridae (Figures 22 et 23). En l'absence du genre-type des Ptychatractidae, *Ptychatractus*, il est difficile de savoir si le nom Ptychatractidae doit être appliqué à la lignée *Exilioidea*, à la lignée *Exilia*, ou à aucune des deux. Une ou deux nouvelles familles devront donc être décrites pour accommoder les topologies obtenues (Annexe 3).

Ces cinq lignées pourraient constituer autant de superfamilles au sein des néogastéropodes, à l'exception éventuellement des Vasidae, qui, comme les Belomitridae et Dolicholatiridae, pourraient être placés dans les Buccinoidea.

Buccinoidea

À condition d'y inclure les Vasidae, les Buccinoidea sont en général retrouvés monophylétiques (Figures 20-21-22 et 23). Plusieurs genres devraient changer de famille : *Taphon* (de Melongenidae vers Fasciolariidae), *Africofusus* (de Fasciolariidae vers Tudiclidae), *Triumphis* (de Pseudolividae vers Pisaniidae), *Monostiolum* (de Pisaniidae vers Prodotiidae), *Pseudamycla* (de Columbellidae vers Nassariidae) et *Troschelia* (de Buccindae vers Colidae). De plus, trois genres nécessiteraient la création de nouvelles familles : *Macron* (actuellement dans les Pseudolividae), *Cyllene* (actuellement dans les Nassariidae) et *Liomesus* (actuellement dans les Buccinidae). Enfin, le spécimen IM-2013-68253, identifié comme un *Dolicholatirus*, ne se groupe pas avec les autres Dolicholatiridae, et constitue une lignée indépendante au sein des Buccinoidea, qui nécessiterait également une nouvelle famille (Annexe 3).

L'inclusion des Columbellidae dans les Nassariidae est interprétée comme un artefact (voir discussion dans (A. E. Fedosov et al., in press), Annexe 1). Les Nassariidae sont ainsi retrouvés monophylétiques, à condition de proposer une nouvelle famille pour le genre *Cyllene* (Annexe 3).

Olivoidea

Au sein des Olivoidea, un seul changement serait à proposer : le transfert du genre *Jaspidella* des Bellolividae vers les Olividae (Annexe 3).

Volutoidea

La super-famille des Volutoidea est non-monophylétique dans tous les arbres (Figures 20-21-22 et 23), avec deux clades soutenus : un clade correspondant à la famille des Cancellariidae et un clade incluant les autres Volutoidea (Volutidae, Marginellidae, Cystiscidae, Marginellonidae) (Annexe 3). Les Volutoidea, sauf Cancellaridae, sont soutenus et monophylétiques dans tous les arbres, et leur position est conservée entre toutes les topologies d'arbres (Figures 20-21-22 et 23). En revanche, la famille des Cancellariidae, bien que monophylétique et soutenue dans l'ensemble des arbres reconstruits, voit sa position changer selon les arbres obtenus : dans les arbres NT123 et NT12, les Cancellariidae se placent hors des Neogastropoda avec un nœud soutenu, alors que dans les arbres AA, les Cancellariidae sont placés en groupe-frère des autres Neogastropoda avec un nœud non soutenu (Figures 20-21-22 et 23). Dans tous les cas, les Cancellariidae seraient exclus des Volutoidea, pour constituer la superfamille des Cancellarioidea (M.-V. Modica et al., 2011; Petit & Harasewych, 2005; Verhecken, 2007).

Les membres des familles Cystiscidae et Marginellidae se mélangent dans les arbres (Annexe 3), mais avant de proposer des transferts de genres d'une famille à une autre, une révision des identifications des spécimens séquencés est à envisager, les spécimens de ces familles étant souvent difficiles à identifier.

Conoidea

Les résultats obtenus suggèrent que plusieurs nouvelles familles devraient être créées au sein des Conoidea :

- une famille pour les genres *Antiplanes, Leucosyrinx* et *Abyssocomitas,* actuellement placés dans les Pseudomelatomidae mais isolés dans une lignée indépendante des Pseudomelatomidae dans tous les arbres (Annexe 3) ;

- jusqu'à quatre nouvelles familles pour accommoder les différentes lignées de Borsoniidae retrouvées dans les arbres (Annexe 3).

De plus, un certain nombre de genres devraient être transférés dans d'autres familles : *Pleurotomoides* (de Clathurellidae vers Raphitomidae), *Paraclathurella* (de Clathurellidae vers Mangeliidae), *Cymakra* (de Mitromorphidae vers Borsoniidae *sensu lato*), *Paraspirotropis* (de Mangeliidae vers Borsoniidae *sensu lato*), *Vexitomina* (de Horaiclavidae vers Pseudomelatomidae), *Austrodrillia* (de Horaiclavidae vers Pseudomelatomidae), *Nquma* (de Horaiclavidae vers Pseudomelatomidae), *Buchema* (de Horaiclavidae vers Pseudomelatomidae), *Darbya* (de Borsoniidae vers Drilliidae), *Fusiturricula* (de Drilliidae vers Pseudomelatomidae) et *Plicisyrinx* (de Pseudomelatomidae vers Horaiclavidae). En outre, l'espèce *Fusiturris pluteata* ne se place pas avec l'espèce-type du genre *Fusiturris* (*F. similis*), et nécessiterait la création d'un nouveau genre au sein de Clavatulidae (Annexe 3).

« Unassigned taxa »

Dans la classification actuelle (WoRMS), 4 taxons ne sont pas assignés à des superfamilles :

- le genre *Enigmavasum* qui, comme expliqué plus haut, pourrait être placé dans les Costellariidae ou constituer une nouvelle famille (Annexe 3) ;

- les familles Harpidae, Babyloniidae et Strepsiduridae, qui correspondent toutes à des lignées indépendantes des autres superfamilles dans les arbres obtenus. Ces trois familles pourraient éventuellement correspondre à trois nouvelles super-familles (Figures 20-21-22 et 23).

1.5. CONCLUSIONS

Les analyses phylogénétiques que nous avons produites vont donc engendrer des changements significatifs dans la classification des néogastéropodes. Ces changements se feront à plusieurs niveaux taxonomiques, avec jusqu'à 6 nouvelles super-familles et 10 nouvelles familles à créer. Le transfert d'une famille entre deux super-familles, ainsi que 18 transferts de genres d'une famille à une autre, sont aussi à considérer.

Nous avons pu mettre en évidence une contradiction entre d'un côté les jeux de données AA, et de l'autre les jeux de données NT123 (Figures 20-21-22 et 23). Les jeux de données NT12 sont congruents avec les jeux de données AA pour les nœuds récents, mais le sont aussi avec les jeux de données NT123 pour les nœuds anciens (en particulier pour les Cancellariidae). Ces résultats suggèrent que la suppression de la 3ème base des codons a permis de réduire l'homoplasie pour les nœuds les plus récents, ce qui expliquerait la congruence avec les jeux de données AA. Cependant, l'homoplasie demeure trop importante pour les nœuds plus anciens, ce qui expliquerait la congruence avec les jeux de données NT123. Même si la topologie AA (avec les Cancellariidae groupe-frère des autres Néogastéropodes) est peu soutenue, elle semblerait plus pertinente, la topologie alternative (NT123 et NT12) étant potentiellement le résultat de phénomènes d'homoplasie. De plus, les données morphologiques et anatomiques à notre disposition suggèrent que les Cancellariidae sont bien des néogastéropodes (voir discussion dans (A. E. Fedosov et al., in press).

Malgré l'effort d'échantillonnage important réalisé pour obtenir cet arbre, il serait nécessaire et important d'inclure des taxons manquants, en particulier :

- le genre *Ptychatractus*, genre-type des Ptychatractidae, qui permettrait d'assigner ce nom de famille à une des lignées qui incluent des genres de Ptychatractidae ;

- les Pseudolividae : les deux spécimens de Pseudolividae, *Triumphis* et *Macron*, présents dans les arbres ne sont en fait pas des Pseudolividae. Les spécimens du genre *Pseudoliva*, genre-type des Pseudolividae, ont été intégrés au jeu de données mais le séquençage n'a pas fonctionné. Cependant, Kantor et al. (Y. Kantor et al., 2017) et ((A. E. Fedosov et al., in press), Annexe 1) ont montré que les Pseudolividae (représenté par le genre *Pseudoliva*) constituent une famille au sein des Olivoidea ;

- un spécimen de Strepsiduridae de meilleure qualité nous permettrait d'obtenir plus d'exons séquencés, le spécimen intégré aux arbres n'ayant que peu d'exons (71), ce qui explique vraisemblablement le peu de soutien pour le positionnement de ce taxon dans les arbres (Figure 20-21-22 et 23).

2. RAPHITOMIDAE

Comme expliqué dans le chapitre 2, nous voulions tester si le jeu d'exons utilisé pouvait résoudre à la fois des relations phylogénétiques profondes (Neogastropoda) et plus récentes, comme au sein de la famille des Raphitomidae.

Nous avons donc fait une sous-sélection parmi les 1728 spécimens que nous avons séquencé dans le jeu de données global, pour reconstruire une phylogénie de la famille des Raphitomidae. C'est une famille qui est encore à l'heure actuelle méconnue et il est nécessaire d'avoir une phylogénie moléculaire mise à jour. De plus, c'est une famille très diversifiée au sein des Conoidea (875 espèces valides listées dans WoRMS), avec une grande quantité d'espèces qui restent à décrire (résultats issus des données de séquençage du gène *cox1*, réalisés dans le cadre du projet HYPERDIVERSE). C'est aussi une famille très variable d'un point de vue morphologique, avec par exemple une grande diversité de protoconques (ce qui est unique chez les Conoidea, voire chez les néogastéropodes), et d'un point de vue anatomique, avec des lignées qui ont perdu secondairement la radula et/ou l'appareil venimeux (A. Fedosov & Kantor, 2007; Y. I. Kantor & Taylor, n.d.).

De nombreux spécimens présents dans les collections du MNHN n'ont pas encore été attribués à un nom d'espèce, voire de genre dans de nombreux cas. Avec l'aide d'un taxonomiste spécialiste du groupe, Peter Stahlschmidt, nous avons réalisé une sélection de spécimens à un rang générique afin d'avoir plusieurs spécimens par genre, dans le but de tester la monophylie des genres et éventuellement d'attribuer des noms de genres aux spécimens « inconnus » actuels. Nous avons sélectionné un ensemble de 331 spécimens afin de reconstruire cette phylogénie intra-familiale (voir chapitre 2 pour les critères de sélection des spécimens).

2.1. JEUX DE DONNEES ET ANALYSES PHYLOGENETIQUES

Comme pour les arbres reconstruits pour les néogastéropodes, deux jeux de données ont été utilisés : un jeu regroupant les spécimens pour lesquels au moins 700 exons avaient été capturés et un regroupant les spécimens pour lesquels au moins 50 exons avaient été capturés. Le jeu de données 700 exons inclut 1065 exons capturés par au minimum 100 spécimens différents, tandis que le jeu de données 50 exons inclut 1066 exons, sur les 1125 exons qui ont été dessinés pour le projet. Le jeu de données « 700 exons » inclut 316 spécimens, plus les exons du génome de

référence de *Conus ventricosus*, donc 317 spécimens au total. Le jeu de données « 50 exons » inclut 323 spécimens (322 spécimens + les exons du génome de référence).

Comme pour le jeu de données des néogastéropodes, nous avons construit trois sous-jeux de données à l'aide du script Python que j'ai décrit dans le chapitre 2. Nous avons également réalisé les mêmes analyses phylogénétiques que pour les jeux de données néogastéropodes (voir ci-dessus). Nous avons reconstruit les arbres phylogénétiques avec IQTree et créé 6 jeux de données différents (NT123-700, NT123-50, NT12-700, NT12-50, AA-700 et AA-50). Nous avons ensuite analysé les jeux de données 700 exons avec des modèles non partitionnés (NP), partitionnés (P) et partitionnés avec un regroupement des exons avec les mêmes vitesses d'évolution (P+Merge). Pour les jeux de données 50 exons, seules les analyses NP et P ont été réalisées (Tableau 9). Enfin, les arbres de gènes qui nous ont permis de reconstruire les super-arbres avec ASTRAL ne seront pas analysés car nous avons obtenu des résultats similaires aux arbres reconstruits pour les néogastéropodes, à savoir des topologies non congruentes ainsi que des supports de branche de mauvaise qualité.

Tableau 9 : Résultats des analyses phylogénétiques pour les Raphitomidae.

Les valeurs en gras des colonnes LogLikelihood, AIC, AICc et BIC sont les meilleures valeurs de chaque sous-jeu de données (NT123-700, NT12-700, AA-700, NT123-50, NT12-50 et AA-50).

taxon	Dataset samples	NT/AA	Partitioned	software	Best Model	LogLikelihood	AIC	AICc	BIC	Free parameters
Raphito	Raphito > 700 exons	NT123	NP	IQ-Tree	GTR+F+I+R11	**-8398065.2835**	16797450.5670	16797454.4434	16804266.4905	660
Raphito	Raphito > 700 exons	NT123	P	IQ-Tree	N.A	-8347118.5894	16726189.1789	16728622.8177	16891175.8374	15976
Raphito	Raphito > 700 exons	NT123	P+Merge	IQ-Tree	N.A	-8360010.5037	**16725991.0073**	**16726071.0333**	**16756817.5706**	2985
Raphito	Raphito > 700 exons	NT12	NP	IQ-Tree	GTR+F+I+R7	**-2619911.7634**	5241127.5269	5241133.1910	5247598.5936	652
Raphito	Raphito > 700 exons	NT12	P	IQ-Tree	N.A	-2582302.9655	**5192311.9310**	**5195110.9577**	5329802.2498	13853
Raphito	Raphito > 700 exons	NT12	P+Merge	IQ-Tree	N.A	-2595527.1559	5195746.3117	5195820.3976	**5219030.2421**	2346
Raphito	Raphito > 700 exons	AA	NP	IQ-Tree	JTT+F+I+R9	**-2489801.3903**	4980936.7807	4980948.7442	**4987091.3862**	667
Raphito	Raphito > 700 exons	AA	P	IQ-Tree	N.A	-2476206.2758	**4963470.5516**	**4964348.8513**	5014488.2635	5529
Raphito	Raphito > 700 exons	AA	P+Merge	IQ-Tree	N.A	-2480827.9312	4964399.8624	4964450.9258	4977059.7106	1372
Raphito	Raphito > 50 exons	NT123	NP	IQ-Tree	GTR+F+I+R11	**-8459714.4576**	16920772.9151	16920776.9316	**16927713.1485**	672
Raphito	Raphito > 50 exons	NT123	P	IQ-Tree	N.A	-8408704.6649	**16849513.3298**	**16851969.5566**	17015294.0224	16052
Raphito	Raphito > 50 exons	NT12	NP	IQ-Tree	GTR+F+I+R7	**-2637860.9925**	5277049.9850	5277055.8565	**5283640.5292**	664
Raphito	Raphito > 50 exons	NT12	P	IQ-Tree	N.A	-2600283.6167	**5228309.2334**	**5231114.1475**	5365986.0997	13871
Raphito	Raphito > 50 exons	AA	NP	IQ-Tree	JTT+F+I+R9	**-2507017.7785**	5015393.5571	5015405.9494	**5021659.2785**	679
Raphito	Raphito > 50 exons	AA	P	IQ-Tree	N.A	-2493312.9611	**4997809.9221**	**4998708.6077**	5049412.1523	5592

Nous avons un total de 15 analyses phylogénétiques qui ont été réalisées pour la famille des Raphitomidae. Les topologies sont globalement assez stables entre les différentes analyses, cependant, les supports de branches ne sont en moyenne pas aussi robustes que pour les arbres des néogastéropodes, ce qui pourrait s'expliquer par le fait que cette famille est plus récente, avec *a priori* une apparition très rapide des différents genres et espèces. Trois clades principaux sont retrouvés (Figure 24) : un premier clade (**Clade A**), avec des spécimens collectés en milieu profond, un second clade (**Clade B**), avec principalement des spécimens côtiers (voir ci-dessous pour plus de détails) et enfin un troisième clade (**Clade C**), avec également des spécimens profonds. La position de ces trois clades principaux dans l'arbre suggère que la colonisation du milieu côtier s'est faite à partir d'une lignée profonde. C'est un phénomène rarement observé, car c'est plutôt le schéma inverse qui est retrouvé habituellement : une colonisation du milieu profond à partir du milieu côtier (Jacobs & Lindberg, 1998; Lorion et al., 2009).

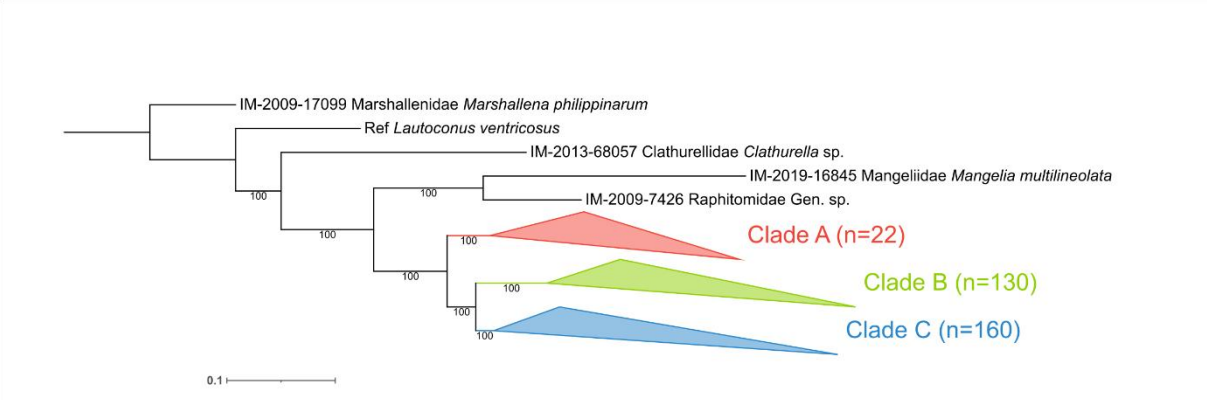


Figure 24 : Phylogénie des Raphitomidae, avec les trois clades principaux retrouvés dans toutes les analyses.

Le triangle rouge se compose des spécimens de milieux profonds, le triangle vert se compose des spécimens de milieux côtiers et enfin le triangle bleu regroupe les spécimens également de milieux profonds.

2.2. ARBRE PHYLOGENETIQUE AA-700-NP

Afin de faciliter la lecture des résultats, je vais présenter seulement l'arbre obtenu avec les données AA-700, avec le modèle NP. Les autres arbres sont disponibles en Annexe 3.

Dans cet arbre, nous avons un spécimen identifié comme un Raphitomidae Gen. sp. qui se place en groupe externe de l'ensemble des autres Raphitomidae. Ce spécimen est groupé avec un spécimen de Mangeliidae que nous avons utilisé afin d'enraciner l'arbre. La position de ce spécimen est inattendue, et nécessitera une vérification de son identification (grâce à la coquille ou la radula).

Clade A : dans ce groupe, composé de 22 spécimens collectés en milieu profond, les genres sont globalement monophylétiques (Figure 25).

- *Risomodaphnella* : ce genre est monophylétique avec 4 spécimens, il se place en groupe-frère d'un groupe comprenant un spécimen du genre *Pleurotomella* et 2 spécimens identifiés comme Gen. sp.

- *Veprecula* : ce genre est monophylétique avec 2 spécimens, il se place en groupe-frère du genre *Teretia* qui est aussi monophylétique et composé de 2 spécimens.

Ensuite vient un groupe composé de 4 spécimens, deux étant identifiés comme Gen. sp. et deux identifiés comme cf. *Teretia* sp.

Le seul représentant du genre *Tatcheriasyrinx* se place en groupe-frère de deux spécimens identifiés comme Gen. sp. et Gen. *polyacantha*. Enfin, 2 spécimens du genre *Famelica*, qui est polyphylétique, se placent en groupe-frère de ce regroupement.

L'unique spécimen du genre *Glaciotomella* se place en groupe-frère des groupes que j'ai cités plus haut. Un autre spécimen du genre *Famelica* se place également en groupe-frère de l'ensemble des spécimens de clade A.

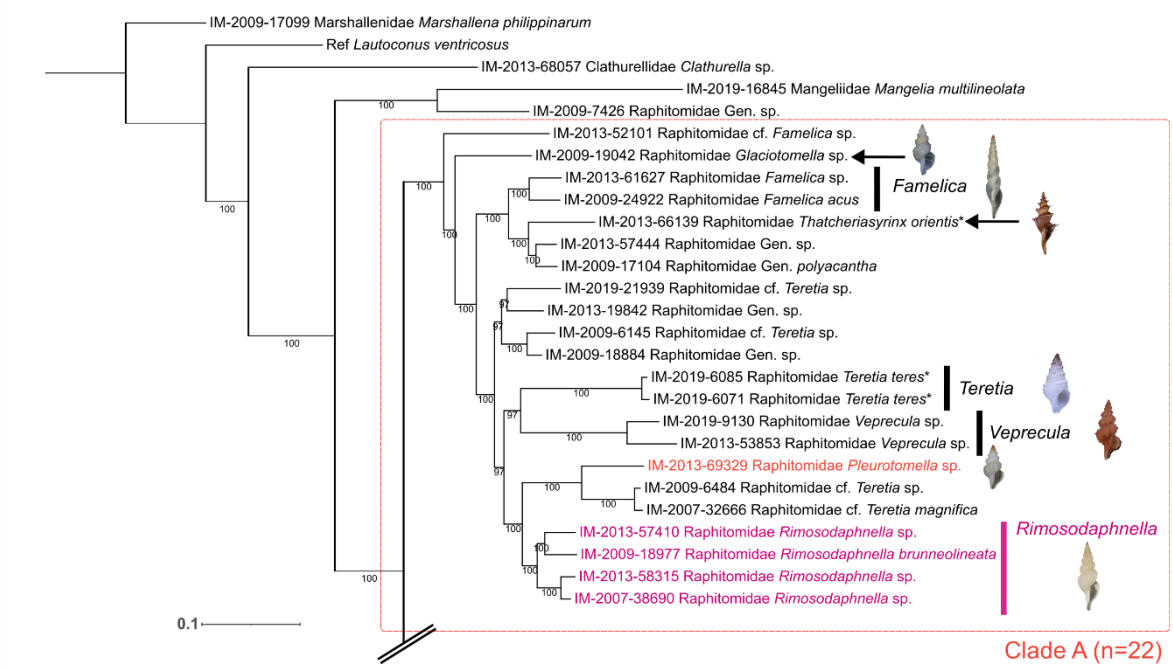


Figure 25 : Relations phylogénétiques au sein du clade A.

Les barres verticales noires regroupent chaque genre monophylétique. La barre verticale de couleur regroupe les spécimens d'un même genre qui est polyphylétique. Les flèches noires indiquent les genres qui sont composés d'un seul spécimen décrit présent dans cet arbre.

Clade B : dans ce groupe, composé de 130 spécimens côtiers, on retrouve les genres (Figure 26) :

- *Hemilienardia* : ce genre est monophylétique avec 8 spécimens, il se place en groupe-frère de deux spécimens du genre *Pleurotomella* ;

- *Raphitoma* : ce genre est monophylétique, il se compose de 6 spécimens et se place en groupe-frère du genre *Cyrillia* qui est aussi monophylétique et qui comprend 3 spécimens ;

- *Leufroyia* : ce genre est monophylétique et se compose de 2 spécimens. Il se place en groupe-frère du genre *Daphnella* qui est aussi monophylétique et qui comprend 2 spécimens également ;

- *Tritonoturris* : ce genre est polyphylétique et comprend 9 spécimens répartis en 3 groupes distincts. Un spécimen se place avec l'unique spécimen du genre *Diaugasma* ; un clade de 4

spécimens monophylétiques se place en groupe-frère de 2 spécimens identifiés comme cf. *Asperdaphne* sp ; enfin, un clade de 4 spécimens dans lequel s'intègre un spécimen identifié comme cf *Asperdaphne* sp ;

- *Pseudodaphnella* : nous avons 40 spécimens de ce genre qui est polyphylétique, avec un clade de 18 spécimens, un autre clade indépendant regroupant 7 spécimens, en groupe-frère d'un clade incluant un spécimen identifié comme Gen. sp., un spécimen de *Clathromangelia,* un spécimen identifié comme cf. *Thetidos* sp. et de 2 spécimens du genre *Exomilus* (monophylétique), et un clade de 18 spécimens de *Pseudodaphnella*, groupe-frère du reste.

Nous avons ensuite un clade de 6 spécimens de *Pseudodaphnella* dans lequel s'intègre 4 spécimens identifiés comme cf. *Austrodaphnella* sp. ; cf. *Hemilienardia idiomorpha* ; Gen. sp. et cf. *Pseudodaphnella reeveana*qui. Les spécimens du genre *Thetidos*, qui est monophylétique, se placent en groupe-frère de cet ensemble.

Dans un autre regroupement au sein du grand clade B, nous avons 5 spécimens de *Pseudodaphnella* monophylétiques. Ils intègrent 2 autres spécimens, l'un identifié Gen. sp. et un second identifié cf. *Neopleurotomoides* sp. Cet ensemble se place en groupe-frère d'un groupe de 2 spécimens de *Hemidaphne*, monophylétique. Tout ce groupe que je viens de décrire est groupe-frère de l'unique spécimen du genre *Microdaphne* présent dans cet arbre.

Dans un autre sous-clade, un groupe de 3 spécimens de *Pseudodaphnella* forme un clade avec un spécimen de *Glyphostomoides*. Ce groupe est placé en groupe-frère d'un sous-clade regroupant l'unique spécimen de *Kuroshiodapne*, l'unique spécimen *d'Austrodaphnella* ainsi que l'un des deux spécimens de *Pagodidaphne*. Enfin, le dernier spécimen de ce groupe est identifié comme Gen. sp.

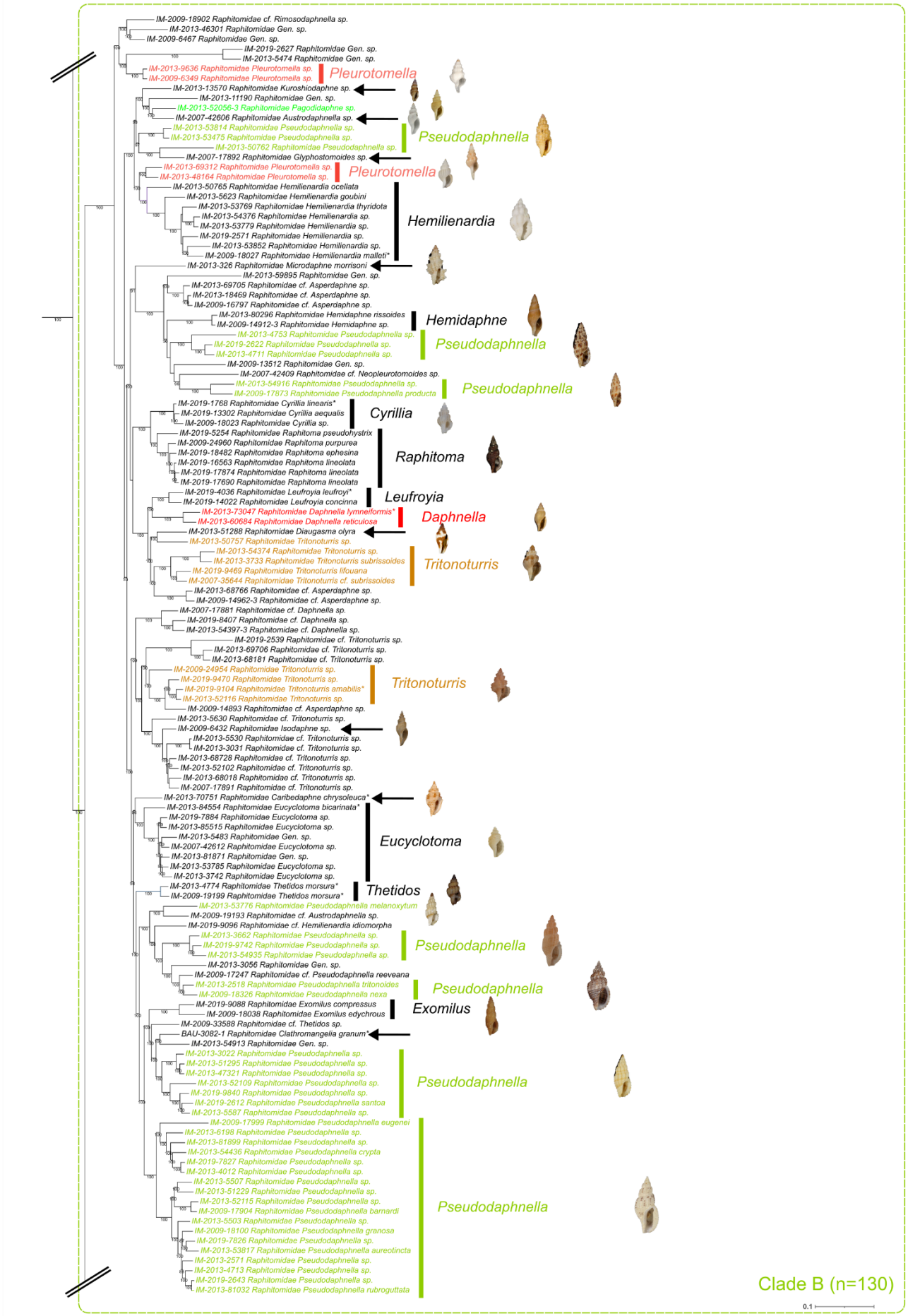
- *Eucyclotoma* : ce genre est monophylétique et comprend 6 spécimens. Ce clade intègre deux spécimens identifiés comme Gen. sp. Ils sont groupe-frère du seul représentant du genre *Caribedaphne*. Cet ensemble se place en groupe-frère du grand ensemble *Pseudodaphnella, Clathromangelia*, *Exomilus* et *Thetidos*.

- *Isodaphne* : l'unique représentant de ce genre se place au sein d'un sous-clade qui regroupe au total 8 spécimens. Les autres spécimens sont tous identifiés comme cf. *Tritonoturris* sp.

Au sein de ce clade B, un clade se place en groupe-frère de toutes les autres lignées que j'ai citées plus haut. Dans ce groupe, on retrouve le genre *Pleurotomella* qui est monophylétique avec 2 spécimens, ainsi que 4 spécimens identifiés comme Gen. sp. ainsi qu'un spécimen identifié comme cf. *Rimosodaphnella* sp.

Figure 26 : Relations phylogénétiques au sein du clade B.

Les flèches noires indiquent les genres pour lesquels un seul spécimen décrit est présent. Les barres verticales noires indiquent les genres monophylétiques et les barres de couleurs indiquent les genres polyphylétiques.


IM-2009-18902 Raphitomidae cf. Rimosodaphnella sp.
IM-2013-46301 Raphitomidae Gen. sp.
IM-2009-6467 Raphitomidae Gen. sp.
IM-2019-2627 Raphitomidae Gen. sp.
IM-2013-5474 Raphitomidae Gen. sp.
IM-2013-9636 Raphitomidae Pleurotomella sp.
IM-2009-6349 Raphitomidae Pleurotomella sp.
Pleurotomella
IM-2013-13570 Raphitomidae Kuroshiodaphne sp.
IM-2013-11190 Raphitomidae Gen. sp.
IM-2013-52056-3 Raphitomidae Pagodidaphne sp.
IM-2007-42606 Raphitomidae Austrodaphnella sp.
IM-2013-53814 Raphitomidae Pseudodaphnella sp.
IM-2013-53475 Raphitomidae Pseudodaphnella sp.
IM-2013-50762 Raphitomidae Pseudodaphnella sp.
Pseudodaphnella
IM-2007-17892 Raphitomidae Glyphostomoides sp.
IM-2013-69312 Raphitomidae Pleurotomella sp.
IM-2013-48164 Raphitomidae Pleurotomella sp.
Pleurotomella
IM-2013-50765 Raphitomidae Hemilienardia ocellata
IM-2013-5623 Raphitomidae Hemilienardia goubini
IM-2013-53769 Raphitomidae Hemilienardia thyridota
IM-2013-54376 Raphitomidae Hemilienardia sp.
IM-2013-53779 Raphitomidae Hemilienardia sp.
IM-2019-2571 Raphitomidae Hemilienardia sp.
IM-2013-53852 Raphitomidae Hemilienardia sp.
IM-2009-18027 Raphitomidae Hemilienardia malleti*
Hemilienardia
IM-2013-326 Raphitomidae Microdaphne morrisoni
IM-2013-59895 Raphitomidae Gen. sp.
IM-2013-69705 Raphitomidae cf. Asperdaphne sp.
IM-2013-18469 Raphitomidae cf. Asperdaphne sp.
IM-2009-16797 Raphitomidae cf. Asperdaphne sp.
IM-2013-80296 Raphitomidae Hemidaphne rissoides
IM-2009-14912-3 Raphitomidae Hemidaphne sp.
Hemidaphne
IM-2013-4753 Raphitomidae Pseudodaphnella sp.
IM-2019-2622 Raphitomidae Pseudodaphnella sp.
IM-2013-4711 Raphitomidae Pseudodaphnella sp.
Pseudodaphnella
IM-2009-13512 Raphitomidae Gen. sp.
IM-2007-42409 Raphitomidae cf. Neopleurotomoides sp.
IM-2013-54916 Raphitomidae Pseudodaphnella sp.
IM-2009-17873 Raphitomidae Pseudodaphnella producta
Pseudodaphnella
IM-2019-1768 Raphitomidae Cyrillia linearis*
IM-2019-13302 Raphitomidae Cyrillia aequalis
IM-2009-18023 Raphitomidae Cyrillia sp.
Cyrillia
IM-2019-5254 Raphitomidae Raphitoma pseudohystrix
IM-2009-24960 Raphitomidae Raphitoma purpurea
IM-2019-18482 Raphitomidae Raphitoma ephesina
IM-2019-16563 Raphitomidae Raphitoma lineolata
IM-2019-17874 Raphitomidae Raphitoma lineolata
IM-2019-17690 Raphitomidae Raphitoma lineolata
Raphitoma
IM-2019-4036 Raphitomidae Leufroyia leufroyi*
IM-2019-14022 Raphitomidae Leufroyia concinna
Leufroyia
IM-2013-73047 Raphitomidae Daphnella lymneiformis*
IM-2013-60684 Raphitomidae Daphnella reticulosa
Daphnella
IM-2013-51288 Raphitomidae Diaugasma olyra
IM-2013-50757 Raphitomidae Tritonoturris sp.
IM-2013-54374 Raphitomidae Tritonoturris sp.
IM-2013-3733 Raphitomidae Tritonoturris subrissoides
IM-2019-9469 Raphitomidae Tritonoturris lifouana
IM-2007-35644 Raphitomidae Tritonoturris cf. subrissoides
Tritonoturris
IM-2013-68766 Raphitomidae cf. Asperdaphne sp.
IM-2009-14962-3 Raphitomidae cf. Asperdaphne sp.
IM-2007-17881 Raphitomidae cf. Daphnella sp.
IM-2019-8407 Raphitomidae cf. Daphnella sp.
IM-2013-54397-3 Raphitomidae cf. Daphnella sp.
IM-2019-2539 Raphitomidae cf. Tritonoturris sp.
IM-2013-69706 Raphitomidae cf. Tritonoturris sp.
IM-2013-68181 Raphitomidae cf. Tritonoturris sp.
IM-2009-24954 Raphitomidae Tritonoturris sp.
IM-2019-9470 Raphitomidae Tritonoturris sp.
IM-2019-9104 Raphitomidae Tritonoturris amabilis*
IM-2013-52116 Raphitomidae Tritonoturris sp.
Tritonoturris
IM-2009-14893 Raphitomidae cf. Asperdaphne sp.
IM-2013-5630 Raphitomidae cf. Tritonoturris sp.
IM-2009-6432 Raphitomidae Isodaphne sp.
IM-2013-5530 Raphitomidae cf. Tritonoturris sp.
IM-2013-3031 Raphitomidae cf. Tritonoturris sp.
IM-2013-68728 Raphitomidae cf. Tritonoturris sp.
IM-2013-52102 Raphitomidae cf. Tritonoturris sp.
IM-2013-68018 Raphitomidae cf. Tritonoturris sp.
IM-2007-17891 Raphitomidae cf. Tritonoturris sp.
IM-2013-70751 Raphitomidae Caribedaphne chrysoleuca*
IM-2013-84554 Raphitomidae Eucyclotoma bicarinata*
IM-2019-7884 Raphitomidae Eucyclotoma sp.
IM-2013-85515 Raphitomidae Eucyclotoma sp.
IM-2013-5483 Raphitomidae Gen. sp.
IM-2007-42612 Raphitomidae Eucyclotoma sp.
IM-2013-81871 Raphitomidae Gen. sp.
IM-2013-53785 Raphitomidae Eucyclotoma sp.
IM-2013-3742 Raphitomidae Eucyclotoma sp.
Eucyclotoma
IM-2013-4774 Raphitomidae Thetidos morsura*
IM-2009-19199 Raphitomidae Thetidos morsura*
Thetidos
IM-2013-53776 Raphitomidae Pseudodaphnella melanoxytum
IM-2009-19193 Raphitomidae cf. Austrodaphnella sp.
IM-2019-9096 Raphitomidae cf. Hemilienardia idiomorpha
IM-2013-3662 Raphitomidae Pseudodaphnella sp.
IM-2019-9742 Raphitomidae Pseudodaphnella sp.
IM-2013-54935 Raphitomidae Pseudodaphnella sp.
Pseudodaphnella
IM-2013-3056 Raphitomidae Gen. sp.
IM-2009-17247 Raphitomidae cf. Pseudodaphnella reeveana
IM-2013-2518 Raphitomidae Pseudodaphnella tritonoides
IM-2009-18326 Raphitomidae Pseudodaphnella nexa
Pseudodaphnella
IM-2019-9088 Raphitomidae Exomilus compressus
IM-2009-18038 Raphitomidae Exomilus edychrous
Exomilus
IM-2009-33588 Raphitomidae cf. Thetidos sp.
BAU-3082-1 Raphitomidae Clathromangelia granum*
IM-2013-54913 Raphitomidae Gen. sp.
IM-2013-3022 Raphitomidae Pseudodaphnella sp.
IM-2013-51295 Raphitomidae Pseudodaphnella sp.
IM-2013-47321 Raphitomidae Pseudodaphnella sp.
IM-2013-52109 Raphitomidae Pseudodaphnella sp.
IM-2019-9840 Raphitomidae Pseudodaphnella sp.
IM-2019-2612 Raphitomidae Pseudodaphnella santoa
IM-2013-5587 Raphitomidae Pseudodaphnella sp.
Pseudodaphnella
IM-2009-17999 Raphitomidae Pseudodaphnella eugenei
IM-2013-6198 Raphitomidae Pseudodaphnella sp.
IM-2013-81899 Raphitomidae Pseudodaphnella sp.
IM-2013-54436 Raphitomidae Pseudodaphnella crypta
IM-2019-7827 Raphitomidae Pseudodaphnella sp.
IM-2013-4012 Raphitomidae Pseudodaphnella sp.
IM-2013-5507 Raphitomidae Pseudodaphnella sp.
IM-2013-51229 Raphitomidae Pseudodaphnella sp.
IM-2013-52115 Raphitomidae Pseudodaphnella sp.
IM-2009-17904 Raphitomidae Pseudodaphnella barnardi
IM-2013-5503 Raphitomidae Pseudodaphnella sp.
IM-2009-18100 Raphitomidae Pseudodaphnella granosa
IM-2019-7826 Raphitomidae Pseudodaphnella sp.
IM-2013-53817 Raphitomidae Pseudodaphnella aureotincta
IM-2013-2571 Raphitomidae Pseudodaphnella sp.
IM-2013-4713 Raphitomidae Pseudodaphnella sp.
IM-2019-2643 Raphitomidae Pseudodaphnella sp.
IM-2013-81032 Raphitomidae Pseudodaphnella rubroguttata
Pseudodaphnella
Clade B (n=130)
0.1

Clade C : ce clade regroupe 160 spécimens de Raphitomidae profonds (Figure 27).

- *Magnella* : ce genre, représenté par 5 spécimens, inclut également le seul représentant du genre *Taranidaphne*. Cet ensemble de 6 spécimens se place en groupe-frère des 2 spécimens du genre *Taranis* qui est monophylétique.

Dans ce clade, nous avons plusieurs genres monophylétiques : le genre *Eubela*, composé de 5 spécimens ; le genre *Miowateria*, composé de 4 spécimens ; le genre *Cryptodaphne* avec 3 spécimens ; le genre *Teretiopsis* qui est composé de 4 spécimens ; le genre *Ootomella*, qui comprend 6 spécimens ; le genre *Paradaphne* avec 5 spécimens qui se placent en groupe-frère d'un spécimen identifié comme cf. *Paradaphne* sp ; le genre *Pagodibela*, qui comprend 4 spécimens ; le genre *Acanthodaphne* avec 4 spécimens ; le genre *Leiosyrinx* qui est composé de 2 spécimens ; le genre *Gladiobela* avec 2 spécimens, ce groupe comprend également 2 autres spécimens identifiés comme Gen. sp.

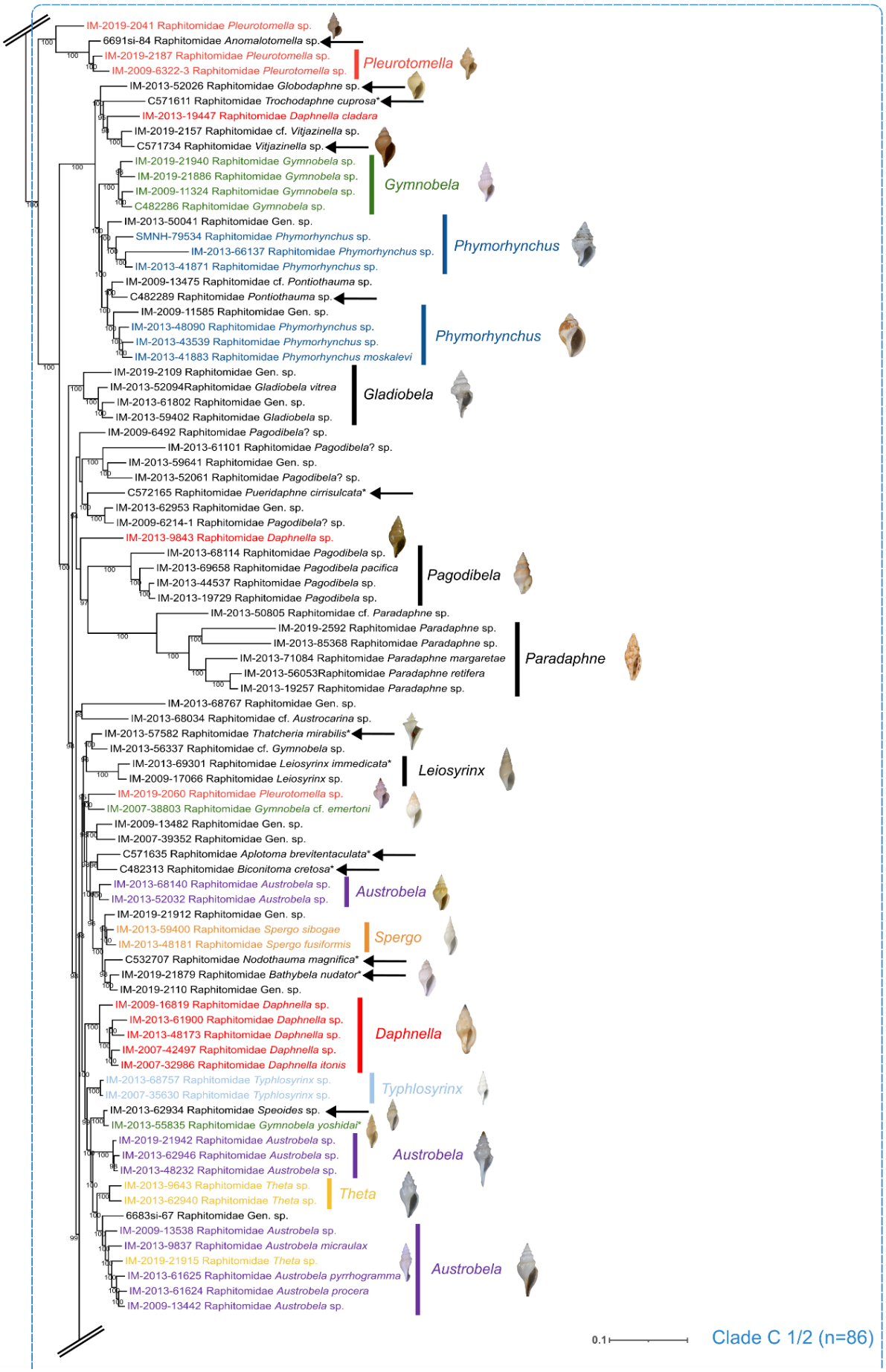
Plusieurs autres genres sont retrouvés quant à eux polyphylétiques : le genre *Gymnobela* qui est composé de 12 spécimens répartis en 7 lignées ; le genre *Austrobela* qui comprend 10 spécimens répartis en 4 lignées ; le genre *Theta* avec 3 spécimens répartis en 2 lignées (dont une incluse dans une lignée d'Austrobela) ; le genre *Typhlosyrinx* avec 3 représentants dans 2 lignées également ; le genre *Daphnella* représenté par 11 spécimens dans 7 lignées; le genre *Lusitanops* qui comprend 2 spécimens ; le genre *Spergo* avec 2 spécimens qui sont monophylétiques, tandis que le 3ème se place avec un spécimen du genre *Typhlosyrinx* ; le genre *Phymorhynchus* dont les 6 spécimens se placent en deux groupes distincts ; le genre *Buccinaria*, dont 4 spécimens sont monophylétiques mais les 2 autres spécimens se placent indépendamment dans l'arbre.

Nous avons dans cet arbre plusieurs genres qui ne sont représentés que par un seul spécimen. C'est le cas pour le genre *Speoides*, ce spécimen se groupe avec un spécimen de *Gymnobela*. Le genre *Bathybela* dont le spécimen se groupe avec un spécimen identifié comme Gen. sp. Le genre *Nodothauma* avec un spécimen seul mais en groupe-frère du seul spécimen de *Bathybela*. Le spécimen du genre *Biconitoma* se groupe avec le seul spécimen du genre *Aplotoma*. Le genre *Thatcheria*, dont le seul spécimen se groupe avec un spécimen identifié comme cf. *Gymnobela* sp. Le genre *Pueridaphne* avec un spécimen qui se place au sein d'un groupe comportant 6 spécimens, 4 étant identifiés comme *Pagodibela* ? et enfin 2 spécimens identifiés comme Gen. sp. Le genre *Pontiothauma* dont le spécimen se groupe avec 1 spécimen identifié comme cf. *Pontiothauma* sp. Le spécimen du genre *Vitjazinella* qui se groupe avec 1 spécimen identifié

comme cf. *Vitjazinella* sp. Le genre *Trochodaphne* dont le seul spécimen se place en groupe de l'unique spécimen du genre *Globodaphne*. Le genre *Anomalotomella*, qui est synonyme du genre *Pleurotomella*, dont l'unique spécimen se place au sein d'un groupe de 3 spécimens de *Pleurotomella*. Enfin, ce clade inclut un grand nombre de spécimens qui ne sont pas identifiés, même au rang du genre, et d'autres qui sont proches d'un genre décrit (« cf. »), mais qui ne se groupent pas avec les autres représentants de ces genres.

Figure 27 : Relations phylogénétiques au sein du clade C.

Les flèches noires indiquent les genres pour lesquels un seul spécimen décrit est représenté dans l'arbre. Les barres verticales colorées indiquent les genres polyphylétiques et les barres verticales noires indiquent les genres monophylétiques.


IM-2019-2041 Raphitomidae Pleurotomella sp.
6691si-84 Raphitomidae Anomalotomella sp.
IM-2019-2187 Raphitomidae Pleurotomella sp.
IM-2009-6322-3 Raphitomidae Pleurotomella sp.
Pleurotomella
IM-2013-52026 Raphitomidae Globodaphne sp.
C571611 Raphitomidae Trochodaphne cuprosa*
IM-2013-19447 Raphitomidae Daphnella cladara
IM-2019-2157 Raphitomidae cf. Vitjazinella sp.
C571734 Raphitomidae Vitjazinella sp.
IM-2019-21940 Raphitomidae Gymnobela sp.
IM-2019-21886 Raphitomidae Gymnobela sp.
IM-2009-11324 Raphitomidae Gymnobela sp.
C482286 Raphitomidae Gymnobela sp.
Gymnobela
IM-2013-50041 Raphitomidae Gen. sp.
SMNH-79534 Raphitomidae Phymorhynchus sp.
IM-2013-66137 Raphitomidae Phymorhynchus sp.
IM-2013-41871 Raphitomidae Phymorhynchus sp.
Phymorhynchus
IM-2009-13475 Raphitomidae cf. Pontiothauma sp.
C482289 Raphitomidae Pontiothauma sp.
IM-2009-11585 Raphitomidae Gen. sp.
IM-2013-48090 Raphitomidae Phymorhynchus sp.
IM-2013-43539 Raphitomidae Phymorhynchus sp.
IM-2013-41883 Raphitomidae Phymorhynchus moskalevi
Phymorhynchus
IM-2019-2109 Raphitomidae Gen. sp.
IM-2013-52094Raphitomidae Gladiobela vitrea
IM-2013-61802 Raphitomidae Gen. sp.
IM-2013-59402 Raphitomidae Gladiobela sp.
Gladiobela
IM-2009-6492 Raphitomidae Pagodibela? sp.
IM-2013-61101 Raphitomidae Pagodibela? sp.
IM-2013-59641 Raphitomidae Gen. sp.
IM-2013-52061 Raphitomidae Pagodibela? sp.
C572165 Raphitomidae Pueridaphne cirrisulcata*
IM-2013-62953 Raphitomidae Gen. sp.
IM-2009-6214-1 Raphitomidae Pagodibela? sp.
IM-2013-9843 Raphitomidae Daphnella sp.
IM-2013-68114 Raphitomidae Pagodibela sp.
IM-2013-69658 Raphitomidae Pagodibela pacifica
IM-2013-44537 Raphitomidae Pagodibela sp.
IM-2013-19729 Raphitomidae Pagodibela sp.
Pagodibela
IM-2013-50805 Raphitomidae cf. Paradaphne sp.
IM-2019-2592 Raphitomidae Paradaphne sp.
IM-2013-85368 Raphitomidae Paradaphne sp.
IM-2013-71084 Raphitomidae Paradaphne margaretae
IM-2013-56053Raphitomidae Paradaphne retifera
IM-2013-19257 Raphitomidae Paradaphne sp.
Paradaphne
IM-2013-68767 Raphitomidae Gen. sp.
IM-2013-68034 Raphitomidae cf. Austrocarina sp.
IM-2013-57582 Raphitomidae Thatcheria mirabilis*
IM-2013-56337 Raphitomidae cf. Gymnobela sp.
IM-2013-69301 Raphitomidae Leiosyrinx immedicata*
IM-2009-17066 Raphitomidae Leiosyrinx sp.
Leiosyrinx
IM-2019-2060 Raphitomidae Pleurotomella sp.
IM-2007-38803 Raphitomidae Gymnobela cf. emertoni
IM-2009-13482 Raphitomidae Gen. sp.
IM-2007-39352 Raphitomidae Gen. sp.
C571635 Raphitomidae Aplotoma brevitentaculata*
C482313 Raphitomidae Biconitoma cretosa*
IM-2013-68140 Raphitomidae Austrobela sp.
IM-2013-52032 Raphitomidae Austrobela sp.
Austrobela
IM-2019-21912 Raphitomidae Gen. sp.
IM-2013-59400 Raphitomidae Spergo sibogae
IM-2013-48181 Raphitomidae Spergo fusiformis
Spergo
C532707 Raphitomidae Nodothauma magnifica*
IM-2019-21879 Raphitomidae Bathybela nudator*
IM-2019-2110 Raphitomidae Gen. sp.
IM-2009-16819 Raphitomidae Daphnella sp.
IM-2013-61900 Raphitomidae Daphnella sp.
IM-2013-48173 Raphitomidae Daphnella sp.
IM-2007-42497 Raphitomidae Daphnella sp.
IM-2007-32986 Raphitomidae Daphnella itonis
Daphnella
IM-2013-68757 Raphitomidae Typhlosyrinx sp.
IM-2007-35630 Raphitomidae Typhlosyrinx sp.
Typhlosyrinx
IM-2013-62934 Raphitomidae Speoides sp.
IM-2013-55835 Raphitomidae Gymnobela yoshidai*
IM-2019-21942 Raphitomidae Austrobela sp.
IM-2013-62946 Raphitomidae Austrobela sp.
IM-2013-48232 Raphitomidae Austrobela sp.
Austrobela
IM-2013-9643 Raphitomidae Theta sp.
IM-2013-62940 Raphitomidae Theta sp.
Theta
6683si-67 Raphitomidae Gen. sp.
IM-2009-13538 Raphitomidae Austrobela sp.
IM-2013-9837 Raphitomidae Austrobela micraulax
IM-2019-21915 Raphitomidae Theta sp.
IM-2013-61625 Raphitomidae Austrobela pyrrhogramma
IM-2013-61624 Raphitomidae Austrobela procera
IM-2009-13442 Raphitomidae Austrobela sp.
Austrobela
0.1
Clade C 1/2 (n=86)


IM-2013-68126 Raphitomidae cf. Acamptodaphne sp.
IM-2007-39368 Raphitomidae cf. Acamptodaphne sp.
IM-2013-58422 Raphitomidae Daphnella sp.
IM-2019-21941 Raphitomidae Rimosodaphnella sp.
IM-2013-68759 Raphitomidae Rimosodaphnella sp.
IM-2013-5572 Raphitomidae Rimosodaphnella sp.
IM-2009-18890 Raphitomidae Rimosodaphnella sp.
Rimosodaphnella
IM-2013-68742 Raphitomidae Lusitanops mitrellaformis
IM-2013-61343 Raphitomidae Gymnobela sp.
IM-2013-60107 Raphitomidae Gymnobela sp.
Gymnobela
IM-2013-62961 Raphitomidae Spergo sp.
IM-2013-19473 Raphitomidae Typhlosyrinx sp.
IM-2013-9997 Raphitomidae Lusitanops sp.
IM-2013-52120 Raphitomidae Daphnella sp.
IM-2009-6010 Raphitomidae cf. Daphnella sp.
IM-2013-18910 Raphitomidae Gen. sp.
IM-2009-19008 Raphitomidae Pagodidaphne sp.
IM-2019-21937 Raphitomidae cf. Lusitanops sp.
IM-2013-44651 Raphitomidae cf. Lusitanops sp.
C519302 Raphitomidae Austrotheta crassidentata*
IM-2019-2082 Raphitomidae Gymnobela sp.
IM-2013-58328 Raphitomidae Gen. sp.
IM-2019-2049 Raphitomidae Teretiopsis sp.
IM-2009-13481 Raphitomidae Teretiopsis sp.
IM-2013-69704 Raphitomidae Teretiopsis sp.
IM-2013-61524 Raphitomidae Teretiopsis sp.
Teretiopsis
IM-2013-19998 Raphitomidae Buccinaria sp.
IM-2013-18610 Raphitomidae Buccinaria sp.
Buccinaria
IM-2013-19477 Raphitomidae cf. Buccinaria sp.
HORS-MNHN-NIWA-72848 Raphitomidae Gen. sp.
IM-2007-38195 Raphitomidae Ootomella cf. martini
IM-2013-18475 Raphitomidae Ootomella sp.
IM-2007-42622 Raphitomidae Ootomella jonkeri
IM-2013-19329 Raphitomidae Ootomella sp.
IM-2013-19306-1 Raphitomidae Ootomella sp.
IM-2009-29354 Raphitomidae Ootomella sp.
Ootomella
IM-2013-68086 Raphitomidae Gymnobela sp.
IM-2013-50242 Raphitomidae Gen. sp.
IM-2009-18323 Raphitomidae Gymnobela sp.
IM-2007-42426 Raphitomidae Gymnobela mitrodeta
Gymnobela
IM-2013-52100 Raphitomidae cf. Rimosodaphnella sp.
IM-2009-19069 Raphitomidae Gen. sp.
IM-2007-35067 Raphitomidae Buccinaria sp.
IM-2013-44619 Raphitomidae Buccinaria sp.
IM-2013-18706 Raphitomidae Buccinaria sp.
IM-2009-6438 Raphitomidae Buccinaria sp.
Buccinaria
IM-2013-68056 Raphitomidae Gen. sp.
IM-2009-6459 Raphitomidae Gen. sp.
IM-2013-9769 Raphitomidae Acanthodaphne sp.
IM-2009-17212 Raphitomidae Acanthodaphne sp.
IM-2013-4781 Raphitomidae Acanthodaphne sp.
IM-2013-16631 Raphitomidae Acanthodaphne sp.
Acanthodaphne
IM-2019-21945 Raphitomidae Gen. sp.
IM-2009-24926 Raphitomidae Gen. sp.
IM-2013-9738 Raphitomidae Cryptodaphne sp.
IM-2009-18992 Raphitomidae Cryptodaphne sp.
IM-2009-18327 Raphitomidae Cryptodaphne sp.
Cryptodaphne
IM-2013-52046 Raphitomidae Miowateria sp.
IM-2009-6461 Raphitomidae Miowateria sp.
IM-2013-9638 Raphitomidae Miowateria sp.
IM-2009-19058 Raphitomidae Miowateria sp.
Miowateria
IM-2019-21876 Raphitomidae Eubela sp.
IM-2013-59591 Raphitomidae Eubela sp.
IM-2019-21938 Raphitomidae Eubela sp.
IM-2013-45521 Raphitomidae Eubela sp.
IM-2013-45511 Raphitomidae Eubela sp.
Eubela
IM-2019-16997 Raphitomidae Taranis moerchii*
IM-2007-42296 Raphitomidae Taranis sp.
Taranis
IM-2013-5364 Raphitomidae Magnella sp.
IM-2009-19059-3 Raphitomidae Magnella sp.
IM-2019-21944 Raphitomidae Magnella sp.
IM-2013-44646 Raphitomidae Magnella sp.
IM-2019-2623 Raphitomidae Taranidaphne sp.
IM-2009-19032 Raphitomidae Magnella sp.
Magnella
0.1
Clade C 2/2 (n=74)

2.3. CHANGEMENTS POTENTIELS DANS LA CLASSIFICATION

- *Rimosodaphnella* : comme décrit précédemment, les spécimens de ce genre se répartissent en deux groupes de 4 spécimens chacun. Un groupe dans le clade A et un groupe dans le clade C. Ces spécimens ont été collectés dans des profondeurs allant de 420 à 770 mètres pour le clade A et entre 410 à 800 mètres pour les spécimens du clade C. Ils sont morphologiquement proches les uns des autres, et leur position est congruente entre les jeux de données AA-700, NT123-700 et NT12-700.

- *Eucyclotoma* : deux spécimens identifiés comme Gen. sp. (IM-2013-81871 et IM-2013-5483) sont compris dans le sous-clade regroupant tous les spécimens du genre *Eucyclotoma*. Ces deux spécimens, après une analyse morphologique, pourraient être attribués à ce genre.

- *Pseudodaphnella* : Les spécimens de ce groupe se séparent en 9 groupes distincts au sein du clade B. Dans un sous-clade regroupant 6 spécimens de *Pseudodaphnella*, il y a 4 spécimens (IM-2009-19193, IM-2019-9096, IM-2013-3056 et IM-2009-17247) qui se placent au sein de ce groupe. Ils sont identifiés comme cf. *Austrodaphnella* sp. ; cf. *Hemilienardia idiomorpha* ; Gen. sp. et cf. *Pseudodaphnella reeveana*. Ces 4 spécimens pourraient être attribués au genre *Pseudophnella*. Cependant, ce genre n'étant pas retrouvé monophylétique dans cet arbre phylogénétique, il serait sans doute nécessaire de réviser la taxonomie de ce genre. Dans un autre groupe, nous avons 5 spécimens de *Pseudodaphnella* qui se groupent avec 1 spécimen identifié comme Gen. sp. (IM-2009-13512) et un spécimen identifié comme cf. *Neopleurotomoides* sp. (IM-2007-42409).

- *Clathromangelia* : un spécimen identifié comme Gen. sp. se place avec le seul spécimen du genre *Clathromangelia*. Ce spécimen est l'espèce-type *Clathromangelia granum*, cela pourrait amener à donner ce nom de genre au spécimen encore non identifié.

- *Tritonoturris* : les 3 groupes de spécimens appartenant à ce genre sont tous présents dans le clade B. Un groupe a en son sein le spécimen de l'espèce-type *Tritonoturris amabilis*. Les spécimens qui n'ont pas encore été attribués à un genre et qui sont groupés avec les *Tritonoturris* devraient être évalués pour éventuellement être assignés à ce genre.

- *Pleurotomella* : ces spécimens sont présents en 4 groupes distincts dans l'arbre. Ils se positionnent dans les trois différents clades : A, B et C. Les 4 spécimens qui se regroupent dans le clade B se placent en deux groupes différents : un groupe de 2 spécimens (IM-2013-9636 et

IM-2009-6349) qui ont été collectés entre 690 et 990 mètres de profondeur, placés en groupe-frère de spécimens qui ont été collectés dans des zones de moins de 20 mètres de profondeur, et un groupe de deux autres spécimens (IM-2013-69312 et IM-2013-48164) qui ont été collectés entre 400 et 700 mètres de profondeur, qui se placent en groupe-frère des spécimens d'*Hemilienardia* qui est un genre côtier. Si la profondeur est un facteur de distinction des clades dans cette phylogénie, alors la position de ces spécimens peut paraître incohérente.

- *Pagodidaphne* : les 2 spécimens de ce genre se placent dans deux clades différents de l'arbre, un dans le clade B (IM-2013-52056) et l'autre dans le clade C (IM-2009-19008). Ils ont été collectés dans les mêmes profondeurs entre 400 et 450 mètres. Le spécimen IM-2013-52056 se place avec l'unique spécimen du genre *Austrodaphnella*. Ce spécimen a été également collecté dans un milieu profond à plus de 600 mètres. La position de ces spécimens n'est pas cohérente avec la distinction entre les clades profonds et le clade côtier. Le spécimen IM-2009-19008 se place avec un spécimen identifié comme Gen. sp. dont l'identification pourrait changer pour devenir *Pagodidaphne* sp.

- *Gladiobela* : les 2 spécimens identifiés comme Gen. sp. qui se groupent avec les *Gladiobela* devrait être attribués au genre *Gladiobela*.

2.4. CONCLUSIONS

L'arbre présenté se compose de 17 genres qui sont monophylétiques, 18 genres représentés par un seul spécimen, qui se groupent parfois avec des spécimens qui n'ont pas encore été assignés à un genre, et enfin 14 genres polyphylétiques parfois présents dans les trois clades A, B et C. C'est le cas pour les genres : *Pleurotomella*, qui inclut un total de 9 spécimens, avec 1 spécimen dans le clade A, 4 spécimens dans le clade B et enfin 4 spécimens dans le clade C ; le genre *Rimosodaphnella* dont 4 spécimens sont monophylétiques dans le clade A et 4 autres spécimens sont monophylétiques dans le clade C ; et enfin le genre *Pagodidaphne* avec 1 spécimen dans le clade B et 1 autre dans le clade C.

Les résultats phylogénétiques pour cette famille suggèrent qu'une révision taxonomique d'une grande partie des genres de Raphitomidae est à prévoir, et qu'un grand nombre de genres nouveaux sont à décrire. Ces résultats confirment, avec un jeu de données bien plus étendu et une meilleure résolution des nœuds, ce que les phylogénies de Raphitomidae préalablement

publiées avaient suggéré (Criscione et al., 2021; Fassio et al., 2019; A. E. Fedosov & Puillandre, 2012). Peter Stahlschmidt, spécialiste du groupe qui a identifié les spécimens inclus dans cette phylogénie, sera présent au MNHN au mois de mai prochain afin de poursuivre l'identification et la description des taxons.

Globalement, nous pouvons remarquer des supports de nœuds plus faibles par rapport à l'arbre des néogastéropodes. Cela peut s'expliquer par des problèmes liés à la variabilité des gènes, qui pourraient être plus adaptés pour la mise en évidence de relations phylogénétiques profondes. Il est également possible que le clade des Raphitomidae s'est diversifié rapidement, ce qui pourrait expliquer les branches très courtes que nous observons, notamment pour le clade C, et par conséquent une difficulté à résoudre ces relations phylogénétiques.

De manière identique aux reconstruction phylogénétique faites pour les néogastéropodes, nous avons deux topologies principales, l'une correspondant aux analyses faites avec les jeux de données AA et NT12, l'autre avec les topologies NT123 (Annexe 3). Comme pour les nœuds les plus récents dans la phylogénie des néogastéropodes, la suppression de la 3ème base de codon dans le jeu de données NT12 aurait donc permis de gérer l'homoplasie, et d'obtenir une topologie similaire à celle obtenue avec le jeu de données AA.

Conclusions et perspectives

1. PHYLOGENIES DES NEOGASTROPODA ET DES RAPHITOMIDAE

Au cours de cette thèse j'ai pu produire plusieurs phylogénies à des échelles taxonomiques différentes. Une première phylogénie à l'échelle des néogastéropodes, basée sur des mitogénomes, a mis en évidence les difficultés de reconstruction phylogénétique à des échelles taxonomiques profondes basées seulement sur ce type de marqueur, et avec un échantillonnage limité. L'utilisation de la méthode de capture d'exons nous a permis d'obtenir une grande quantité de données afin de reconstruire les différentes phylogénies décrites dans le chapitre 4. Le taux important de capture des exons pour les 3 batchs de séquençage, pour la plupart des spécimens, nous a permis d'intégrer des spécimens de toutes les super-familles et familles (sauf les Pseudolividae) de néogastéropodes mais aussi des super-familles de Caenogastropoda non néogastéropodes proches, avec une efficacité comparable entre le groupe interne et le groupe externe (même si l'analyse plus détaillée présentée dans le chapitre 3 montre une corrélation négative entre la distance génétique et la nombre d'exons reconstruits). J'ai pu mettre ainsi en évidence l'efficacité de la méthode de capture d'exons pour reconstruire des phylogénies profondes.

Concernant l'échantillonnage, avec plus de 1250 genres valides de néogastéropodes (WoRMS), sans compter tous les genres qu'il reste à décrire, contre « seulement » 774 inclus dans la phylogénie, un effort pourra être fait à l'avenir. Certains genres-clés (voir chapitre 4) devront cependant être ciblés en priorité, pour faciliter la conversion de la phylogénie en une nouvelle classification des néogastéropodes. Néanmoins, les résultats obtenus permettent d'ores et déjà d'envisager la description de nouvelles familles (Figure 28) et l'ajout de ces nouveaux spécimens devrait également entraîner la découverte de nouvelles lignées, aux rangs génériques et familiaux.

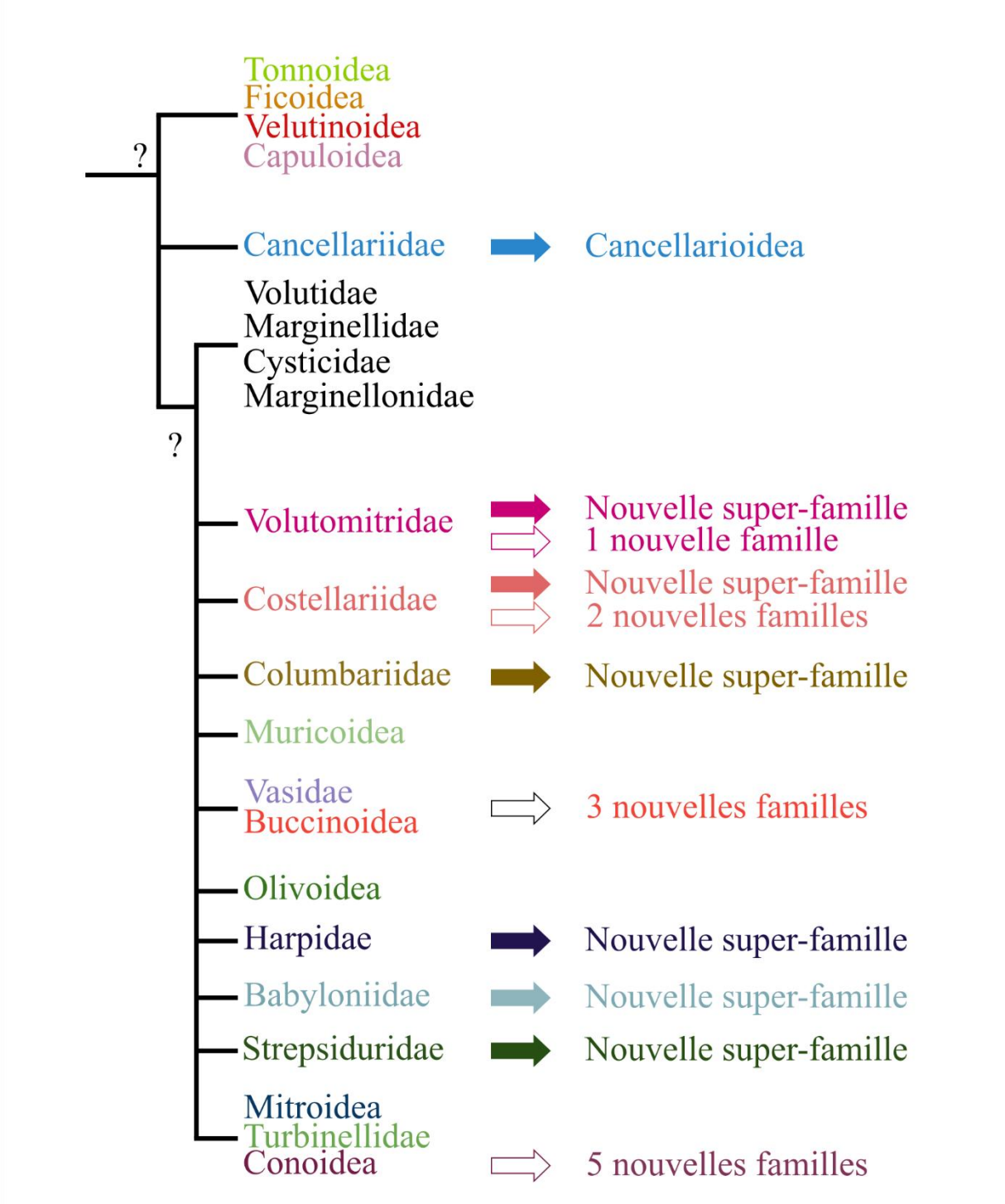


Figure 28 : Résumé des hypothèses phylogénétiques.

Hypothèses les plus robustes pour les néogastéropodes, avec les deux irrésolutions majeures (« ? ») et les potentielles nouvelles familles et super-familles.

Dans la perspective de compléter l'échantillonnage, le travail sur les spécimens anciens de collection est très prometteur. En effet, les premiers tests d'extraction d'ADN à partir des coquilles de mollusque que j'ai effectué au cours de cette thèse doivent être poursuivis. Cela nous donnerait accès à des genres, voire des familles, qui ne seront potentiellement plus jamais collectés, ou qui n'ont jamais été collectés vivants. Nous pourrions alors utiliser les collections plus anciennes du MNHN, qui sont les plus grandes collections du monde en termes de quantité de spécimens de mollusques, et en particulier de néogastéropodes. Les spécimen-types présents dans les collections nous permettraient d'attribuer des noms à des lignées, même pour les espèces délimitées moléculairement. De même, il est à noter que de nombreux spécimens qui n'avaient pas été fixés spécifiquement pour des analyses ADN ont fourni des résultats satisfaisants, permettant de les intégrer aux phylogénies.

Il serait important de vérifier les topologies différentes que nous avons obtenues dans les phylogénies des néogastéropodes, pour les jeux de données NT123, NT12, et AA. Est-ce que ce sont les modèles de reconstruction phylogénétique que nous avons utilisés qui ont créé un biais et qui nous ont donné ces résultats ? Ou est-ce qu'il existe un réel effet du retrait de la 3ème base de chaque codon ? D'après les topologies obtenues, la topologie des arbres reconstruits à partir des séquences d'acides aminés semblent être la plus congruente avec les phylogénies publiées par le passé, et surtout avec les connaissances disponibles sur la morphologie et l'anatomie des néogastéropodes. En effet, l'inclusion des spécimens de Cancellariidae au sein des néogastéropodes, plutôt que leur positionnement en groupe-frère du complexe Tonnoidea, Ficoidea, Velutinoidea et Capuloidea, semble le plus parcimonieux au regard des données morpho-anatomiques : elle suggère que les spécimens de cette famille partagent les mêmes traits morphologiques que les autres néogastéropodes, tels que la présence des glandes salivaires accessoires, la valve de Leiblein ainsi que la glande rectale (Andrews, 1991; Y. I. Kantor & Fedosov, 2009).

Il est important de noter que les hypothèses phylogénétiques des néogastéropodes obtenues au cours de ma thèse sont particulièrement congruentes avec la phylogénie obtenue par Fedosov et al. ((A. E. Fedosov et al., in press), Annexe 1). L'échantillonnage, quoique bien plus réduit dans l'article de Fedosov et al, est en partie recouvrant avec celui que j'ai utilisé, mais les exons séquencés ne sont pas du tout les mêmes. Cela nous a permis d'obtenir des jeux de données indépendants mais congruents, avec des résultats nouveaux, notamment pour le placement de certains genres.

Pour la phylogénie des Raphitomidae, les résultats sont plus compliqués à analyser. Nous avons encore beaucoup de problèmes d'identification, mais aussi de nombreux genres à réviser et à décrire. La comparaison avec les phylogénies précédentes n'est pas simple, car elles étaient bien moins complètes : nous avons en effet produit la première phylogénie à grande échelle pour cette famille, non centrée sur un sous-groupe.

Dans les deux cas (Neogastropoda et Raphitomidae), les phylogénies obtenues vont nécessiter de nombreux échanges avec les spécialistes des taxons, d'abord pour confirmer et affiner les identifications taxonomiques pour les Raphitomidae, et ensuite pour proposer des nouvelles classifications en accord avec les phylogénies. Le jeu de données néogastéropodes, qui a été complété par plusieurs collègues afin de renforcer l'échantillonnage (avec plusieurs représentants par genre (voir chapitre 2), fera certainement l'objet de plusieurs publications. En plus d'une publication pour l'ensemble des néogastéropodes, avec une révision aux rangs super-famille et famille, la qualité de l'échantillonnage et des résultats obtenus permettent d'envisager des révisions aux rangs sous-famille et genre. Par exemple, pour les Muricidae, l'échantillonnage utilisé dans ma thèse est bien plus complet que ceux publiés précédemment (Russini et al., 2023), et fera l'objet d'une publication séparée, avec une révision complète de la classification intra-familiale. La stratégie sera similaire pour les Costellariidae, Colubrariidae, Velutinoidea, Ovulidae, en particulier.

2. BILAN ET PERSPECTIVES METHODOLOGIQUES

Au cours des réflexions autour du design des sondes de capture, nous avons fait le choix de repartir de zéro, et de ne pas utiliser les jeux d'exons définis préalablement dans l'équipe. Ce choix s'explique par la publication d'un génome de référence de néogastéropodes (Pardos-Blas et al., 2021), juste avant le début de ma thèse, mais aussi du séquençage de plusieurs transcriptomes issus de spécimens de différentes super-familles de néogastéropodes produits par l'équipe et nouvellement disponibles. Réduire la quantité de données manquantes était également un objectif, afin de faciliter la reconstruction du jeu de données des exons capturés après le séquençage. Une quantité importante de données manquantes a impacté les reconstructions phylogénétiques passées qui ont été déjà réalisées dans l'équipe (Abdelkrim et al., 2018; A. E. Fedosov et al., in press; Zaharias et al., in press), Annexe 1). Cela a engendré le retrait d'une proportion importante d'exons du jeu de données final, et d'un nombre parfois

important de spécimens. Afin de reconstruire des phylogénies les plus complètes et résolues possibles, le retrait de certains spécimens de groupes taxonomiques rares par exemple, était un problème que nous voulions minimiser. Tout le travail en amont du séquençage (voir chapitre 2) nous a permis d'obtenir de très bons résultats pour la quantité ainsi que la qualité des exons que nous avons obtenus. De plus, c'est au cours du design des exons que j'ai pu me former à l'écriture de scripts Python afin de pouvoir travailler sur les différentes étapes du pipeline de pré-séquençage. Cela m'a permis aussi de poursuivre ce travail lors de l'analyse des données et l'amélioration du pipeline post-séquençage (Chapitre 2).

L'utilisation d'un nouveau pipeline de design des sondes de capture, et l'augmentation de l'efficacité de la capture, nous a permis d'envisager de réduire le nombre d'exons ciblés (de quelques milliers dans la majorité des jeux de données publiés, à un peu plus de 1000 dans notre cas). Les résultats ont montré que cela n'a pas impacté notre capacité à reconstruire des phylogénies supportées, tout en augmentant le nombre de sondes utilisées par exons, et donc réduisant ainsi la quantité de données manquantes. Nous pouvons envisager à l'avenir de réaliser des phylogénies à large échelle à un prix relativement raisonnable pour une qualité et une quantité de données difficilement atteignable avec une autre méthode de séquençage. Dans le cadre de ma thèse, nous avons estimé qu'un échantillon, pour la préparation de banque, la capture et le séquençage, revient à 60-70€. C'est un tarif très compétitif étant donné la quantité et la qualité des données produites (Zaharias et al., 2020). Les autres approches, qui fourniraient plus de données, sont soit significativement plus onéreuses (génomes complets, ou même « genome skimming »), soit nécessitent du matériel frais avec un ARN préservé (transcriptomes).

Au-delà de la forte proportion de données manquantes, qui est un problème majeur lors des projets passés de l'équipe, mais également classique dans les articles publiés, avec un grand nombre de *loci* ciblés écartés du jeu de données final car contenant trop de données manquantes, nous avons aussi essayé de limiter la présence de paralogues. Lors du design des sondes de capture, nous avons éliminé les séquences codantes dans les transcriptomes de néogastéropodes qui pouvaient avoir un match en BLAST avec des protéines différentes du génome de référence (voir figure 6 du chapitre 2). Nous avons ainsi été très strict dans le choix des exons (quitte à limiter le nombre total d'exons, comme discuté ci-dessus), pour minimiser le risque d'inclure des marqueurs de familles multigéniques.

Lors de la reconstruction des exons en post-séquençage, nous n'avons réalisé aucun traitement spécifique afin de gérer la paralogie. Cependant, j'ai pu examiner plusieurs centaines d'alignements dans notre jeu de données final ainsi que reconstruire des arbres de gènes correspondants et je n'ai pas pu mettre en évidence des patterns qui pourraient suggérer la présence de paralogues.

En effet, le travail approfondi réalisé pour retrouver les bons cadres de lecture nous a permis d'obtenir des alignements avec peu d'indels, et il n'a pas été détecté dans les arbres des clades soutenus qui contrediraient fortement la topologie attendue (ou au moins les nœuds connus) pour les néogastéropodes. Néanmoins, cela ne prouve pas l'absence de paralogues dans notre jeu de données, au moins pour certains *loci*, et l'utilisation d'outils de détection de paralogues (comme Orthofinder (Emms & Kelly, 2019)) est à envisager pour confirmer nos hypothèses. De plus, depuis le début de ma thèse et le design des sondes, plusieurs génomes complets de néogastéropodes ont été produits (Farhat et al., 2023), et nous pourrions désormais utiliser ces génomes pour mettre en évidence la présence éventuelle de paralogues.

Enfin, l'intégration d'un transcriptome de la famille des Ovulidae, qui est une famille de Caenogastropoda non néogastéropodes, nous permet d'envisager l'intégration de nouveaux transcriptomes dans le dessin de nouvelles sondes de capture. Nous avons pu constater que les résultats (nombre d'exons capturés) pour cette famille sont aussi bons que pour les groupes de néogastéropodes pour lesquels nous avions plus de transcriptomes disponibles. Le jeu d'exons que nous avons dessiné au cours de cette thèse pourra donc être utilisé pour produire de nouvelles phylogénies à des échelles taxonomiques profondes. En effet, l'ajout de nouveaux transcriptomes dans les nouveaux groupes que nous souhaiterions cibler peut permettre d'avoir une capture d'exons suffisamment efficace pour ajouter des spécimens à la phylogénie globale, taxonomiquement plus étendue que les néogastéropodes. Les exons ciblés à l'aide des sondes de capture sont les mêmes entre tous les groupes. C'est un point important à noter : pour capturer ces mêmes exons, qui devraient être conservés entre des lignées éloignées au sein des Caenogastropoda, voire des Gastéropodes, il suffit de produire de nouvelles sondes, à l'aide de nouveaux transcriptomes, sans avoir besoin de cibler de nouveaux exons.

L'objectif principal de la thèse, à savoir la production d'une phylogénie la plus complète et résolue possible des néogastéropodes, est donc atteint, même si, comme discuté précédemment, cette phylogénie n'est certainement pas la dernière qui sera produite pour ce taxon. Cette phylogénie va donc permettre la poursuite du travail dans le cadre du projet HYPERDIVERSE.

Des analyses macroévolutives, avec l'ajout des données *cox1* pour intégrer la diversité spécifique à la phylogénie, se feront sur la base de la topologie obtenue pendant ma thèse. Dans ce contexte, des analyses de datation de l'arbre seront envisagées, avec l'appui des paléontologues spécialistes des néogastéropodes, qui montreront que les relations phylogénétiques reconstruites sont en accord ou non avec le registre fossile abondant des néogastéropodes.

Bibliographie

Abalde, S., Tenorio, M. J., Afonso, C. M. L., & Zardoya, R. (2017). Mitogenomic phylogeny of cone snails endemic to Senegal. *Molecular Phylogenetics and Evolution*, *112*, 79–87. https://doi.org/10.1016/j.ympev.2017.04.020

Abdelkrim, J., Aznar-Cormano, L., Fedosov, A. E., Kantor, Y. I., Lozouet, P., Phuong, M. A., Zaharias, P., & Puillandre, N. (2018). Exon-Capture-Based Phylogeny and Diversification of the Venomous Gastropods (Neogastropoda, Conoidea). *Molecular Biology and Evolution*, *35*(10), 2355–2374. hal-02002406v1. https://doi.org/10.1093/molbev/msy144

Andrews, E. B. (1991). THE FINE STRUCTURE AND FUNCTION OF THE SALIVARY GLANDS OF *NUCELLA LAPILLUS* (GASTROPODA: MURICIDAE). *Journal of Molluscan Studies*, *57*(1), 111–126. https://doi.org/10.1093/mollus/57.1.111

Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., & Pupko, T. (2012). FastML: A web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research*, *40*(W1), W580–W584. https://doi.org/10.1093/nar/gks498

Bandel, K. (1984). *THE RADULAE OF CARIBBEAN AND OTHER MESOGASTROPODA AND NEOGASTROPODA*. 199.

Bandyopadhyay, P. K., Stevenson, B. J., Cady, M. T., Olivera, B. M., & Wolstenholme, D. R. (2006). Complete mitochondrial DNA sequence of a Conoidean gastropod, Lophiotoma (Xenuroturris) cerithiformis: Gene order and gastropod phylogeny. *Toxicon*, *48*(1), 29–43. https://doi.org/10.1016/j.toxicon.2006.04.013

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Barco, A., Claremont, M., Reid, D. G., Houart, R., Bouchet, P., Williams, S. T., Cruaud, C., Couloux, A., & Oliverio, M. (2010). A molecular phylogenetic framework for the Muricidae, a diverse family of carnivorous gastropods. *Molecular Phylogenetics and Evolution*, *56*(3), 1025–1039. 152. https://doi.org/10.1016/j.ympev.2010.03.008

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., & Good, J. M. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, *13*(1), 403. https://doi.org/10.1186/1471-2164-13-403

Blaimer, B. B., Brady, S. G., Schultz, T. R., Lloyd, M. W., Fisher, B. L., & Ward, P. S. (2015). Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: A case study of formicine ants. *BMC Evolutionary Biology*, *15*(1), 271. https://doi.org/10.1186/s12862-015-0552-5

Blaimer, B. B., Lloyd, M. W., Guillory, W. X., & Brady, S. G. (2016). Sequence Capture and Phylogenetic Utility of Genomic Ultraconserved Elements Obtained from Pinned Insect Specimens. *PLOS ONE*, *11*(8), e0161531. https://doi.org/10.1371/journal.pone.0161531

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bondarev I. (2001). Description of a new cone species (Conus evansi) from the Red Sea, Dahlak (Gastropoda, Conidae). *La Conchiglia*, *33*(299), 25–26.

Borowiec, M. L. (2016). AMAS: A fast tool for alignment manipulation and computing of summary statistics. *PeerJ*, *4*, e1660. https://doi.org/10.7717/peerj.1660

Bose, U., Suwansa-ard, S., Maikaeo, L., Motti, C. A., Hall, M. R., & Cummins, S. F. (2017). Neuropeptides encoded within a neural transcriptome of the giant triton snail Charonia tritonis, a Crown-of-Thorns Starfish predator. *Peptides*, *98*, 3–14. https://doi.org/10.1016/j.peptides.2017.01.004

Bossert, S., Murray, E. A., Almeida, E. A. B., Brady, S. G., Blaimer, B. B., & Danforth, B. N. (2019). Combining transcriptomes and ultraconserved elements to illuminate the phylogeny of Apidae. *Molecular Phylogenetics and Evolution*, *130*, 121–131. https://doi.org/10.1016/j.ympev.2018.10.012

Bouchet, P., Héros, V., Lozouet, P., & Maestrati, P. (2008). A quarter-century of deep-sea malacological exploration in the South and West Pacific: Where do we stand? How far to go? In V. Héros, R. H. Cowie, & P. Bouchet (Eds.), *Tropical Deep-Sea Benthos* (Vol. 25, pp. 9–40). Muséum national d'Histoire naturelle.

Bouchet, P., Kantor, Y. I., Sysoev, A. V., & Puillandre, N. (2011). A new operational classification of the Conoidea (Gastropoda). *Journal of Molluscan Studies*, *77*(3), 273–308. 160. https://doi.org/10.1093/mollus/eyr017

Bouchet, P., Lozouet, P., Maestrati, P., & Heros, V. (2002). Assessing the magnitude of species richness in tropical marine environments: Exceptionally high numbers of molluscs at a New Caledonia site. *Biological Journal of the Linnean Society*, *75*(4), 421–436. 112. https://doi.org/10.1046/j.1095-8312.2002.00052.x

Bouchet, P., Lozouet, P., & Sysoev, A. (2009). An inordinate fondness for turrids. *Deep Sea Research Part II: Topical Studies in Oceanography*, *56*(19–20), 1724–1731. 146. https://doi.org/10.1016/j.dsr2.2009.05.033

Bouchet, P., Rocroi, J.-P., Hausdorf, B., Kaim, A., Kano, Y., Nützel, A., Parkhaev, P., Schrödl, M., & Strong, E. E. (2017). Revised Classification, Nomenclator and Typification of Gastropod and Monoplacophoran Families. *Malacologia*, *61*(1–2), 1–526. hal-03929819v1. https://doi.org/10.4002/040.061.0201

Bouchet, P., & Strong, E. E. (2010). Historical name-bearing types in marine molluscs: An impediment to biodiversity studies. In A. Polaszek (Ed.), *Systema Naturae 250—The Linnaean Ark* (First Edition, pp. 63–74). CRC Press; 150. https://repository.si.edu/bitstream/handle/10088/11295/iz_Bouchet_Strong_2010_Historical_types.pdf?sequence=1&isAllowed=y

Bragg, J. G., Potter, S., Bi, K., & Moritz, C. (2016). Exon capture phylogenomics: Efficacy across scales of divergence. *Molecular Ecology Resources*, *16*(5), 1059–1068. https://doi.org/10.1111/1755-0998.12449

Branstetter, M. G., Longino, J. T., Ward, P. S., & Faircloth, B. C. (2017). Enriching the ant tree of life: Enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. *Methods in Ecology and Evolution*, *8*(6), 768–776. https://doi.org/10.1111/2041-210X.12742

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Cantu, V. A., Sadural, J., & Edwards, R. (2019). *PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets* [Preprint]. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.27553v1

Choi, E. H., Choi, N. R., & Hwang, U. W. (2021). The mitochondrial genome of an Endangered freshwater snail *Koreoleptoxis nodifila* (Caenogastropoda: Semisulcospiridae) from South Korea. *Mitochondrial DNA Part B*, *6*(3), 1120–1123. https://doi.org/10.1080/23802359.2021.1901626

Choquet, M., Smolina, I., Dhanasiri, A. K. S., Blanco-Bercial, L., Kopp, M., Jueterbock, A., Sundaram, A. Y. M., & Hoarau, G. (2019). Towards population genomics in non-model species with large genomes: A case study of the marine zooplankton *Calanus finmarchicus*. *Royal Society Open Science*, *6*(2), 180608. https://doi.org/10.1098/rsos.180608

Claremont, M., Houart, R., Williams, S. T., & Reid, D. G. (2013). A molecular phylogenetic framework for the Ergalataxinae (Neogastropoda: Muricidae). *Journal of Molluscan Studies*, *79*(1), 19–29. https://doi.org/10.1093/mollus/eys028

Claremont, M., Reid, D. G., & Williams, S. T. (2011). Evolution of corallivory in the gastropod genus Drupella. *Coral Reefs*, *30*(4), 977–990. https://doi.org/10.1007/s00338-011-0788-5

Claremont, M., Vermeij, G. J., Williams, S. T., & Reid, D. G. (2013). Global phylogeny and new classification of the Rapaninae (Gastropoda: Muricidae), dominant molluscan predators on tropical rocky seashores. *Molecular Phylogenetics and Evolution*, *66*(1), 91–102. https://doi.org/10.1016/j.ympev.2012.09.014

Colgan, D. J., Ponder, W. F., Beacham, E., & Macaranas, J. (2007). Molecular phylogenetics of Caenogastropoda (Gastropoda: Mollusca). *Molecular Phylogenetics and Evolution*, *42*(3), 717–737. https://doi.org/10.1016/j.ympev.2006.10.009

Couto, D. R., Bouchet, P., Kantor, Y. I., Simone, L. R. L., & Giribet, G. (2016). A multilocus molecular phylogeny of Fasciolariidae (Neogastropoda: Buccinoidea). *Molecular Phylogenetics and Evolution*, *99*, 309–322. https://doi.org/10.1016/j.ympev.2016.03.025

Criscione, F., Hallan, A., Puillandre, N., & Fedosov, A. (2021). Where the snails have no name: A molecular phylogeny of Raphitomidae (Neogastropoda: Conoidea) uncovers vast unexplored diversity in the deep seas of temperate southern and eastern Australia. *Zoological Journal of the Linnean Society*, *191*(4), 961–1000. hal-02970382v1. https://doi.org/10.1093/zoolinnean/zlaa088

Cunha, R. L., Castilho, R., Rüber, L., & Zardoya, R. (2005). Patterns of Cladogenesis in the Venomous Marine Gastropod Genus Conus from the Cape Verde Islands. *Systematic Biology*, *54*(4), 634–650. https://doi.org/10.1080/106351591007471

Cunha, R. L., Grande, C., & Zardoya, R. (2009). Neogastropod phylogenetic relationships based on entire mitochondrial genomes. *BMC Evolutionary Biology*, *9*(1), 210. https://doi.org/10.1186/1471-2148-9-210

Cunha, R. L., Tenorio, M. J., Afonso, C., Castilho, R., & Zardoya, R. (2007). Replaying the tape: Recurring biogeographical patterns in Cape Verde Conus after 12 million years: RECURRING BIOGEOGRAPHICAL PATTERNS IN CAPE VERDE CONUS. *Molecular Ecology*, *17*(3), 885–901. https://doi.org/10.1111/j.1365-294X.2007.03618.x

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., & Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences*, *110*(39), 15758–15763. https://doi.org/10.1073/pnas.1314445110

deMaintenon, M. J. (1999). Phylogenetic Analysis of the Columbellidae (Mollusca: Neogastropoda) and the Evolution of Herbivory from Carnivory. *Invertebrate Biology*, *118*(3), 258. https://doi.org/10.2307/3226997

Der Sarkissian, C., Möller, P., Hofman, C. A., Ilsøe, P., Rick, T. C., Schiøtte, T., Sørensen, M. V., Dalén, L., & Orlando, L. (2020). Unveiling the Ecological Applications of Ancient DNA From Mollusk Shells. *Frontiers in Ecology and Evolution*, *8*, 37. https://doi.org/10.3389/fevo.2020.00037

Der Sarkissian, C., Pichereau, V., Dupont, C., Ilsøe, P. C., Perrigault, M., Butler, P., Chauvaud, L., Eiríksson, J., Scourse, J., Paillard, C., & Orlando, L. (2017). Ancient DNA analysis identifies marine mollusc shells as new metagenomic archives of the past. *Molecular Ecology Resources*, *17*(5), 835–853. https://doi.org/10.1111/1755-0998.12679

Derkarabetian, S., Benavides, L. R., & Giribet, G. (2019). Sequence capture phylogenomics of historical ethanol-preserved museum specimens: Unlocking the rest of the vault. *Molecular Ecology Resources*, *19*(6), 1531–1544. https://doi.org/10.1111/1755-0998.13072

Duda, T. F., & Kohn, A. J. (2005). Species-level phylogeography and evolutionary history of the hyperdiverse marine gastropod genus Conus. *Molecular Phylogenetics and Evolution*, *34*(2), 257–272. https://doi.org/10.1016/j.ympev.2004.09.012

Duda, T. F., Kohn, A. J., & Palumbi, S. R. (2001). Origins of diverse feeding ecologies within Conus, a genus of venomous marine gastropods. *Biological Journal of the Linnean Society*, *73*(4), 391–409. https://doi.org/10.1111/j.1095-8312.2001.tb01369.x

Duda, T. F., & Lee, T. (2009). Ecological Release and Venom Evolution of a Predatory Marine Snail at Easter Island. *PLoS ONE*, *4*(5), e5558. https://doi.org/10.1371/journal.pone.0005558

Duda, T. F., & Palumbi, S. R. (2004). Gene expression and feeding ecology: Evolution of piscivory in the venomous gastropod genus *Conus*. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *271*(1544), 1165–1174. https://doi.org/10.1098/rspb.2004.2708

Duda, T. F., & Rolán, E. (2004). Explosive radiation of Cape Verde Conus, a marine species flock: MARINE SPECIES FLOCK IN CAPE VERDE. *Molecular Ecology*, *14*(1), 267–272. https://doi.org/10.1111/j.1365-294X.2004.02397.x

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Espiritu, D. J. D., Watkins, M., Dia-Monje, V., Cartier, G. E., Cruz, L. J., & Olivera, B. M. (2001). Venomous cone snails: Molecular phylogeny and the generation of toxin diversity. *Toxicon*, *39*(12), 1899–1916. https://doi.org/10.1016/S0041-0101(01)00175-1

Faircloth, B. C., Branstetter, M. G., White, N. D., & Brady, S. G. (2015). Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among H ymenoptera. *Molecular Ecology Resources*, *15*(3), 489–501. https://doi.org/10.1111/1755-0998.12328

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. *Systematic Biology*, *61*(5), 717–726. https://doi.org/10.1093/sysbio/sys004

Farhat, S., Modica, M. V., & Puillandre, N. (2023). Whole Genome Duplication and Gene Evolution in the Hyperdiverse Venomous Gastropods. *Molecular Biology and Evolution*, *40*(8), msad171. https://doi.org/10.1093/molbev/msad171

Fassio, G., Russini, V., Pusateri, F., Giannuzzi-Savelli, R., Høisæter, T., Puillandre, N., Modica, M. V., & Oliverio, M. (2019). An assessment of *Raphitoma* and allied genera (Neogastropoda: Raphitomidae). *Journal of Molluscan Studies*, *85*(4), 413–424. hal-02970454v1. https://doi.org/10.1093/mollus/eyz022

Fedesov, A. E., & Kantor, Y. I. (2008). Toxoglossan gastropods of the subfamily Crassispirinae (Turridae) lacking a radula, and a discussion of the status of the subfamily Zemaciinae. *Journal of Molluscan Studies*, *74*(1), 27–35. https://doi.org/10.1093/mollus/eym042

Fedosov, A. E., Caballer Gutierrez, M., Buge, B., Sorokin, P. V., Puillandre, N., & Bouchet, P. (2019). Mapping the missing branch on the neogastropod tree of life: Molecular phylogeny of marginelliform gastropods. *Journal of Molluscan Studies*, *85*(4), 439–451. hal-02559712v1. https://doi.org/10.1093/mollus/eyz028

Fedosov, A. E., Malcolm, G., Terryn, Y., Gorson, J., Modica, M. V., Holford, M., & Puillandre, N. (2019). Phylogenetic classification of the family Terebridae (Neogastropoda: Conoidea). *Journal of Molluscan Studies*, *85*(4), 359–388. hal-02559725v1. https://doi.org/10.1093/mollus/eyz004

Fedosov, A. E., & Puillandre, N. (2012). Phylogeny and taxonomy of the *Kermia–Pseudodaphnella* (Mollusca: Gastropoda: Raphitomidae) genus complex: a remarkable radiation via diversification of larval development. *Systematics and Biodiversity*, *10*(4), 447–477. https://doi.org/10.1080/14772000.2012.753137

Fedosov, A. E., Zaharias, P., Lemarcis, T., Modica, M. V., Holford, M., Oliverio, M., Kantor, Y. I., & Puillandre, N. (in press). *Phylogenomics of Neogastropoda: The backbone hidden in the bush.*

Fedosov, A., & Kantor, Y. (2007). Toxoglossan gastropods of the subfamily Crassispirinae (Turridae) lacking a radula, and a discussion of the status of the subfamily Zemaciinae. *Journal of Molluscan Studies*, *74*(1), 27–35. https://doi.org/10.1093/mollus/eym042

Fedosov, A., Puillandre, N., Herrmann, M., Kantor, Y., Oliverio, M., Dgebuadze, P., Modica, M. V., & Bouchet, P. (2018). The collapse of Mitra: Molecular systematics and morphology of the Mitridae (Gastropoda: Neogastropoda). *Zoological Journal of the Linnean Society*, *183*(2), 253–337. hal-03926162. https://doi.org/10.1093/zoolinnean/zlx073

Fedosov, A., Puillandre, N., Kantor, Y., & Bouchet, P. (2015). Phylogeny and systematics of mitriform gastropods (Mollusca: Gastropoda: Neogastropoda): Phylogeny of Mitriform Gastropods. *Zoological Journal of the Linnean Society*, *175*(2), 336–359. https://doi.org/10.1111/zoj.12278

Ferreira, S., Ashby, R., Jeunen, G.-J., Rutherford, K., Collins, C., Todd, E. V., & Gemmell, N. J. (2020). DNA from mollusc shell: A valuable and underutilised substrate for genetic analyses. *PeerJ*, *8*, e9420. https://doi.org/10.7717/peerj.9420

Fourdrilis, S., de Frias Martins, A. M., & Backeljau, T. (2018). Relation between mitochondrial DNA hyperdiversity, mutation rate and mitochondrial genome evolution in Melarhaphe neritoides (Gastropoda: Littorinidae) and other Caenogastropoda. *Scientific Reports*, *8*(1), 17964. https://doi.org/10.1038/s41598-018-36428-7

Galindo, L. A., Puillandre, N., Strong, E. E., & Bouchet, P. (2014). Using microwaves to prepare gastropods for DNA barcoding. *Molecular Ecology Resources*, *14*(4), 700–705. 176. https://doi.org/10.1111/1755-0998.12231

Galindo, L. A., Puillandre, N., Utge, J., Lozouet, P., & Bouchet, P. (2016). The phylogeny and systematics of the Nassariidae revisited (Gastropoda, Buccinoidea). *Molecular Phylogenetics and Evolution*, *99*, 337–353. https://doi.org/10.1016/j.ympev.2016.03.019

Gamba, C., Hanghøj, K., Gaunitz, C., Alfarhan, A. H., Alquraishi, S. A., Al-Rasheid, K. A. S., Bradley, D. G., & Orlando, L. (2016). Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. *Molecular Ecology Resources*, *16*(2), 459–469. https://doi.org/10.1111/1755-0998.12470

Gamba, C., Jones, E. R., Teasdale, M. D., McLaughlin, R. L., Gonzalez-Fortes, G., Mattiangeli, V., Domboróczki, L., Kővári, I., Pap, I., Anders, A., Whittle, A., Dani, J., Raczky, P., Higham, T. F. G., Hofreiter, M., Bradley, D. G., & Pinhasi, R. (2014). Genome flux and stasis in a five millennium transect of European prehistory. *Nature Communications*, *5*(1), 5257. https://doi.org/10.1038/ncomms6257

Geist, J., Wunderlich, H., & Kuehn, R. (2008). Use of mollusc shells for DNA-based molecular analyses. *Journal of Molluscan Studies*, *74*(4), 337–343. https://doi.org/10.1093/mollus/eyn025

Goulding, T. C., Yeung, N. W., & Hayes, K. A. (2021). Historical DNA from Museum Shell Collections: Evaluating the Suitability of Dried Micromollusks for Molecular Systematics. *American Malacological Bulletin*, *38*(2). https://doi.org/10.4003/006.038.0209

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Harasewych, M. G., Sei, M., Wirshing, H. H., & Uribe, J. E. (2019). The complete mitochondrial genome of Neptuneopsis gilchristi G.B. Sowerby III, 1898 (Neogastropoda: Volutidae: Calliotectinae). *THE NAUTILUS*, *133*, 7.

Hayashi, S. (2005). The molecular phylogeny of the Buccinidae (Caenogastropoda: Neogastropoda). *MOLLUSCAN RESEARCH*, *25*, 14.

Hedtke, S. M., Morgan, M. J., Cannatella, D. C., & Hillis, D. M. (2013). Targeted Enrichment: Maximizing Orthologous Gene Comparisons across Deep Evolutionary Time. *PLoS ONE*, *8*(7), e67908. https://doi.org/10.1371/journal.pone.0067908

Holford, M., Puillandre, N., Terryn, Y., Cruaud, C., Olivera, B., & Bouchet, P. (2009). Evolution of the Toxoglossa Venom Apparatus as Inferred by Molecular Phylogeny of the Terebridae. *Molecular Biology and Evolution*, *26*(1), 15–25. 143. https://doi.org/10.1093/molbev/msn211

Huang, X., & Madan, A. (1999). CAP3: A DNA Sequence Assembly Program. *Genome Research*, *9*(9), 868–877. https://doi.org/10.1101/gr.9.9.868

Hugall, A. F., O'Hara, T. D., Hunjan, S., Nilsen, R., & Moussalli, A. (2016). An Exon-Capture System for the Entire Class Ophiuroidea. *Molecular Biology and Evolution*, *33*(1), 281–294. https://doi.org/10.1093/molbev/msv216

Hunter, J. P. (1998). Key innovations and the ecology of macroevolution. *Trends in Ecology & Evolution*, *13*(1), 31–36. https://doi.org/10.1016/S0169-5347(97)01273-1

Inäbnit, T., Jochum, A., Slapnik, R., & Neubert, E. (2021). New genetic data reveals a new species of Zospeum in Bosnia (Gastropoda, Ellobioidea, Carychiinae). *ZooKeys*, *1071*, 175–193. https://doi.org/10.3897/zookeys.1071.66417

Jacobs, D. K., & Lindberg, D. R. (1998). Oxygen and evolutionary patterns in the sea: Onshore/offshore trends and recent recruitment of deep-sea faunas. *Proceedings of the National Academy of Sciences*, *95*(16), 9396–9401. https://doi.org/10.1073/pnas.95.16.9396

Jiang, J., Yuan, H., Zheng, X., Wang, Q., Kuang, T., Li, J., Liu, J., Song, S., Wang, W., Cheng, F., Li, H., Huang, J., & Li, C. (2019). Gene markers for exon capture and phylogenomics in ray-finned fishes. *Ecology and Evolution*, *9*(7), 3973–3983. https://doi.org/10.1002/ece3.5026

Kantor, Y., Fedosov, A. E., Puillandre, N., Bonillo, C., & Bouchet, P. (2017). Returning to the roots: Morphology, molecular phylogeny and classification of the Olivoidea (Gastropoda: Neogastropoda). *Zoological Journal of the Linnean Society*, *180*(3), 493–541. hal-03921031v1. https://doi.org/10.1093/zoolinnean/zlw003

Kantor, Y. I. (n.d.). *ANATOMICAL BASIS FOR THE ORIGIN AND EVOLUTION OF THE TOXOGLOSSAN MODE OF FEEDING*. 16.

Kantor, Y. I., & Fedosov, A. (2009). Morphology and development of the valve of Leiblein: Possible evidence for paraphyly of the Neogastropoda. *THE NAUTILUS*, *123*(3), 10.

Kantor, Y. I., Fedosov, A. E., Kosyan, A. R., Puillandre, N., Sorokin, P. A., Kano, Y., Clark, R., & Bouchet, P. (2022). Molecular phylogeny and revised classification of the Buccinoidea (Neogastropoda). *Zoological Journal of the Linnean Society*, *194*(3), 789–857. hal-03321428v1. https://doi.org/10.1093/zoolinnean/zlab031

Kantor, Y. I., Lozouet, P., Puillandre, N., & Bouchet, P. (2014). Lost and found: The Eocene family Pyramimitridae (Neogastropoda) discovered in the Recent fauna of the Indo-Pacific. *Zootaxa*, *3754*(3), 239–276. 175. https://doi.org/10.11646/zootaxa.3754.3.2

Kantor, Y. I., & Puillandre, N. (2021). Rare, deep-water and similar: Revision of Sibogasyrinx (Conoidea: Cochlespiridae). *European Journal of Taxonomy*, *773*, 19–60. hal-03360999v1. https://doi.org/10.5852/ejt.2021.773.1509

Kantor, Y. I., Puillandre, N., Fraussen, K., Fedosov, A., & Bouchet, P. (2013). Deep-water Buccinidae (Gastropoda: Neogastropoda) from sunken wood, vents and seeps: molecular phylogeny and taxonomy. *Journal of the Marine Biological Association of the United Kingdom*, *93*(8), 2177–2195. 173. https://doi.org/10.1017/S0025315413000672

Kantor, Y. I., Puillandre, N., Rivasseau, A., & Bouchet, P. (2012). Neither a buccinid nor a turrid: A new family of deep-sea snails for Belomitra P. Fischer, 1883 (Mollusca, Neogastropoda), with a review of Recent Indo-Pacific species. *Zootaxa*, *3496*(1), 1–64. 167. https://doi.org/10.11646/zootaxa.3496.1.1

Kantor, Y. I., Strong, E. E., & Puillandre, N. (2012). A new lineage of Conoidea (Gastropoda: Neogastropoda) revealed by morphological and molecular data. *Journal of Molluscan Studies*, *78*(3), 246–255. https://doi.org/10.1093/mollus/eys007

Kantor, Y. I., & Taylor, I. D. (n.d.). *Foregut anatomy and relationships of raphitomine gastropods (Gastropoda: Conoidea: Raphitominae)*. 1.

Kantor, Y. I., & Taylor, J. D. (1991). Evolution of the toxoglossan feeding mechanism: New information on the use of the radula. *Journal of Molluscan Studies*, *57*(1), 129–134. https://doi.org/10.1093/mollus/57.1.129

Kantor, Y., Sirenko, B., Zvonareva, S. S., & Fedosov, A. (2022). Taxonomic status of genera of Buccininae (Neogastropoda, Buccinidae) updated based on molecular data with description of new species and corrections of nomenclature of Buccinum. *European Journal of Taxonomy*, *817*, 11–34. hal-03949236v1. https://doi.org/10.5852/ejt.2022.817.1759

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kocot, K. M., Cannon, J. T., Todt, C., Citarella, M. R., Kohn, A. B., Meyer, A., Santos, S. R., Schander, C., Moroz, L. L., Lieb, B., & Halanych, K. M. (2011). Phylogenomics reveals deep molluscan relationships. *Nature*, *477*, 452–456. https://doi.org/10.1038/nature10382

Kraus, N. J., Corneli, P. S., Watkins, M., Bandyopadhyay, P. K., Seger, J., & Olivera, B. M. (2011). Against expectation: A short sequence with high signal elucidates cone snail phylogeny. *Molecular Phylogenetics and Evolution*, *58*(2), 383–389. https://doi.org/10.1016/j.ympev.2010.11.020

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., Lin, D., Fang, H., Zhu, A., & Gao, Y. (2010). Species identification and phylogenetic analysis of genus Nassarius (Nassariidae) based on mitochondrial genes. *Chinese Journal of Oceanology and Limnology*, *28*(3), 565–572. https://doi.org/10.1007/s00343-010-9031-4

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Lorion, J., Duperron, S., Gros, O., Cruaud, C., & Samadi, S. (2009). Several deep-sea mussels and their associated symbionts are able to live both on wood and on whale falls. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1654), 177–185. https://doi.org/10.1098/rspb.2008.1101

Machkour-M'Rabet, S., Hanes, M. M., Martínez-Noguez, J. J., Cruz-Medina, J., & García-De León, F. J. (2021). The queen conch mitogenome: Intra- and interspecific mitogenomic variability in Strombidae and phylogenetic considerations within the Hypsogastropoda. *Scientific Reports*, *11*(1), 11972. https://doi.org/10.1038/s41598-021-91224-0

Magoc, T., & Salzberg, S. L. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, *27*(21), 2957–2963. https://doi.org/10.1093/bioinformatics/btr507

McCormack, J. E., Tsai, W. L. E., & Faircloth, B. C. (2016). Sequence capture of ultraconserved elements from bird museum specimens. *Molecular Ecology Resources*, *16*(5), 1189–1203. https://doi.org/10.1111/1755-0998.12466

Meyer, M., & Kircher, M. (2010). Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harbor Protocols*, *2010*(6), pdb.prot5448. https://doi.org/10.1101/pdb.prot5448

Mirarab, S., Reaz, R., Bayzid, Md. S., Zimmermann, T., Swenson, M. S., & Warnow, T. (2014). ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics*, *30*(17), i541–i548. https://doi.org/10.1093/bioinformatics/btu462

Modica, M. V., Bouchet, P., Cruaud, C., Utge, J., & Oliverio, M. (2011). Molecular phylogeny of the nutmeg shells (Neogastropoda, Cancellariidae). *Molecular Phylogenetics and Evolution*, *59*(3), 685–697. 158. https://doi.org/10.1016/j.ympev.2011.03.022

Modica, M. V., Kosyan, A. R., & Oliverio, M. (2009). The relationships of the enigmatic gastropod Tritonoharpa (Neogastropoda): New data on early neogastropod evolution? *THE NAUTILUS*, *123*(3).

Modica, M. V., Reinoso Sánchez, J., Pasquadibisceglie, A., Oliverio, M., Mariottini, P., & Cervelli, M. (2018). Anti-haemostatic compounds from the vampire snail Cumia reticulata: Molecular cloning and in-silico structure-function analysis. *Computational Biology and Chemistry*, *75*, 168–177. https://doi.org/10.1016/j.compbiolchem.2018.05.014

Modica, M.-V., Verhecken, A., & Oliverio, M. (2011). The relationships of the enigmatic neogastropod Loxotaphrus (Cancellariidae). *New Zealand Journal of Geology and Geophysics*, *54*(1), 115–124. https://doi.org/10.1080/00288306.2011.537610

Moles, J., & Giribet, G. (2021). A polyvalent and universal tool for genomic studies in gastropod molluscs (Heterobranchia). *Molecular Phylogenetics and Evolution*, *155*, 106996. https://doi.org/10.1016/j.ympev.2020.106996

Morton, B. (2003). [No title found]. *Molluscan Research*, *23*(3), 239. https://doi.org/10.1071/MR03008

Morton, B., & Jones, D. S. (2003). THE DIETARY PREFERENCES OF A SUITE OF CARRION-SCAVENGING GASTROPODS(NASSARIIDAE, BUCCINIDAE) IN PRINCESS ROYAL HARBOUR, ALBANY, WESTERNAUSTRALIA. *Journal of Molluscan Studies*, *69*(2), 151–156. https://doi.org/10.1093/mollus/69.2.151

Nam, H. H., Corneli, P. S., Watkins, M., Olivera, B., & Bandyopadhyay, P. (2009). Multiple genes elucidate the evolution of venomous snail-hunting Conus species. *Molecular Phylogenetics and Evolution*, *53*(3), 645–652. https://doi.org/10.1016/j.ympev.2009.07.013

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Olivera, B. M., Fedosov, A., Imperial, J. S., & Kantor, Y. (2017). Physiology of Envenomation by Conoidean Gastropods. In S. Saleuddin & S. Mukai (Eds.), *Physiology of Molluscs* (1st ed., Vol. 1, pp. 153–188). Apple Academic Press; hal-03943153v1. https://doi.org/10.1201/9781315207124-5

Oliverio, M. (2009). Diversity of Coralliophilinae (Mollusca, Neogastropoda, Muricidae) at Austral Islands (South Pacific). *Zoosystema*, *31*(4), 759–789.

Oliverio, M., Cervelli, M., & Mariottini, P. (2002). ITS2 rRNA evolution and its congruence with the phylogeny of muricid neogastropods (Caenogastropoda, Muricoidea). *Molecular Phylogenetics and Evolution*, *25*(1), 63–69. https://doi.org/10.1016/S1055-7903(02)00227-0

Oliverio, M., & Modica, M. V. (2010). Relationships of the haematophagous marine snail *Colubraria* (Rachiglossa: Colubrariidae), within the neogastropod phylogenetic framework. *Zoological Journal of the Linnean Society*, *158*(4), 779–800. https://doi.org/10.1111/j.1096-3642.2009.00568.x

Osca, D., Templado, J., & Zardoya, R. (2015). Caenogastropod mitogenomics. *Molecular Phylogenetics and Evolution*, *93*, 118–128. https://doi.org/10.1016/j.ympev.2015.07.011

Pardos-Blas, J. R., Irisarri, I., Abalde, S., Afonso, C. M. L., Tenorio, M. J., & Zardoya, R. (2021). The genome of the venomous snail *Lautoconus ventricosus* sheds light on the origin of conotoxin diversity. *GigaScience*, *10*(5), giab037. https://doi.org/10.1093/gigascience/giab037

Pereira, C. M., Rosado, J., Seabra, S. G., Pina-Martins, F., Paulo, O. S., & Fonseca, P. J. (2010). *Conus pennaceus*: A phylogenetic analysis of the Mozambican molluscan complex. *African Journal of Marine Science*, *32*(3), 591–599. https://doi.org/10.2989/1814232X.2010.538157

Petit, R. E., & Harasewych, M. (2005). Catalogue of the superfamily Cancellarioidea Forbes and Hanley, 1851 (Gastropoda: Prosobranchia). *Zootaxa*.

Phuong, M. A., & Mahardika, G. N. (2018). Targeted sequencing of venom genes from cone snail genomes improves understanding of conotoxin molecular evolution. *Molecular Biology and Evolution*, *35*(5), 1210–1224.

Phuong, M. A., Mahardika, G. N., & Alfaro, M. E. (2016). Dietary breadth is positively correlated with venom complexity in cone snails. *BMC Genomics*, *17*(1), 401. https://doi.org/10.1186/s12864-016-2755-6

Ponder, W. F., Colgan, D. J., Healy, J. M., Alexander, N., Simone, L. R. L., & Mielke, E. E. (2008). Caenogastropoda. In W. Ponder (Ed.), *Phylogeny and Evolution of the Mollusca* (pp. 331–383). University of California Press. https://doi.org/10.1525/california/9780520250925.003.0013

Ponder, W. F., & Lindberg, D. R. (1997). Towards a phylogeny of gastropod molluscs: An analysis using morphological characters. *Zoological Journal of the Linnean Society*, *119*(2), 83–265. https://doi.org/10.1111/j.1096-3642.1997.tb00137.x

Psonis, N., Vardinoyannis, K., & Poulakakis, N. (2022). High-throughput degraded DNA sequencing of subfossil shells of a critically endangered stenoendemic land snail in the Aegean. *Molecular Phylogenetics and Evolution*, *175*, 107561. https://doi.org/10.1016/j.ympev.2022.107561

Puillandre, N., Bouchet, P., Duda, T. F., Kauferstein, S., Kohn, A. J., Olivera, B. M., Watkins, M., & Meyer, C. (2014). Molecular phylogeny and evolution of the cone snails (Gastropoda, Conoidea). *Molecular Phylogenetics and Evolution*, *78*, 290–303. https://doi.org/10.1016/j.ympev.2014.05.023

Puillandre, N., Duda, T. F., Meyer, C. P., Olivera, B. M., & Bouchet, P. (2015). One, four or 100 genera? A new classification of the cone snails. *Journal of Molluscan Studies*, *81*(1), 1–23. https://doi.org/10.1093/mollus/eyu055

Puillandre, N., Kantor, Y. I., Sysoev, A. V., Couloux, A., Meyer, C. P., Rawlings, T., Todd, J. A., & Bouchet, P. (2011). The dragon tamed? A molecular phylogeny of the Conoidea (Gastropoda). *Journal of Molluscan Studies*, *77*(3), 259–272. 159. https://doi.org/10.1093/mollus/eyr015

Russini, V., Fassio, G., Modica, M. V., deMaintenon, M. J., & Oliverio, M. (2017). An assessment of the genus *Columbella* Lamarck, 1799 (Gastropoda: Columbellidae) from eastern Atlantic. *Zoosystema*, *39*(2), 197–212. https://doi.org/10.5252/z2017n2a2

Russini, V., Fassio, G., Nocella, E., HOUART, R., Barco, A., Puillandre, N., Lozouet, P., Modica, M. V., & Oliverio, M. (2023). Whelks, rock-snails, and allied: A new phylogenetic framework for the family Muricidae

(Mollusca: Gastropoda). *The European Zoological Journal*, *90(2)*, 856–868. https://doi.org/doi.org/10.1080/24750263.2023.2283517

Ryu, T., Seridi, L., & Ravasi, T. (2012). The evolution of ultraconserved elements with different phylogenetic origins. *BMC Evolutionary Biology*, *12*(1), 236. https://doi.org/10.1186/1471-2148-12-236

Shokralla, S., Gibson, J. F., Nikbakht, H., Janzen, D. H., Hallwachs, W., & Hajibabaei, M. (2014). Next-generation DNA barcoding: Using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular Ecology Resources*, *14*(5), 892–901. https://doi.org/10.1111/1755-0998.12236

Silva, P. C., Malabarba, M. C., Vari (In memoriam), R., & Malabarba, L. R. (2019). Comparison and optimization for DNA extraction of archived fish specimens. *MethodsX*, *6*, 1433–1442. https://doi.org/10.1016/j.mex.2019.06.001

Simakov, O., Marletaz, F., Cho, S.-J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo, D.-H., Larsson, T., Lv, J., Arendt, D., Savage, R., Osoegawa, K., De Jong, P., Grimwood, J., Chapman, J. A., Shapiro, H., Aerts, A., Otillar, R. P., Terry, A. Y., … Rokhsar, D. S. (2013). Insights into bilaterian evolution from three spiralian genomes. *Nature*, *493*(7433), 526–531. https://doi.org/10.1038/nature11696

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. https://doi.org/10.1093/bioinformatics/btv351

Smith, S. A., Wilson, N. G., Goetz, F. E., Feehery, C., Andrade, S. C. S., Rouse, G. W., Giribet, G., & Dunn, C. W. (2011). Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*, *480*(7377), 364–367. https://doi.org/10.1038/nature10526

Starrett, J., Derkarabetian, S., Hedin, M., Bryson, R. W., McCormack, J. E., & Faircloth, B. C. (2017). High phylogenetic utility of an ultraconserved element probe set designed for Arachnida. *Molecular Ecology Resources*, *17*(4), 812–823. https://doi.org/10.1111/1755-0998.12621

Strong, E. E. (2003). Refining molluscan characters: Morphology, character coding and a phylogeny of the Caenogastropoda. *Zoological Journal of the Linnean Society*, *137*(4), 447–554. https://doi.org/10.1046/j.1096-3642.2003.00058.x

Taylor, J. D. (1993). Dietary and anatomical specialization of mitrid gastropods (Mitridae) at Rottnest Island, Western Australia. *Proceedings of the Fifth International Marine Biological Workshop: The Marine Flora and Fauna of Rottnest Island, Western Australia*, 583–599.

Teasdale, L. C., Ko, F., Murray, K. D., O'Hara, T., & Moussalli, A. (2016). *Identification and qualification of 500 nuclear, single-copy, orthologous genes for the Eupulmonata (Gastropoda) using transcriptome sequencing and exon capture*. 17.

Tin, M. M.-Y., Economo, E. P., & Mikheyev, A. S. (2014). Sequencing Degraded DNA from Non-Destructively Sampled Museum Specimens for RAD-Tagging and Low-Coverage Shotgun Phylogenetics. *PLoS ONE*, *9*(5), e96793. https://doi.org/10.1371/journal.pone.0096793

Trevisan, B., Alcantara, D. M. C., Machado, D. J., Marques, F. P. L., & Lahr, D. J. G. (2019). Genome skimming is a low-cost and robust strategy to assemble complete mitochondrial genomes from ethanol preserved specimens in biodiversity studies. *PeerJ*, *7*, e7543. https://doi.org/10.7717/peerj.7543

Uribe, J. E., Zardoya, R., & Puillandre, N. (2018). Phylogenetic relationships of the conoidean snails (Gastropoda: Caenogastropoda) based on mitochondrial genomes. *Molecular Phylogenetics and Evolution*, *127*, 898–906. hal-02002442v1. https://doi.org/10.1016/j.ympev.2018.06.037

Vallejo, B. (2005). Inferring the mode of speciation in Indo-West Pacific Conus (Gastropoda: Conidae): Conus speciation in the IWP. *Journal of Biogeography*, *32*(8), 1429–1439. https://doi.org/10.1111/j.1365-2699.2005.01260.x

Van Valen, L. (1965). Morphological Variation and Width of Ecological Niche. *The American Naturalist*, *99*(908), 377–390. https://doi.org/10.1086/282379

Van Valen, L. (1971). ADAPTIVE ZONES AND THE ORDERS OF MAMMALS. *Evolution*, *25*(2), 420–428. https://doi.org/10.1111/j.1558-5646.1971.tb01898.x

Vaux, F., Hills, S. F. K., Marshall, B. A., Trewick, S. A., & Morgan-Richards, M. (2018). Genome statistics and phylogenetic reconstructions for Southern Hemisphere whelks (Gastropoda: Buccinulidae). *Data in Brief*, *16*, 172–181. https://doi.org/10.1016/j.dib.2017.11.021

Verhecken, A. (2007). Revision of the Cancellariidae (Mollusca, Neogastropoda, Cancellarioidea) of the eastern Atlantic (40 degrees N-40 degrees S) and the Mediterranean. *Zoosystema*, *29*(2), 281–364.

Vermeij, G. J. (2024). *Shell-based genus-level reclassification of the Family Vasidae (Mollusca: Neogastropoda)*.

Walton, K., Scarsbrook, L., Mitchell, K. J., Verry, A. J. F., Marshall, B. A., Rawlence, N. J., & Spencer, H. G. (2023). Application of palaeogenetic techniques to historic mollusc shells reveals phylogeographic structure in a New Zealand abalone. *Molecular Ecology Resources*, *23*(1), 118–130. https://doi.org/10.1111/1755-0998.13696

Wang, J.-G., Zhang, D., Jakovlić, I., & Wang, W.-M. (2017). Sequencing of the complete mitochondrial genomes of eight freshwater snail species exposes pervasive paraphyly within the Viviparidae family (Caenogastropoda). *PLOS ONE*, *12*(7), e0181699. https://doi.org/10.1371/journal.pone.0181699

Wang, Q., Liu, H., Yue, C., Xie, X., Li, D., Liang, M., & Li, Q. (2021). Characterization of the complete mitochondrial genome of *Ficus variegata* (Littorinimorpha: Ficidae) and molecular phylogeny of Caenogastropoda. *Mitochondrial DNA Part B*, *6*(3), 1126–1128. https://doi.org/10.1080/23802359.2021.1901628

WoRMS Editorial Board. (2017). *World Register of Marine Species. Available from http://www.marinespecies.org at VLIZ. Accessed yyyy-mm-dd.* [Darwin Core Archive]. VLIZ. https://doi.org/10.14284/170

Yang, D. Y., Eng, B., Waye, J. S., Dudar, J. C., & Saunders, S. R. (1998). Improved DNA extraction from ancient bones using silica-based spin columns. *American Journal of Physical Anthropology*, *105*(4), 539–543. https://doi.org/10.1002/(SICI)1096-8644(199804)105:4<539::AID-AJPA10>3.0.CO;2-1

Yang, M., Dong, D., & Li, X. (2021). The complete mitogenome of *Phymorhynchus* sp. (Neogastropoda, Conoidea, Raphitomidae) provides insights into the deep-sea adaptive evolution of Conoidea. *Ecology and Evolution*, ece3.7582. https://doi.org/10.1002/ece3.7582

Zaharias, P., Kantor, Y. I., Fedosov, A. E., & Puillandre, N. (in press). *Coupling DNA barcodes and exon-capture to resolve the phylogeny of Turridae (Gastropoda, Conoidea)*.

Zaharias, P., Pante, E., Gey, D., Fedosov, A. E., & Puillandre, N. (2020). Data, time and money: Evaluating the best compromise for inferring molecular phylogenies of non-model animal taxa. *Molecular Phylogenetics and Evolution*, *142*, 106660. hal-02458233v1. https://doi.org/10.1016/j.ympev.2019.106660

Zapata, F., Wilson, N. G., Howison, M., Andrade, S. C. S., Jörger, K. M., Schrödl, M., Goetz, F. E., Giribet, G., & Dunn, C. W. (2014a). *Data from: Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda* (Version 1, p. 68327139 bytes) [dataset]. Dryad. https://doi.org/10.5061/DRYAD.5BC98

Zapata, F., Wilson, N. G., Howison, M., Andrade, S. C. S., Jörger, K. M., Schrödl, M., Goetz, F. E., Giribet, G., & Dunn, C. W. (2014b). Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda.

Proceedings of the Royal Society B: Biological Sciences, *281*(1794), 20141739. https://doi.org/10.1098/rspb.2014.1739

Zou, S., Li, Q., & Kong, L. (2011). Additional gene data and increased sampling give new insights into the phylogenetic relationships of Neogastropoda, within the caenogastropod phylogenetic framework. *Molecular Phylogenetics and Evolution*, *61*(2), 425–435. https://doi.org/10.1016/j.ympev.2011.07.014

ANNEXE 1 :

Fedosov A, Zaharias P, **Lemarcis T**, Modica MV, Holford M, Oliverio M, Kantor Y, Bouchet P, Puillandre N. Phylogenomics of the Neogastropoda: the backbone hidden in the bush. Syst. Biol., in press.

Phylogenomics of Neogastropoda: the backbone hidden in the bush

Alexander E. Fedosov*[1,2], Paul Zaharias[2], Thomas Lemarcis[2], Maria Vittoria Modica[2,3], Mandë Holford[4,5,6], Marco Oliverio[2,7], Yuri I. Kantor[2,8], Nicolas Puillandre[2]

[1] Department of Zoology, Swedish Museum of Natural History, Box 50007, 10405, Stockholm, Sweden.

[2] Institut Systématique Evolution Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, Paris, France.

[3] Department of Biology and Evolution of Marine Organisms, Stazione Zoologica Anton Dohrn, Naples, Italy.

[4] Department of Chemistry, Hunter College, Belfer Research Building, City University of New York, New York, USA.

[5] Department of Invertebrate Zoology, the American Museum of Natural History, New York, USA.

[6] PhD programs in Biology, Biochemistry, and Chemistry, The Graduate Center of the City University of New York, New York, USA.

[7] Department of Biology and Biotechnologies "Charles Darwin", Sapienza University of Rome. Zoology, Rome, Italy.

[8] A.N. Severtsov Institute of Ecology and Evolution, Russian Academy of Sciences, Moscow, Russia

* Corresponding author (Alexander Fedosov, Email: fedosovalexander@gmail.com)

Abstract. – The molluscan order Neogastropoda encompasses over 15,000 almost exclusively marine species playing important roles in benthic communities and in the economies of coastal countries. Neogastropoda underwent intensive cladogenesis in early stages of diversification, generating a 'bush' at the base of their evolutionary tree, that has been hard to resolve even with high throughput molecular data. In the present study to resolve the bush, we use a variety of phylogenetic inference methods and a comprehensive exon capture dataset of 1,817 loci (79.6% data occupancy) comprising 112 taxa of 48 out of 60 Neogastropoda families. Our results show consistent topologies and high support in all analyses at (super)family level, supporting monophyly of Muricoidea, Mitroidea, Conoidea, and, with some reservations, Olivoidea and Buccinoidea. Volutoidea and Turbinelloidea as currently circumscribed are clearly paraphyletic. Despite our analyses consistently resolving most backbone nodes, three prove problematic: First, uncertain placement of Cancellariidae, as the sister group to either a Ficoidea-Tonnoidea clade, or to the rest of Neogastropoda, leaves monophyly of Neogastropoda unresolved. Second, relationships are contradictory at the base of the major 'core Neogastropoda' grouping. Third, coalescence-based analyses reject monophyly of the Buccinoidea in relation to Vasidae. We analysed phylogenetic signal of targeted loci in relation to potential biases, and we propose most probable resolutions in the latter two recalcitrant nodes. The uncertain placement of Cancellariidae may be explained by orthology violations due to differential paralog loss shortly after the whole genome duplication, which should be resolved with a curated set of longer loci.

Keywords: Phylogenomics, Phylogenetic conflict, Mollusca, Targeted enrichment, Cancellariidae, marine mollusks.

Running title: Backbone phylogeny of the Neogastropoda.

Understanding patterns of lineage relatedness is a fundamental task of life science and is the ultimate goal of the Tree of Life (TOL) initiative (Hinchliff et al., 2015). While the introduction of high-throughput sequencing technologies was initially believed to render TOL reconstruction a rather technical task depending mainly on adequate lineage sampling, it has become evident that the process is severely challenged by a phenomenon figuratively named 'bushes in the tree of life' (Rokas and Carroll, 2006). This pattern typically occurs in lineages that have undergone multiple cladogenesis events in a short time span (Rokas and Carroll, 2006). Because the amount of phylogenetic signal is proportional to the TOL stem lengths, short stems require increasingly large amount of data to be resolved (Lanyon, 1988), and the inference of true topology in these segments of a tree is increasingly confounded by homoplasy (Takezaki et al., 2004). Nevertheless, quickly radiating lineages are among the most interesting to investigate, because intensive cladogenesis is a signature of evolutionary success of the lineages (Hunter, 1998), and understanding the origin of their prosperity requires a robust phylogenetic hypothesis (Whitfield and Lockhart, 2007; Prum et al., 2015).

Being the second most species-rich phylum, Mollusca encompasses taxa with remarkable diversity of body plans (Wanninger and Wollesen, 2019; Modica et. al, 2019; Kocot et al., 2020; Ponder et al. 2021) and unresolved or contentious relationships (Cunha et al., 2022; Uribe et al., 2022). Molluscan phylogenetics is challenged by the coexistence of uncertainties regarding the placement of ancient lineages, many of them being extinct (Sutton et al., 2016; Wanninger and Wollesen, 2019), and a plethora of relatively recent successful radiations. The largest marine gastropod order, the Neogastropoda, is perhaps the most conspicuous example of the latter situation. Having radiated in late Cretaceous and early Cenozoic, in the context of the Mesozoic Marine Revolution (Vermeij, 1977), the Neogastropoda flourished in Cenozoic seas. Currently the Neogastropoda exhibit a tremendous species richness with over 15,000 species, corresponding to about one fifth of the present-day molluscan diversity (MolluscaBase, available at https://www.molluscabase.org/). The vast majority of neogastropod species are carnivores. Being slow in motion, many lineages have developed a unique array of biochemical adaptations to mediate interactions with their prey and predators (Olivera et al., 2014; Ponte and Modica, 2017; Kuznetsova et al., 2022). Deadly venoms of cone-snails, comprising a high number of structurally and pharmacologically diversified neuropeptides referred to as conotoxins, are the best-known example of these biochemical adaptations. The unique pharmacological properties of conotoxins and their relevance for drug development (Safavi-Hemami et al., 2019) fuel the increasing multidisciplinary interest in Neogastropoda. However, the lack of a robust phylogenetic hypothesis of Neogastropoda is an impediment to the

systematic investigation of the translational applications of their bioactive compounds. Therefore, reconstructing the phylogeny of the neogastropod order will not only enable a reassessment of neogastropod systematics, but also streamline evolutionary and biochemical research on this successful molluscan lineage.

The 60 currently recognized Neogastropoda families are classified into seven superfamilies (Bouchet et al., 2017, with updates as per MolluscaBase); however, monophyly of the order remains questionable, and interrelationships among its main taxa poorly understood. The published studies addressing Neogastropoda phylogenetics suffered complementary flaws. Morphology-based cladistic analyses (e.g. Riedel, 2000; Simone, 2011) were misled by the wide-spread homoplasies in character evolution. Molecular phylogenies based on the Sanger approach (e.g. Zou et al., 2011; Fedosov et al., 2019) lacked resolution at deep nodes, due to the clearly insufficient number of characters included. In turn, phylogenomic studies (Osca et al., 2015; Abdelkrim et al., 2018; Cunha and Giribet, 2019; Lemarcis et al., 2022) had incomplete and unbalanced taxon sampling and/or suffered from the limitations inherent to mitogenome-based phylogenomics (Duchêne et al., 2011). The goal of this study is to resolve backbone Neogastropoda relationships through extensive lineage sampling and the application of a leading-edge phylogenomic approach to data generation and analysis. We successfully reconstructed a largely supported phylogenetic framework for the Neogastropoda, establishing for the first time affinities of previously enigmatic lineages. While our results suggest major revisions in the systematics of the Neogastropoda, their formal implementation extends beyond the scope of the present work.

Materials and Methods

Bait design and taxa sampling

Details of the probe kit design, taxonomic sampling, lab work and a comprehensive account on the initial stages of the data analysis are provided in the Supplementary Material (10.5061/dryad.8931zcrx5). Briefly, 46 transcriptomes of 32 caenogastropod species were used for bait design (Zaharias et al., 2020; Lemarcis et al., 2022). All transcriptomes were (re)-assembled as detailed in Fassio et al. (2019) and then aligned against the genome of *Lottia gigantea* to identify exon/intron boundaries (Abdelkrim et al., 2018). Then we identified a subset of 4,456 exons (>180-bp) spanning approximately 1.3 Mb that were present in at least two families of Conoidea, and in at least three non-conoidean transcriptomes. The empirical exon sequences (i.e., those present in analyzed transcriptomes) were used alongside reconstructed ancestral sequences (in fast-evolving loci) for probe design, producing a set of 42,011 2x tiling 100-bp baits. After duplicate removal, the final set comprised 40,040 baits developed into a MyBait generation-5 biotinylated probes kit (Mycroarray, Arbor Biosciences, CA).

We obtained Ethanol-preserved tissue samples of 135 taxa, covering 51 families of Neogastropoda and related lineages (Fig. 1, Supplementary Table S1), and complemented these data with 12 transcriptomes that had the highest BUSCO completeness (Waterhouse et al., 2018). Library preparation was performed in three batches: the protocol detailed in Abdelkrim et al. (2018) was used for the specimens in the 1st and 2nd batches, while the KAPA protocol was used for the 3rd batch specimens. The libraries were paired-end sequenced on Illumina HiSeq 4000 and Illumina NovaSeq platforms, with read length of 100 and 150 bp, respectively. The final number of reads per library ranged from 851,299 (MNHN IM-2013-43718, *Glabella rosadoi*) to 44,946,221 (MNHN IM-2013-48309, *Xenophora* sp.), with a median number of 11,4517,88 reads per library.

Data assembly and loci recovery

Data assembly and processing generally followed Abdelkrim et al. (2018). To maximize recovery of targeted loci for exon capture datasets we used two assemblers, SPAdes v3.14 (Bankevich et al., 2012) and TRINITY v2.9 (Grabherr et al., 2011), whereas only TRINITY was used for the transcriptomes. For each exon-capture sample, SPAdes and TRINITY assemblies were merged and clustered by running CD-HIT (Fu et al., 2012) with 99% identity.

We associated assembled contigs with targets using BLASTn, (e-value 1e-20) and used Exonerate v2.2.0 under the est2genome model to redefine boundaries of the targeted exons

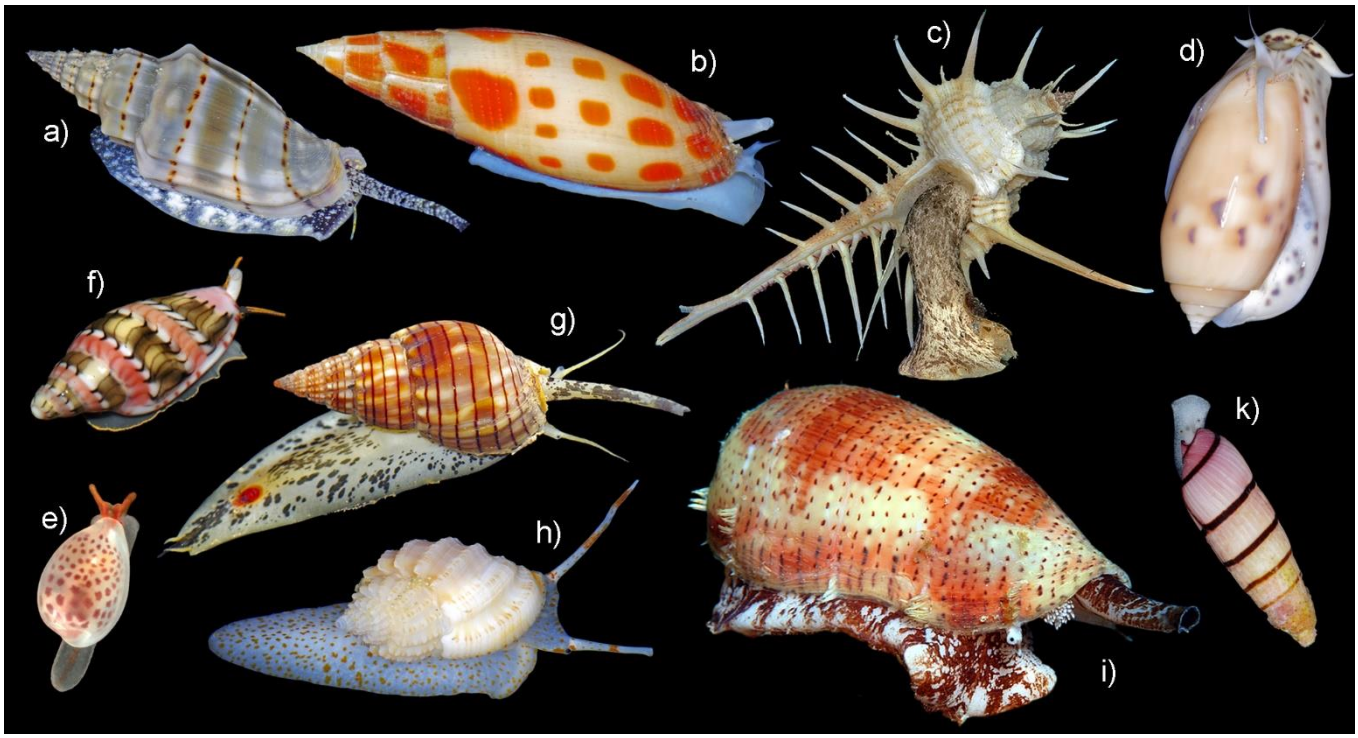


Figure 1. Living members of major Neogastropoda lineages. a. *Vexillum gruneri* (Costellariidae); b. *Mitra mitra* (Mitridae); c. *Murex tenuirostrum* (Muricidae), d. *Oliva amethystina* (Olividae); e. *Ticofurcilla* sp. (Cystiscidae), f. *Marginella festiva* (Marginellidae); g. *Nassarius glans* (Nassariidae); h. *Scalptia contabulata* (Cancellariidae), i. *Conus tulipa* (Conidae), k. *Myurella pygmaea* (Terebridae).

(Abdelkrim et al., 2018). For each sample, all contigs that generated BLAST hits against the exon library were extracted from the assembly. The quality-trimmed reads were mapped against the contigs of interest with bowtie v2.2.7 (Langmead and Salzberg, 2012) to assess capture efficiency. We used samtools v1.9 and bcftools v1.3 (Li et al., 2009) for single-nucleotide polymorphism calling, aiming to assess heterozygosity in the captured sequences. Sites with coverage <4 were masked as 'N', followed by removal of short sequences (length <= 70% of target length), low quality sequences ('N' comprising >30% of sequence length). Sequences with heterozygosity > 2 standard deviations from the mean were also removed.

Orthology assessment

The sequences of interest were sorted by target identity, and then aligned using MAFFT v7.407 (Katoh and Standley 2013) with *G-INS-i*, and *–adjust_direction* option enabled. The alignments were then translated using MACSE v2.06 (Ranwez et al., 2018), and the obtained amino-acid sequences sorted back by sample for orthogroup identification with ORTHOFINDER v2.5.4 (Emms and Kelly, 2019). The thirty most complex orthogroups comprising multiple sequences for nearly all samples were removed. Gene trees were reconstructed for the remaining 3,000 orthogroups with >= 65 samples represented, using RAxML under the GTRGAMMA model with 100 bootstrap replicates (Stamatakis, 2006). When a sample was represented by multiple sequences in an orthogroup alignment, we first used a custom Python script (S10-3) to remove residual cross-contamination based on the orthogroup tree topology, coverage data, and sequences lengths, and then selected the largest 1:1 ortholog subtree using PhyloPyPruner v1.2.6 (Thalén, 2018). We removed terminal long branches using the custom Python script S10-4, and end-trimmed the alignments using TRIMAL v1.2 (Capella-Gutiérrez et al., 2009). The 112 taxa with highest data occupancy (i.e. the smallest amount of missing data (*highDO* taxa) were retained for downstream analyses.

Matrix assembly and phylogenetic analyses

The 1,817 orthogroup alignments comprising >=35 aa sites with >=70 highDO taxa included generated the matrix NEO70 (total 125,508 sites, 20.4% missing data). To further reduce missing data, a subset of 731 alignments comprising >= 95 highDO taxa were selected to build the matrix NEO95 (total 52,805 sites, 11.4% missing data). We used RAxML with a PROTGAMMALG4X model and 20 rapid bootstraps (Cunha et al., 2022) for a second-round gene tree reconstruction, and further subsampled the matrix NEO95 using GenesortR (Mongiardino Koch, 2021). GenesortR first scores all loci based on seven parameters reflecting

phylogenetic ‘usefulness’, and then removes loci that could bias phylogenetic reconstructions. The obtained matrix NEO95-GSR500 consisted of the 500 ‘best’ loci and included 37,958 aligned amino-acid sites.

Multispecies coalescent phylogenies were reconstructed from three respective sets of gene trees by using both ASTRAL III (v5.6.3 - Zhang et al., 2018) and ASTEROID v1.0 (Morel et al., 2023). Maximum Likelihood phylogenies were reconstructed with IQ-TREE v.2.2.1 (Minh et al. 2020), performed on both gene-partitioned (IQ-part) and on unpartitioned matrices with best-fit profile mixture models (IQ-PMM). In partitioned analyses, best fit models were estimated for edge-unlinked partitions, and the partitions with compatible model parameters merged prior to the tree search (**-st AA -msub nuclear -ninit 10 -bb 1500 -sp partition_file -m MFP+MERGE -rcluster 10 -madd LG4M, LG4X -mrate G, R, E**). Due to the prohibitive runtimes of the MFP-MERGE mode, partition merging was not performed on the NEO70 dataset. In the IQ-PMM analyses, the command line of Cunha et al. (2022) was run to identify best fit exchange matrix (**-st AA - msub nuclear -ninit 10 -bb 1500 -m MFP -mset LG, WAG -rcluster 10 -mfreq F+C40/60 -mrate G, R**). Sixty mixture classes (C60) were enabled for NEO95 matrices. In contrast, we only allowed forty mixture classes (C40) for NEO70 due to 1 TB RAM limitation of our phylogenetic server.

We performed Bayesian inference by running PhyloBayes v4.1 (Lartillot et al., 2013) on the two NEO95 matrices, under CAT-GTR model, disregarding constant sites. Each analysis was run in four chains, and terminated once convergence criteria (accessed with tracecomp) were achieved for at least two chains (8,771 and 11,760 generations for NEO95 and NEO95-GSR500 matrices respectively).

To visualize overall similarities among the obtained tree topologies, we first ran a custom python script S10-6 to retrieve all unique clades comprising two or more taxa from the trees from the analyses described above (Fig. 2a), and then compiled a clade presence-absence (coded as 1 and 0) matrix for these 14 trees. This matrix was subjected to Principal component analysis (PCA) using PAST (Hammer et al., 2001).

Our phylogenetic analyses repeatedly recovered alternative topologies at three backbone nodes. The nodes that produced conflicting topologies define (i) the placement of Cancellariidae (referred to as baseNEO), (ii) the first offshoot of Core Neogastropoda (baseCore) (iii) the affinities of early branching Buccinoidea (baseBuc) – Figure 2a-e. To understand the source of support for these conflicting hypotheses, for each contradictory relationship we performed site-wise phylogenetic signal measures (ΔSLS) as detailed by Shen et al. (2017). First, six analyses under constrained topologies were run on the matrix NEO95

(two for each node, one under ML-PMM, another with partitions) to obtain best-scoring alternative topologies. Then for each pair of alternative trees, SLS (per site likelihood score) was calculated under respective model (PROTGAMMALG4X was run as unpartitioned model), using raxmlHPC with –f G option.

We calculated site-wise phylogenetic signal (ΔSLS) by subtracting an alternative topology' SLS from the main topology' SLS (therefore, positive ΔSLS values are those supporting the main topology). We employed Approximately Unbiased (AU) test in CONSEL (Shimodaira and Hasegawa, 2001) to check if one topology is significantly better than the alternative. By summing up ΔSLS values for each locus, we computed ΔGLS values as a proxy of gene-wise phylogenetic signal (custom Python script S10-7). In addition to ΔGLS, we calculated standard deviation for ΔSLS values of each locus, and we used proportion of SD(ΔSLS) to ΔGLS as a measure of noise in the phylogenetic signal. If this proportion exceeded 10 for a locus (suggesting highly dissimilar site-wise signals, summing up to close-to-zero ΔGLS), this locus was excluded as bearing contradictory signal. The remaining loci were divided in three subsets: 10% loci with lowest ΔGLS, 10% loci with highest ΔGLS, and the remaining 80%. Then we performed a t-test to find out whether there was a significant difference among the subsets in respect to potential biases (Saturation, Compositional heterogeneity, Evolutionary rate), assessed by GenesortR.

Results

Support for Neogastropoda superfamilies and families

The composition and relationships within the superfamily level clades are highly congruent among the 17 trees reconstructed from the three analyzed datasets (Supplementary figs. S2-S15). All our analyses support the monophyly of Muricoidea (=Muricidae), Mitroidea, and Conoidea. The remaining four superfamilies are consistently recovered as paraphyletic. Volutoidea comprises two unrelated clusters: Cancellariidae and Volutidae plus marginelliform gastropods (Cystiscidae and Marginellidae). The Panamanian species *Triumphis distorta* traditionally placed in Pseudolividae but unequivocally recovered within Buccinoidea, violates reciprocal monophyly of Olivoidea and Buccinoidea. Whereas Olivoidea excluding *Triumphius* is consistently monophyletic, relationships at the base of Buccinoidea are contradictory (see below). The Turbinelloidea taxa form five unrelated highly supported clades: (i) Volutomitridae plus *Exilioidea* (Ptychatractidae), (ii) Costellariidae plus *Exilia* (Ptychatractidae), (iii) Columbariidae, (iv) Vasidae, and (v) Turbinellidae. The extant families Harpidae and Babyloniidae, currently not assigned to superfamilies (Mollusca base accessed on 17th of July

2023), represented by respectively three and one species in our dataset, do not show consistent affinities with any other lineage. Of the 25 tonnoidean and neogastropod families represented by two or more species in our dataset, monophyly is consistently rejected for Pseudolividae (see above) and Ptychatractidae (with *Exilioidea* always being sister group to Volutomitridae, and *Exilia* the sister group to Costellariidae). Furthermore, in five analyses, Volutidae is retrieved as paraphyletic in relation to the marginelliform clade (Fig. 2a), and in six (all coalescence-based) Nassariidae is paraphyletic in relation to other Buccinoidea (for support values of the Volutidae and Nassariidae nodes see Fig. 2a). One important finding among the family level relationships is the placement of Columbellidae within the Buccinoidea as a sister to the Colubrariidae-Colidae-Prosiphonidae-Eosiphonidae clade, which was strongly supported in all our analyses.

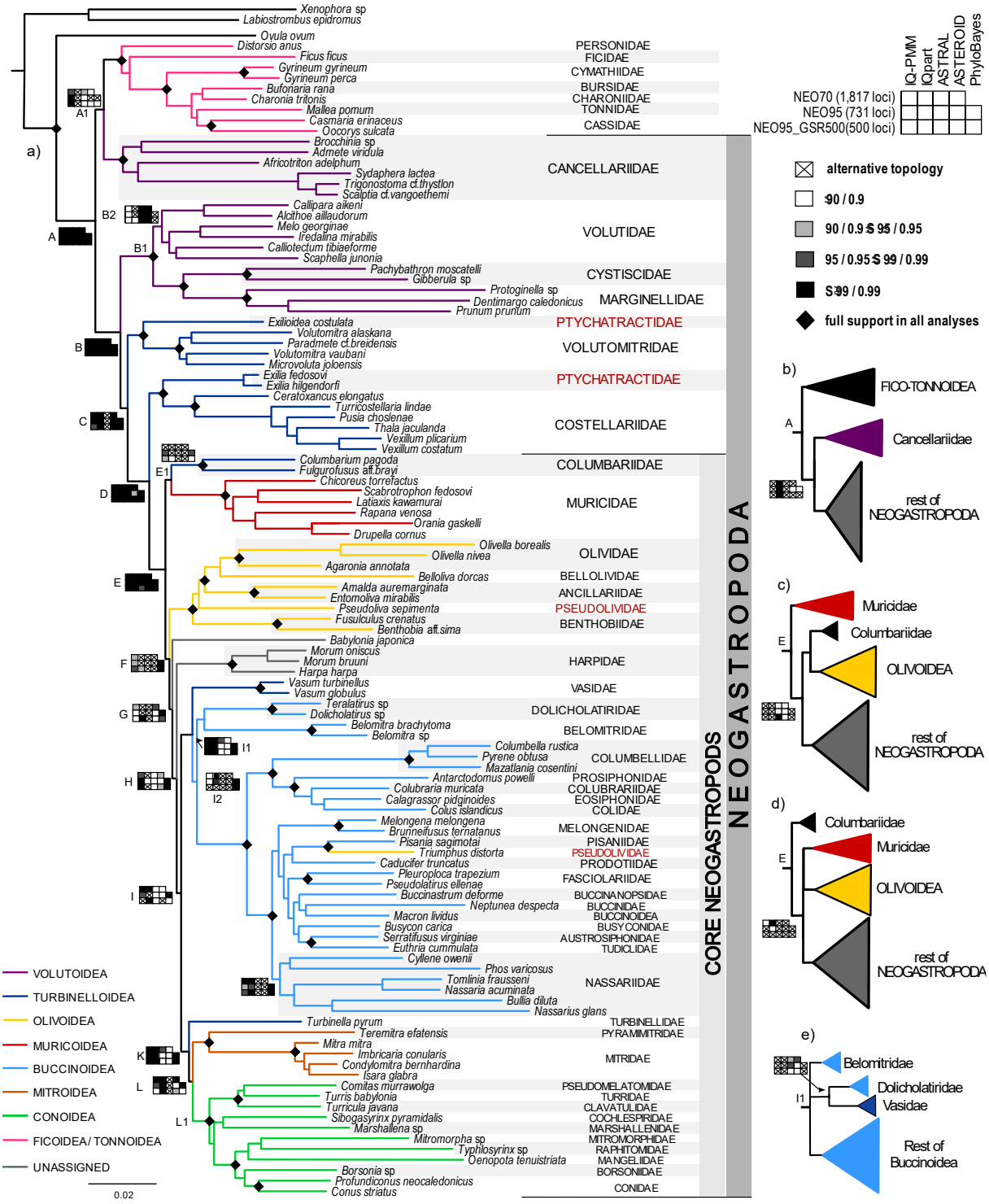


Figure 2. a. IQ-TREE-PMM tree generated with the NEO95 matrix; families consistently paraphyletic shown in red. Node supports for deep nodes summarized from 14 analyses as shown on top-right inset, IQ-PMM – IQ tree under Profile Mixture Model; tree branches are color-coded according to the current superfamily classification – bottom-left inset; scale as probability of substitution per site; b – e. Alternative topologies, and their support; b. Placement of Cancellariidae; c, d. Base of the 'core Neogastropoda'; e. Base of Buccinoidea.

Backbone relationships of Neogastropoda

To address backbone Neogastropoda relationships, we select the IQ-PMM tree obtained from the NEO95 matrix (Fig. 2a), which features the most frequently sampled topology at each backbone node (denoted as A-L), and at the base of Neogastropoda (node A1). We recognize as problematic nodes those with a consistently sampled alternative topology, criteria for consistency being: recovered (i) in at least three analyses, (ii) with at least two different inference methods, and (iii) with moderate or high support in at least one analysis. The first such problematic node concerns the placement of the family Cancellariidae at the base of Neogastropoda, either as a sister group to the Ficoidea–Tonnoidea clade (Fig. 2a), or to the rest of the Neogastropoda (Fig. 2b). The first topology receives high support in all IQ-PMM analyses, the second in the partitioned IQ-Tree analyses, whereas the coalescence-based and PhyloBayes inferences lack support for the placement of Cancellariidae.

The remaining neogastropod taxa always form a maximally supported clade (node B), and the topology at the three deepest nodes D–E is consistent and highly supported across most analyses. These nodes correspond to the consecutively branching off (i) Volutidae plus marginelliform gastropods (C), (ii) Volutomitridae plus *Exilioidea* (D), and (iii) Costellariidae plus *Exilia* (E). The remaining taxa are always recovered in a highly supported cluster (node E), which we refer to from here onwards as 'core Neogastropoda'. This clade comprises seven major lineages corresponding to (1) family Columbariidae, (2) family Muricidae, (3) superfamily Olivoidea (except *Triumphius*), (4) family Babyloniidae, (5) family Harpidae, (6) BV clade (Buccinoidea including *Triumphius* and Vasidae), and (7) TMC clade, (Turbinellidae, (Mitroidea, Conoidea)).

Three conflicting topologies at the base of core Neogastropoda (nodes E1, F) correspond to either Muricidae (Fig. 2c), or Columbariidae (Fig. 2d), or Muricidae and Columbariidae (most consistently recovered, Fig. 2a), being the sister group to all other core lineages. The latter topology is supported by two IQ-PMM and both PhyloBayes analyses, and invariably places the Olivoidea as the next branching lineage. In contrast, the partitioned IQ-TREE analyses (except the NEO95-500 matrix) favour Columbariidae as the first branching core lineage (Fig. 2c), whereas all coalescence-based analyses place Muricidae at the base of the core radiation, and Columbariidae as a sister group to Olivoidea, though usually without support.

The affinities among the four remaining lineages are generally more consistent and suggest a sister relationship between the BV and TMC clades, with Harpidae being a sister group to (BV, TMC), and Babyloniidae a sister group to (Harpidae, (BV, TMC)). Two further

problematic nodes, I2 and L1, concern relationships at the base of the buccinoidean and conoidean radiations respectively. The conflicting topologies at the base of Buccinoidea concern affinities of the early branching buccinoidean families Belomitridae and Dolicholatiridae. In all coalescence-based and some ML inferences either both these families, or only Dolicholatiridae appear more closely related to Vasidae than to the rest of the Buccinoidea (Fig. 2e, coalescence-based analyses, moderately supported, or lacking support).

Finally, a fourth major uncertainty affecting relationships at the base of Conoidea is a topology in which Cochlespiridae is the sister group to all other conoideans (Abdelkrim et al., 2018). This result is recovered in both PhyloBayes analyses, ASTEROID and IQ-PMM (both on the matrix NEO95_500), whereas the majority of the analyses suggest a sister relationship between Cochlespiridae and Marshallenidae. Possibly, this persistent grouping is an LBA artifact that is efficiently countered by 'cat gtr' model in Phylobayes (Uribe et al. 2018). Since the present study focuses on the relationships among the major Neogastropoda lineages, and the relationships within Conoidea have recently been addressed with phylogenomics (Abdelkrim et al., 2018), we have reduced the taxon coverage in this lineage. Having noted the robustly supported monophyly of the Conoidea in all analyses, we did not examine the sources of conflict among its lineages.

Sources of phylogenetic conflict

The PCA performed on the matrix summarising clade presence-absence (Supplementary Fig. S16) shows that topology at conflicting nodes depends more on the phylogenetic inference method, than on the matrix used. The two first principal components explained 45.7% of the observed variation. The first PC clearly separates the coalescence-based and concatenation-based analyses, indicating that ASTEROID trees are overall slightly more congruent with the ML- and Bayesian trees. The second PC separates the partitioned IQ-TREE trees (on top of the plot), the IQ-PMM trees, and the PhyloBayes trees, but also bears some signal of the matrix analysed: for each inference method, NEO95-500 trees are placed on the diagram lower than the trees obtained from the larger datasets NEO70 and NEO95.

The AU tests on the ΔSLS values calculated under GAMMALG4X did not prefer one of the conflicting topologies over another in any comparison (Supplementary Table S2). For the partitioned data, only the main topology at the nodes I1/I2 (monophyletic Buccinoidea) fit to the data significantly better than the respective alternative topologies. The t-test suggests that regardless of the query node, the loci with a strong ΔGLS on average show higher evolutionary rate (the reason why they offer some resolution), and under partitioned model, are more likely

to be affected by both compositional heterogeneity and saturation (Supplementary Table S2, Fig. 3). Under GAMMALG4X, the loci with a strong signal favouring affinity of the Cancellariidae with Ficoidea-Tonnoidea are clearly fewer, and they are significantly more affected by both saturation and high compositional heterogeneity (Fig. 3 d, j). Furthermore, loci with strong signal favouring Vasidae-Dolicholatiridae-Belomitridae affinity at the base of Buccinoidea show higher levels of saturation compared to those with strong support for the main topology (Fig. 3 l).

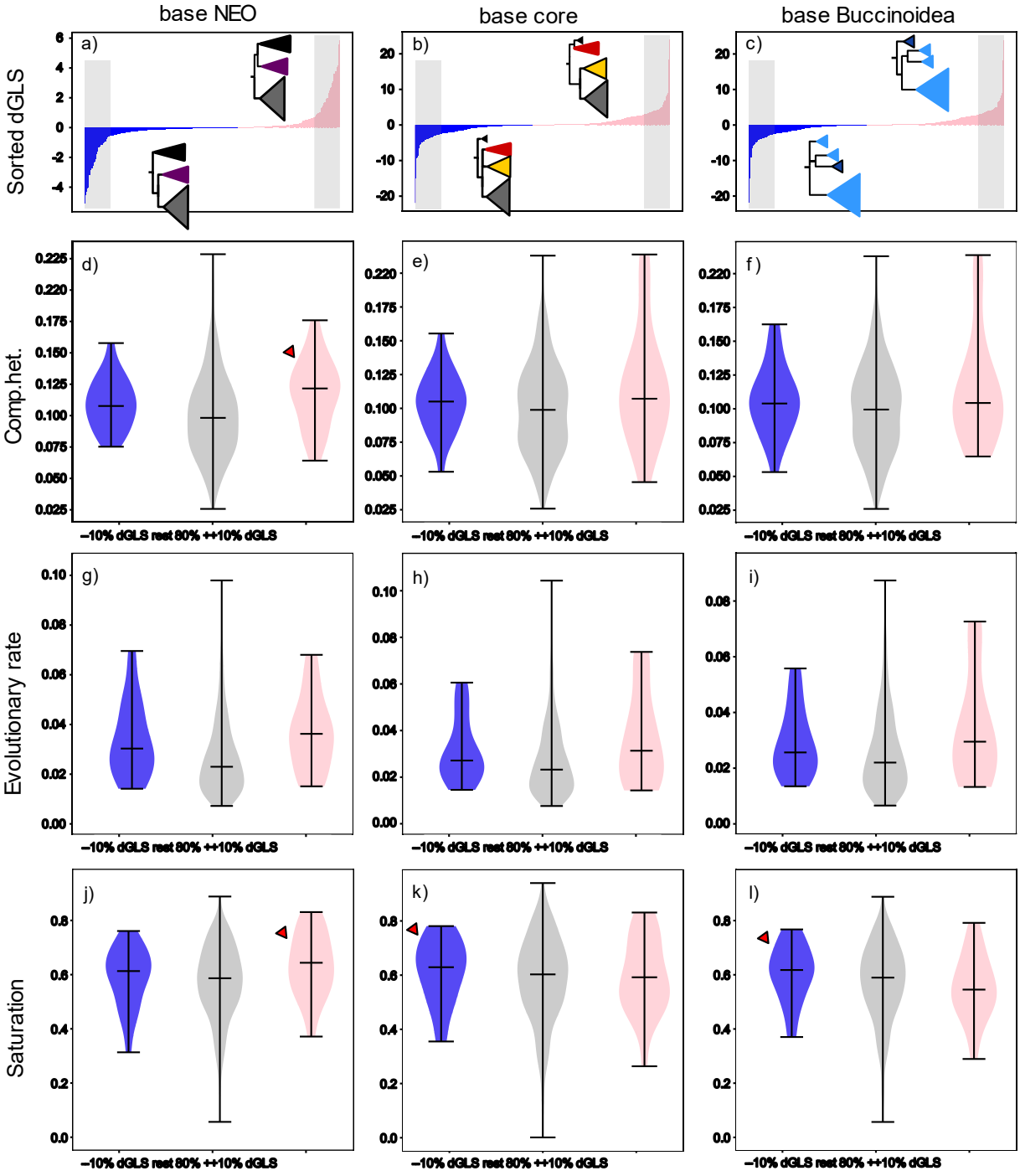


Figure 3. Phylogenetic signal (ΔGLS) supporting alternative topologies in three contradictory nodes. a-c. distribution of ΔGLS values in the 731 loci of the NEO95 matrix under GAMMAPROTLG4X model; pink bars representing loci supporting major topology (recovered in IQ-PMM analysis), blue supporting the alternative topology (from constrained topology in IQ-PMM analysis) – both respective topologies shown; grey zones mark 10% of loci with strongest ΔGLS signal for one or another topology. d-l. Loci metrics, compositional heterogeneity (second row), evolutionary rate (third row), and saturation at third codon position (bottom row) in three groups of loci by ΔGLS: 10% loci with strongest ΔGLS support for the alternative topology (blue), 80% of loci with weak ΔGLS values, irrespective of supported topology (grey), 10% loci with strongest ΔGLS support for the main topology (pink). Red triangles mark values mentioned in the text.

Discussion

Relationships at the base of Neogastropoda

Recent studies on various metazoan lineages shared the common conclusion that the presence of conflicting signals is an inherent property of phylogenomic datasets (Betancur-R. et al., 2019; Parins-Fukuchi et al., 2021; Cunha et al., 2022; Mongiardino Koch et al., 2023) and suggested that the true topology could be identified by accounting for technical errors and exploring sources of the conflicts (but see Mongiardino Koch et al., 2023). In our analysis, the most challenging conflict concerns the placement of Cancellariidae, either as a sister to the rest of Neogastropoda, or to Ficoidea-Tonnoidea, as supported by partitioned and ML-PMM analyses, respectively. We demonstrate that the latter topology may, at least partly, be driven by loci with high levels of saturation and compositional heterogeneity.

One further factor adding to uncertainty at this node is the inevitably difficult orthology inference due to the whole genome duplication (WGD) event that pre-dated the neogastropod radiation, confirmed by karyological (Hallinan and Lindberg, 2011) and whole genome data (Pardos-Blas et al., 2021; Farhat et al., 2023). Although redundant gene copies are usually quickly lost, the clades that have diverged shortly after a WGD event may differentially retain paralogs, leading to inaccurate phylogeny estimates (Xiong et al., 2022). Hence, the contradictory signals regarding the placement of Cancellariidae may be explained by a differential pattern of paralogs loss in Tonnoidea, Cancellariidae and the rest of Neogastropoda, as the separation of these lineages was likely one of the first major splits that followed the WGD (Hallinan and Lindberg, 2011; Farhat et al., 2023). Because the probability of the gene loss is proportional to the internal branch length (Xiong et al., 2022), the extent of the differential gene loss should be less in the pair Cancellariidae / Ficoidea-Tonnoidea compared to the pair Cancellariidae / rest of Neogastropoda, which are invariably separated by a higher sum of branch lengths (custom Python script S10-8, Supplementary Table S3). As a result, there would exist a pool of alignments where gene copies in Cancellariidae are orthologous to those in Ficoidea-Tonnoidea but not in the rest of Neogastropoda, and these alignments would expectedly favour the Cancellariidae / Ficoidea-Tonnoidea grouping.

It is noteworthy that none of our analyses recovered Cancellariidae as a sister to Tonnoidea plus Neogastropoda, a placement supported by recent mitogenomic phylogenies (Osca et al., 2015; Lemarcis et al., 2022) but based on a very limited sampling of Cancellariidae. Morphological data generally supports monophyly of Neogastropoda. However, the key anatomical traits for understanding Neogastropoda evolution, radula and valve of Leiblein, are highly aberrant in Cancellariidae (Modica et al., 2011), and they do not provide any clues on

the affinities of this enigmatic lineage. Further genomic data, whole genome assemblies, or a carefully curated set of longer loci would be instrumental for disentangling the relationships at the base of Neogastropoda radiation.

Relationships within the core Neogastropoda

We examined deep relationships within the order Neogastropoda based on both an unprecedented taxonomic coverage (112 neogastropod taxa representing 48 families) and a representative genomic sampling (from 1,817 loci with ~20.4% of missing data to 731 loci with 11.6 % missing data only). Although we failed to recover a single topology for the Neogastropoda tree, high support was retrieved for most backbone nodes, allowing to localize uncertainty to four specific nodes. Three of them are associated with the origin of remarkably species-rich radiations: the core Neogastropoda, the superfamily Buccinoidea and the superfamily Conoidea.

Within core Neogastropoda, the sister relationship of Columbariidae and Muricidae is morphologically plausible, albeit their similarities are mainly limited to shared plesiomorphies (Kantor 2002). The most frequently sampled alternative topology (Fig. 2c) results from the coalescence-based analyses, however, with low support values at query nodes. Furthermore, recent findings casted doubts on the ability of summary-based approaches to accurately resolve deep and intricate phylogenies (e.g., Gatesy and Springer, 2014). Therefore, we regard this alternative topology as rather unlikely. Similarly, the only analysis supporting Columbariidae as a sister group to all other core lineages (NEO70, partitioned IQ-TREE; fig. 1d) relies on a larger proportion of missing data, with inference performed on very short loci, resulting in the overall unrealistically high bootstrap support values (Thomson and Brown, 2022). Therefore, we consider the topology where the Columbariidae-Muricidae lineage represent the first offshoot within core Neogastropoda as the most probable. This topology is the most frequently sampled and is supported in nearly half concatenation-based inferences.

Two very short branches at the base of the Buccinoidea separate first the Vasidae and then the Dolicholatiridae and Belomitridae from the main stem of Buccinoidea. Vasidae, Dolicholatiridae, and Belomitridae share somewhat similar radulae, with bicuspidate lateral teeth (Medinskaya et al., 1996). However, Dolicholatiridae and Belomitridae, similarly to all other Buccinoidea, lack accessory salivary glands and an anal gland, whereas the latter is present in Vasidae. While it is tempting to speculate that the loss of accessory salivary glands and anal gland in Dolicholatiridae, Belomitridae and all other Buccinoidea is a result of single evolutionary event supporting their affinity, a shared loss of a trait cannot be considered

evidence of affinity (Strong and Lipscomb 1999). Therefore, the anatomical evidence is inconclusive, as to whether Dolicholatiridae-Belomitridae are closer to the Vasidae or to the major Buccinoidea clade. Phylogenetic uncertainty here is likely due to the series of very short branches followed by a longer one leading to the major Buccinoidea. Topology resolution in proximity of such patterns is susceptible to a biased signal from loci affected by saturation (Breinholt & Kawahara, 2013), and indeed we detected higher levels of saturation in loci with a strong signal for Vasidae-Dolicholatiridae-Belomitridae grouping. This result, and the generally consistent support for monophyletic Buccinoidea in our concatenation-based analyses, prompt us to consider this topology (Fig. 2a) as the most probable.

Rapid diversification of the core Neogastropoda coincided with the dramatic paleo-climatic events of the late Cretaceous and the K-Pg boundary (Vermeij, 1977). This period was marked by the origin of many lineage-specific morphological innovations, mainly associated with the dynamic evolution of foregut underpinning the diversification of feeding strategies in Neogastropoda (Ponder, 1973; Kantor, 2002). Parins-Fukuchi et al. (2023) suggested that the complex evolutionary patterns of genes linked to bursts of morphological disparity could also complicate phylogenetic inference. Similar to Neogastropoda, the evolutionary histories of two iconic vertebrate radiations – birds and mammals – suffer from a lack resolution at the phylogenetic splits typically aligned with the K-Pg boundary. Remarkably, even with significantly more genomic resources available in these lineages, certain relationships remain challenging to address due to pervasive phylogenomic conflicts. Nonetheless, we anticipate that the present phylogeny will serve as a valuable guide for future expansion of genomic resources for Neogastropoda. This expansion is crucial for understanding the evolutionary history of this remarkable group of marine invertebrates.

Relationships of Neogastropoda and their implications for systematics

Our findings unequivocally support the monophyly of five Neogastropod superfamilies: Conoidea, Muricoidea, Mitroidea, Olivoidea and Buccinoidea (with the reassignment of *Triumphius* from the Olivoidea to the Buccinoidea). Within Buccinoidea, we confidently place the previously disputed Columbellidae as the sister group to the Colubrariidae-Colidae-Prosiphonidae-Eosiphonidae clade. Furthermore, all the concatenation-based inferences confirmed the monophyly of Nassariidae, questioned by Kantor et al. (2022). Notably, we identify Vasidae for the first time as the sister group to the Buccinoidea. Indeed, the affinity of Vasidae and Buccinoidea *sensu* Kantor et al. (2022) is recovered in all our analyses, and has a

much stronger support than the Buccinoidea clade itself. Based on this outcome, we propose the inclusion of Vasidae in the superfamily Buccinoidea.

The scope of the superfamily Volutoidea must be restricted to the content of the clade including Volutidae and marginelliform gastropods (Fedosov et al., 2019). Future investigations are required to validate the monophyly of Volutidae and ascertain the placement of enigmatic taxa such as the families Granulinidae and Marginellonidae. The family Cancellariidae should definitely be assigned to a separate superfamily Cancellarioidea, as previously proposed by Ponder (1973), and Bouchet & Rocroi (2005).

Our analyses reveal the polyphyly of Turbinelloidea (*sensu* Fedosov et al., 2017), a result that necessitates profound revisions to neogastropod systematics. Some changes, such as the inclusion of Columbariidae in Muricoidea, and Vasidae in the Buccinoidea, can be readily inferred from the present phylogeny, others yet to be proposed. The existing scheme with eight superfamilies leaves out of superfamilies at least four major lineages retrieved in our analyses. Therefore, the establishment of four new superfamilies to accommodate i) Volutomitridae plus *Exilioidea*, ii) Costellariidae plus *Exilia*, iii) Babyloniidae, and iv) Harpidae, emerges as the most reliable systematic arrangement based on the reconstructed tree topology.

Data accessibility

Raw reads data (both transcriptomic and genomic) are available under the NCBI Bioproject PRJNA885117. Phylogenetic matrices, gene alignments and trees, output of the orthology inference software, as well as the original scripts are available as supplementary data at Dryad https://doi.org/10.5061/dryad.8931zcrx5.

Acknowledgments

Specimens were obtained during research cruises and expeditions organized by the MNHN and ProNatura International as part of the Our Planet Reviewed program, and by the MNHN and the Institut de Recherche pour le Développement as part of the Tropical Deep-Sea Benthos program (see Supp. Acknowledgments). We are grateful to Bruce Marshall (NMNZ, Wellington), Nerida Wilson (WAM, Perth), Mandy Reid (AMS, Sydney), Katrin Linse (British Antarctic Survey, Cambridge), Yasunori Kano (NSMTo. Tokyo), Paolo Albano (Stazione Zoologica 'Anton Dohrn', Napoli), Gustav Paulay (NHMF), Douglas Eernisse (California State University, Fullerton), Miroslav Harasewych and Dr Ellen Strong (USNM, Washington), Ivan Nehaev (St. Petersbourg State University), Anastassya Maiorova (NSCMB, Vladivostok) and Sofia Zvonareva (IPEE RAS) for providing specimens for the present study. We are grateful to

Barbara Buge (MNHN) for assistance in specimen curation, and to the team of SSM (UAR2700 - MNHN) for support in the lab. We are immensely grateful to Philippe Bouchet for providing access to the MNHN specimens, funds, and encouraging completion of the present study. We are grateful to two anonymous reviewers for their invaluable comments on the manuscript.

Funding

The present work was supported by the grants MSF to AF, RSF 16-14-10118 to YK and AF and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 865101) to NP.

References

Abdelkrim J., Aznar-Cormano L., Fedosov A., Kantor Y., Lozouet P., Phuong M., Zaharias P., Puillandre N. 2018. Exon-capture based phylogeny and diversification of the venomous gastropods (Neogastropoda, Conoidea). Mol. Biol. Evol. 35:2355–2374.

Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M., Nikolenko S.I., Pham S., Prjibelski A.D., Pyshkin A.V., Sirotkin A.V., Vyahhi N., Tesler G., Alekseyev M.A., Pevzner P.A. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. 19:455–477.

Betancur-R. R., Arcila D., Vari R.P., Hughes L.C., Oliveira C., Sabaj M.H., Ortí G. 2019. Phylogenomic incongruence, hypothesis testing, and taxonomic sampling: The monophyly of characiform fishes. Evolution. 73:329–345.

Bouchet P, Rocroi J–P. 2005. Classification and nomenclator of gastropod families, Malacologia 47: 1–397.

Bouchet P., Rocroi J.-P., Hausdorf B., Kaim A., Kano Y., Nützel A., Parkhaev P., Schrödl M., Strong E.E. 2017. Revised classification, nomenclator and typification of gastropod and monoplacophoran families. Malacologia 61:1–526.

Breinholt J.W., Kawahara A.Y. 2013. Phylotranscriptomics: saturated third codon positions radically influence the estimation of trees based on next-gen data. Genome Biol. Evol. 5:2082–2092.

Capella-Gutiérrez S., Silla-Martínez J.M., Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25:1972–1973.

Cunha T.J., Giribet G. 2019. A congruent topology for deep gastropod relationships. Proc.Biol. Sci. 286:20182776.

Cunha T.J., Reimer J.D., Giribet G. 2022. Investigating Sources of Conflict in Deep Phylogenomics of Vetigastropod Snails. Syst. Biol. 71:1009–1022.

Duchêne S., Archer F.I., Vilstrup J., Caballero S., Morin P.A. 2011. Mitogenome Phylogenetics: The Impact of Using Single Regions and Partitioning Schemes on Topology, Substitution Rate and Divergence Time Estimation. PLoS One. 6: e27138.

Emms D.M., Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20:238.

Farhat S., Modica M.V., Puillandre N. Whole genome duplication and gene evolution in the hyperdiverse venomous gastropods. Mol. Biol. Evol. accepted manuscript.

Fassio G., Modica M.V., Mary L., Zaharias P., Fedosov A.E., Gorson J., Kantor Y.I., Holford M., Puillandre N. 2019. Venom Diversity and Evolution in the Most Divergent Cone Snail Genus *Profundiconus*. Toxins 11: 623.

Fedosov A.E., Caballer Gutierrez M., Buge B., Boyer F., Sorokin P.V., Puillandre N., Bouchet P. 2019. Mapping the missing branch on Neogastropoda tree of life: molecular phylogeny of marginelliform gastropods. J. Moll. Stud. 85:439–451.

Fedosov A.E., Puillandre N., Herrmann M., Dgebuadze P., Bouchet P. 2017. Phylogeny, systematics and evolution of the family Costellariidae (Gastropoda: Neogastropoda). Zool. J. Linn. Soc. 179:541–626.

Fu L., Niu B., Zhu Z., Wu S., Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 28:3150–3152.

Gatesy J., Springer M.S. 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Mol. Phylogenet. Evol. 80:231–266.

Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–652.

Hallinan N.M., Lindberg D.R. 2011. Comparative Analysis of Chromosome Counts Infers Three Paleopolyploidies in the Mollusca. Genome Biol. Evol. 3:1150–1163.

Hammer Ø., Harper D.A.T., Ryan P.D. 2001. PAST: Paleontological statistics software package for education and data analyses. *Paleontol. Electron*. 4:1–9.

Hinchliff C.E., Smith S.A., Allman J.F., Burleigh J.G., Chaudhary R., Coghill L.M., Crandall K.A., Deng J., Drew B.T., Gazis R., Gude K., Hibbett D.S., Katz L.A., Iv H.D.L., McTavish E.J., Midford P.E., Owen C.L., Ree R.H., Rees J.A., Soltis D.E., Williams T., Cranston K.A. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc. Natl. Acad. Sci. U.S.A. 112: 12764–12769.

Kantor Y.I. 2002. Morphological prerequisites for understanding neogastropod phylogeny. Boll. Malacol. Suppl. 4:161–174.

Kantor Y.I., Fedosov A.E., Kosyan A.R., Puillandre N., Sorokin P.A., Kano Y., Clark R., Bouchet P. 2022. Molecular phylogeny and revised classification of the Buccinoidea (Neogastropoda). Zool. J. Linn. Soc. 194:789–857.

Katoh K., Standley D.M. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30:772–780.

Kocot K.M., Poustka A.J., Stöger I. Halanych K.M., Schrödl M. 2020. New data from Monoplacophora and a carefully-curated dataset resolve molluscan relationships. Sci. Rep. 10, 101.

Kuznetsova K.G., Zvonareva S.S., Ziganshin R., Mekhova E.S., Dgebuadze P., Yen D.T.H., Nguyen T.H.T., Moshkovskii S.A., Fedosov A.E. 2022. Vexitoxins: conotoxin-like venom peptides from predatory gastropods of the genus *Vexillum*. Proc. Biol. Sci. 289:20221152

Langmead B., Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 9:357–359.

Lanyon S.M. 1988. The Stochastic Mode of Molecular Evolution: What Consequences for Systematic Investigations? Auk. 105:565–573.

Lartillot N., Rodrigue N., Stubbs D., Richer J. 2013. PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. Syst. Biol. 62:611–615.

Lemarcis T., Fedosov A.E., Kantor Y.I., Abdelkrim J., Zaharias P., Puillandre N. 2022. Neogastropod (Mollusca, Gastropoda) phylogeny: A step forward with mitogenomes. Zool. Scr. 51:550–561.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–2079.

Medinskaya A.I., Harasewych M.G., Kantor Y.I. 1996. On the anatomy of *Vasum muricatum* (Born, 1778) (Neogastropoda, Turbinellidae). Ruthenica. 5:131–138.

Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., von Haeseler A., Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37:1530–1534.

Modica MV, Gorson J, Fedosov AE, Malcolm G, Terryn Y, Puillandre N, Holford M. 2019. Macroevolutionary analyses suggest environmental factors, not venom apparatus, play key role in Terebridae marine snail diversification. Sys. Bio. 69: 413–430

Modica M.V., Bouchet P., Cruaud C., Utge J., Oliverio M. 2011. Molecular phylogeny of the nutmeg shells (Neogastropoda, Cancellariidae). Mol. Phylogenet. Evol. 59:685–697.

Mongiardino Koch N. 2021. Phylogenomic Subsampling and the Search for Phylogenetically Reliable Loci. Mol. Biol. Evol. 38:4025–4038.

Mongiardino Koch N., Tilic E., Miller A.K., Stiller J., Rouse G.W. 2023. Confusion will be my epitaph: genome-scale discordance stifles phylogenetic resolution of Holothuroidea. Proc. Biol. Sci. 290:20230988.

Morel B., Williams T.A., Stamatakis A. 2023. Asteroid: a new algorithm to infer species trees from gene trees under high proportions of missing data. Bioinformatics. 39: btac832.

Olivera B.M., Showers Corneli P., Watkins M., Fedosov A. 2014. Biodiversity of Cone Snails and Other Venomous Marine Gastropods: Evolutionary Success Through Neuropharmacology. Annu. Rev. Anim. Biosci. 2:487–513.

Osca D., Templado J., Zardoya R. 2015. Caenogastropod mitogenomics. Mol. Phylogenet. Evol. 93:118–128.

Pardos-Blas J.R., Irisarri I., Abalde S., Afonso C.M.L., Tenorio M.J., Zardoya R. 2021. The genome of the venomous snail *Lautoconus ventricosus* sheds light on the origin of conotoxin diversity. GigaScience. 10: giab037.

Parins-Fukuchi C., Stull G.W., Smith S.A. 2021. Phylogenomic conflict coincides with rapid morphological innovation. Proc. Natl. Acad. Sci. U.S.A. 118: e2023058118.

Ponder W.F. 1973. The origin and evolution of Neogastropoda. Malacologia. 12:295–338.

Ponder W.F., Lindberg D.R., Ponder J.M. 2020. *Biology and Evolution of the Mollusca*. Volume 1. Boca Raton. 1-900.

Ponte G., Modica M.V. 2017. Salivary glands in predatory mollusks: Evolutionary considerations. Fronti. Physiol. 8:1–8.

Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature. 526:569–573.

Ranwez V., Douzery E.J.P., Cambon C., Chantret N., Delsuc F. 2018. MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. Mol. Biol. Evol. 35:2582–2584.

Riedel F. *Ursprung und Evolution der „höheren" Caenogastropoda*. Berlin. 1-265.

Rokas A., Carroll S.B. 2006. Bushes in the Tree of Life. PLOS Biol. 4:1899–1904.

Safavi-Hemami H., Brogan S.E., Olivera B.M. 2019. Pain therapeutics from cone snail venoms: From Ziconotide to novel non-opioid pathways. J. Proteomics. 190:12–20.

Shen X.-X., Hittinger C.T., Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1:0126.

Shimodaira H., Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics. 17:1246–1247.

Simone L.R.L. 2011. Phylogeny of the Caenogastropoda (Mollusca), based on comparative morphology. Arq. Zool. 42:161–323.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688–2690.

Strong E.E., Lipscomb D. 1999. Character Coding and Inapplicable Data. Cladistics 15:363–371.

Sutton M., Perales-Raya C., Gilbert I. 2016. A phylogeny of fossil and living neocoleoid cephalopods. Cladistics. 32:297–307.

Takezaki N., Figueroa F., Zaleska-Rutczynska Z., Takahata N., Klein J. 2004. The Phylogenetic Relationship of Tetrapod, Coelacanth, and Lungfish Revealed by the Sequences of Forty-Four Nuclear Genes. Mol. Biol. Evol. 21:1512–1524.

Thalén F. 2018. PhyloPyPruner: tree-based orthology inference for phylogenomics with new methods for identifying and excluding contamination. Lund University Student Papers. 8963554.

Thomson R.C., Brown J.M. 2022. On the Need for New Measures of Phylogenomic Support. Syst. Biol. 71:917–920.

Uribe J.E., González V.L., Irisarri I., Kano Y., Herbert D.G., Strong E.E., Harasewych M.G. 2022. A Phylogenomic Backbone for Gastropod Molluscs. Syst Biol. 71: 1271-1280.

Uribe J.E., Zardoya R., Puillandre N. 2018. Phylogenetic relationships of the conoidean snails (Gastropoda: Caenogastropoda) based on mitochondrial genomes. Mol. Phylogenet. Evol. 127:898–906.

Vermeij G.J. 1977. The Mesozoic marine revolution: evidence from snails, predators and grazers. Paleobiology. 3:245–258.

Wanninger A., Wollesen T. 2019. The evolution of molluscs. *Biol Rev*. 94:102–115.

Waterhouse R.M., Seppey M., Simão F.A., Manni M., Ioannidis P., Klioutchnikov G., Kriventseva E.V., Zdobnov E.M. 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. 35:543–548.

Whitfield J.B., Lockhart P.J. 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol. 22:258–265.

Xiong H., Wang D., Shao C., Yang X., Yang J., Ma T., Davis C.C., Liu L., Xi Z. 2022. Species Tree Estimation and the Impact of Gene Loss Following Whole-Genome Duplication. Syst. Biol. 71: 1348-1361.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 19:153.

Zou S., Li Q., Kong L. 2011. Additional gene data and increased sampling give new insights into the phylogenetic relationships of Neogastropoda, within the caenogastropod phylogenetic framework. Mol. Phylogenet. Evol. 61:425–435.

Title: Phylogenomics of Neogastropoda: the backbone hidden in the bush

Authors: Alexander E. Fedosov, Paul Zaharias, Thomas Lemarcis, Maria Vittoria Modica, Mande Holford, Marco Oliverio, Yuri I. Kantor, Nicolas Puillandre

Supplementary material

Bait design, laboratory procedures and recovery of the targeted loci

A total of 46 transcriptomes of 33 caenogastropod species were used for bait design. Of them, 24 venom gland transcriptomes of the superfamily Conoidea members have been published before (Gonzales and Saloma, 2014; Gorson et al., 2015; Zaharias et al., 2020; Fedosov et al., 2021), with most transcriptomes (21) representing Turridae (Zaharias et al., 2020). Transcriptomes of further 11 species were generated recently (Lemarcis et al., 2022) to cover phylogenetic diversity of the order. All transcriptomes were (re)-assembled using TRINITY v.2.11.0 with default parameters, using an embedded TRIMMOMATIC plugin, and skipping SALMON run (--no_salmon). The quality trimming parameters were as follows: 'TrueSeq3-PE.fa:2:40:15 SLIDINGWINDOW:4:15 LEADING:3 TRAILING:3 MINLEN:36'. The transcriptomic assemblies were clustered using CD-Hit (identity 0,99) to reduce assembly redundancy, and then aligned against the genome of *Lottia gigantea* (Simakov et al., 2013) to identify exon/intron boundaries (Phuong and Mahardika, 2018). Then we identified a subset of 4,456 exons (>180-bp) spanning approximately 1.3 Mb that were present in at least 2 families of Conoidea, and at least in 3 non-conoidean transcriptomes. Their sequences were extracted from the assembled contigs, translated using MACSE (Ranwez et al., 2018), and the obtained amino acid fasta files were aligned with MAFFT v6.240 with automatic strategy selection (Katoh and Standley, 2013). They were then concatenated into a supermatrix to build a preliminary phylogeny of the available transcriptomes (using RAxML with GTRGAMMA model on an unpartitioned aa matrix – Fig S1). Target recovery is known to reduce notably when targeted sequences are too divergent from the baits (Schott et al., 2017). One possible solution for highly diversified taxa is breaking down a phylogeny into a set of narrower clades, and designing a separate probe kit for each of them (Hugall et al., 2016). Alternatively, ancestral sequences of the targets can be inferred, and used for the probe design, which was shown to increase the capture efficiency in divergent lineages (Hugall et al., 2016; Abdelkrim et al., 2018). Because monophyly has not been supported for a number of neogastropod taxa,

and relationships of many lineages remain unclear, here we applied the latter strategy, guided by the obtained preliminary tree topology (Fig. S1). Because the target recovery in cross-species capture is more efficient when each locus is targeted by multiple probes with slightly varied sequence (Schott et al., 2017), we performed ancestral sequence reconstruction in five increasingly more restricted taxa sets. These were (i) *Ovula ovum* vs all ingroup Neogastropoda species, (ii) *Distorsio anus* vs all ingroup Neogastropoda species, (iii) *Babylonia areolata* vs all other 30 ingroup neogastropods, (iv) *Chicoreus torrefactus* and *Oliva sericea* vs other 28 ingroup neogastropods (but excluding *B. areolata*), (v) *Cantharus melanostomus* and *Vasum turbinellus* vs 28 ingroup neogastropods (excluding *B. areolata*, *Chicoreus torrefactus*, and *Oliva sericea*). Furthermore, two species, *Gymnobela* sp. and *Typhlosyrinx sp.* (Conoidea: Raphitomidae) generated long branches (Fig. S1), so we performed a sixth comparison: *Gymnobela* sp. and *Typhlosyrinx* sp. vs other 29 ingroup neogastropods. Therefore, each exon was represented in the dataset by up to six reconstructed ancestral sequences. This set of ancestral exon sequences were used to design 42,011 2x tiling 100-bp baits, which after the duplicate removal, produced a final set of 40,040 baits developed into a MyBait generation 5 biotinylated probes kit (Mycroarray, Arbor Biosciences).

Taxonomic sampling, library preparation, hybridization, and sequencing

We obtained Ethanol-preserved tissue samples of 141 taxa, covering all major phylogenetic lineages of the Neogastropoda, Tonnoidea + Ficoidea (six families of ten recognized), and the 'higher' Caenogastropoda' superfamiliy Stromboidea (2) – Table S1. We also analyzed 55 published and 3 newly generated transcriptomic datasets, but of these, only 12 with highest BUSCO completeness (Waterhouse et al., 2018), assessed against the mollusca_odb10 loci set, were selected for phylogeny, and eventually only nine retained for the final analyses. After the taxa with low (<50% in the final loci set) were removed (see below), we performed final analyses on the set of 112 taxa (Table S1), representing 48 neogastropod families out of 60 currently recognized (see below).

The library preparation was performed in three batches: the protocol of (Meyer & Kircher, 2009), with modifications as per (Abdelkrim et al., 2018) was used for the specimens in the 1st and 2nd batches, that were pooled after indexing into eight and six pools respectively. The

KAPA protocol was used for the thirty 3rd batch specimens, pooled in three pools. The hybridization was performed over 48 hours for all pools; following the pooled libraries cleanup and amplification; the bioanalyzer traces were generated to estimate the fragment length distribution, and Qubit was used for library quantifications. All 1st batch pools were sequenced on one lane of Illumina HiSeq 4000 (100 bp PE reads), and later pools 3-8 were recaptured and sequenced again on one lane of Illumina NovaSeq (150bp PE reads). All 2nd batch specimens were sequenced on one lane of Illumina HiSeq 4000 (150bp PE reads), and all 3rd batch specimens were sequenced on one lane of Illumina NovaSeq (150bp PE reads). The number of reads per library ranged from 851,299 (IM-2013-43718 *Glabella rosadoi*) to 44,946,221 (IM-2013-48309, *Xenophora* sp.), with an average of 11,4517,88 reads per library.

Loci assembly and recovery in targeted enrichment and transcriptomic data

The data assembly and further processing generally followed the protocol used by Abdelkrim et al. (2018), and utilized same Python 2 scripts with technical modifications to adapt them to our system. The original scripts not used by Abdelkrim et al., (2018) are mentioned explicitly, and provided in Supplementary data 10. We quality trimmed reads using Trimmomatic v0.39 (ILLUMINACLIP enabled, seed mismatch threshold = 2, palindrome clip threshold = 40, simple clip threshold = 15; SLIDING WINDOW enabled, window size = 4, quality threshold = 20; MINLEN = 36; LEADING = 15; TRAILING = 15) and used FLASH v1.2.11 (Magoc and Salzberg, 2011) to merge mate reads. To maximize recovery of targeted loci for exon capture datasets we used two assemblers, SPAdes v3.14 (Bankevich et al., 2012) with a set of kmer-sizes defined automatically depending on the read lengths, and Trinity v2.11 (Grabherr et al., 2011) with default kmer=25, and skipping the salmon run. The resulting captured data assemblies were merged using the Linux “cat” command, and the super-assembly redundancy was then reduced by performing contig clustering with CAP3 and CD-HIT (Fu et al., 2012) with percent identity = 99. Only Trinity was used for the transcriptomes, using same parameters as described for transcriptomic data above, and again, following a CD-Hit clustering with the identity of 0,99. We used BLASTN v2.2.31 (evalue 1e-20, and word size = 11) to identify contigs corresponding to the targeted loci, and used Exonerate v2.2.0 under the est2genome model to redefine boundaries of the target exons, since these were originally inferred from a

phylogenetically very divergent *Lottia*. Once we had reconstituted exon boundaries, we again used BLASTN against these (minimal BLAST e-value of 20, wordsize 11) to extract all putative target sequences from the assemblies. We mapped trimmed reads to the extracted target sequences using BOWTIE v2.2.7 (Langmead and Salzberg, 2012) with –very-sensitive-local and –no-discordant options to calculate percent of the on-target reads (Table S2), and thus evaluate efficiency of the capture. We marked duplicates with picard-tools v2.1.1 (http://broadinstitute.github.io/picard; last accessed July 29, 2018). The *mpileup* function of samtools v1.9 (Li et al., 2009), piped with the bcftools v1.3 *call* command generated variant calling files, which were used to filter out targets with high heterozygosity (> 2 Standard deviation from the mean) or with high proportion (>30%) of low coverage bases (4x threshold for base masking as 'N'). All surviving targets were combined based on the target identity PythonScript S10-1, trimmed to match the BLAST coordinates, and aligned using MAFFT v7.407 (Katoh and Standley, 2013) with *G-INS-i* algorithm and *–adjust_direction* flag. The obtained nucleotide per-target alignments were translated using MACSE v2.06 (Ranwez et al., 2018) *translateNT2AA* program, within sample duplicate aa sequences removed, and the remaining ones pooled back by sample (PythonScript S10-2). The orthology inference was performed by OrthoFinder v2.5.4 (Emms and Kelly, 2019), and identified 5,944 orthogroups (OG, Supp_data1). The largest 30 OGs containing on average >4 non-duplicate sequences per sample were discarded, and the following 3,000 OGs (OGs 000031-003030, comprising sequences of >68 taxa) were kept for downstream analyses. First filtering steps were performed on the nucleotide data, and once best sequence per sample per locus was identified, the clean alignments were again translated for phylogenomics. The OG gene trees were inferred from nucleotide data using RAxML v8.2.10 (Stamatakis 2006), with random seed and 100 rapid bootstrap replicates (`raxmlHPC-PTHREADS-SSE3 p 12345 x 12345 -#100 –m GTRGAMMA`). We used an original Python script S10-3 to first remove putative cross-contaminations, and then select the largest subtree with PhyloPyPruner. To detect potential cross-contaminations, we identified all clusters, comprising (nearly) identical sequences (branch length <0.0001). Then we checked, whether this cluster is included in a clade at any of the seven deeper nodes that has a bootstrap support >50. Because the previously published phylogenies included many of the species / specimens analyzed herein (Barco et al., 2010; Modica et al., 2011; Kantor et al., 2012, 2017, 2022; Fedosov et al., 2015, 2017, 2019; Abdelkrim et al., 2018; Strong et al., 2019), their placement within the family-level

taxa was *a priori* known. So, if a supported *parent* clade, with a nested cluster of (nearly) identical sequences was detected, we tentatively assigned it to one of 22 well established phylogenetic lineages of the Neogastropoda (Table S1). Those (nearly) identical sequences were excluded as potential contaminants that (i) do not match the supported *parent* clade annotation, and (ii) show lower coverage compared to those sequences of the supported *parent* clade that match its annotation by >2 SD from the mean of the latter. The detected cross-contaminations were removed from the alignment files, and pruned from the gene trees using newick utilities 'nw_prune' command. Finally, the largest subtree (LS) method of PhyloPyPruner was run to retain 1:1 ortholog tree and the respective sequences (`'phylopypruner --dir '+writedir+' --min-len 50 --mask longest --outgroup IM-2013-48309 IM-2013-53691 Ovula_ovum --prune LS --min-taxa 60 --min-gene-occupancy 45 --min-otu-occupancy 62'`). Then TRIMAL v1.2 (Capella-Gutiérrez et al. 2009) was used to remove the start-end alignment columns with gaps in >50% of sequences.

When sample-duplicates were removed, we performed second round of OG-tree reconstruction (using the same parameters of RaXML), to guide the long branch removal. Originally, we also approached this task with PhyloPyPrunner –trim-lb command, with a cutoff branch length of +/- 4.5 SD from median. We noticed that under certain conditions, use of this tool leads to unjustified removal of genuine sequences representing divergent lineages, resulting in lineage-specific increase of missing data. Therefore, we used custom Python script S10-4, to perform a more flexible pruning of long terminal branches. The branch lengths of some taxa showed a bimodal distribution, because the branch lengths of singleton records (i.e., when a sample is the sole representative of its phylogenetic lineage in the locus alignment) notably exceed branch lengths of non-singletons (i.e., when other representative samples of the same lineage are present in the tree). Therefore, for each sample, we computed statistics of branch lengths separately for singleton records and for non-singletons. We removed only those branches that exceeded both, (i) the median branch length of the sample across trees by more than 2.7 SD (if a sample is a sole representative of its lineage in a given tree, then its branch length was compared to singleton statistics, if not, to non-singletons), and (ii) the median terminal branch length computed across the given tree, also by more than 2.7 SD. Thus filtered loci were translated (MACSE, as detailed above), and used

for the final gene tree inference as detailed by Cunha & Giribet (2019). We examined the gene trees resulting from the reconstruction, in particular, to estimate the topologies obtained for the most complex OGs. In a series of gene trees, an obviously artificial topology was recovered: with a few clusters of identical sequences not reflecting the known patterns of lineages relationships, and likely arising from the degree of sequence heterogeneity being by far higher than our solutions could cope with. We noted that gene trees showing such pattern could be detected by a bi-modal distribution of uncorrected pairwise distances among samples, and implemented the test of the uncorrected pairwise distances distribution in a custom Python script S10-5. We applied it first to the gene trees of Cunha et al. (2021) curated loci set to ensure that no locus will be identified as potentially problematic. Then the loci with bi-modal distribution of uncorrected pairwise distances were removed from our data, generating a starting dataset of 1817 loci, each comprising >= 35 aa sites, and >= 70 taxa, with a total of 125,508 aa sites, and 20.8% of missing data. This dataset comprised 112 taxa, (minimal data occupancy 49.8%, median data occupancy 19.9%)

References

Abdelkrim J., Aznar-Cormano L., Fedosov A., Kantor Y., Lozouet P., Phuong M., Zaharias P., Puillandre N. 2018. Exon-capture based phylogeny and diversification of the venomous gastropods (Neogastropoda, Conoidea). *Molecular Biology and Evolution*. 35:2355–2374.

Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M., Nikolenko S.I., Pham S., Prjibelski A.D., Pyshkin A.V., Sirotkin A.V., Vyahhi N., Tesler G., Alekseyev M.A., Pevzner P.A. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*. 19:455–477.

Capella-Gutiérrez S., Silla-Martínez J.M., Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 25:1972–1973.

Cunha T.J., Giribet G. 2019. A congruent topology for deep gastropod relationships. *Proceedings of the Royal Society B*. 286:20182776.

Emms D.M., Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*. 20:238.

Fassio G., Modica M.V., Mary L., Zaharias P., Fedosov A.E., Gorson J., Kantor Y.I., Holford M., Puillandre N. 2019. Venom Diversity and Evolution in the Most Divergent Cone Snail Genus *Profundiconus*. *Toxins*. 11:623.

Fedosov A., Zaharias P., Puillandre N. 2021. A phylogeny-aware approach reveals unexpected venom components in divergent lineages of cone snails. *Proceedings of the Royal Society B* . 288:20211017.

Fu L., Niu B., Zhu Z., Wu S., Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 28:3150–3152.

Gonzales D.T.T., Saloma C.P. 2014. A bioinformatics survey for conotoxin-like sequences in three turrid snail venom duct transcriptomes. *Toxicon*. 92:66–74.

Gorson J., Ramrattan G., Verdes A., Wright E.M., Kantor Y.I., Srinivasan R.R., Musunuri R., Packer D., Albano G., Qui W.-G., Holford M. 2015. Molecular diversity and gene evolution of the venom arsenal of Terebridae predatory marine snails. *Genome Biology and Evolution*. 7:1761–1778.

Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology,* 29:644–652.

Hugall A.F., O'Hara T., Hunjan S., Nilsen R., Moussalli A. 2016. An Exon-Capture System for the Entire Class Ophiuroidea. *Molecular Biology and Evolution*, 33:281–294.

Katoh K., Standley D.M. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* . 30:772–780.

Lemarcis T., Fedosov A.E., Kantor Y.I., Abdelkrim J., Zaharias P., Puillandre N. 2022. Neogastropod (Mollusca, Gastropoda) phylogeny: A step forward with mitogenomes. *Zoologica Scripta*. 51:550–561.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 25:2078–2079.

Magoc T., Salzberg S.L. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*. 27:2957–2963.

Phuong M.A., Mahardika G.N. 2018. Targeted sequencing of venom genes from cone snail genomes improves understanding of conotoxin molecular evolution. *Molecular Biology and Evolution*. 35:1210–1224.

Ranwez V., Douzery E.J.P., Cambon C., Chantret N., Delsuc F. 2018. MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. Molecular Biology and Evolution. 35:2582–2584.

Schott R.K., Panesar B., Card D.C., Preston M., Castoe T.A., Chang B.S.W. 2017. Targeted Capture of Complete Coding Regions across Divergent Species. *Genome Biology and Evolution*. 9:398–414.

Simakov O., Marletaz F., Cho S.-J., Edsinger-Gonzales E., Havlak P., Hellsten U., Kuo D.-H., Larsson T., Lv J., Arendt D., Savage R., Osoegawa K., de Jong P., Grimwood J., Chapman J.A., Shapiro H., Aerts A., Otillar R.P., Terry A.Y., Boore J.L., Grigoriev I.V., Lindberg D.R., Seaver E.C., Weisblat D.A., Putnam N.H., Rokhsar D.S. 2012. Insights into bilaterian evolution from three spiralian genomes. *Nature*. 493:526.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 22:2688–2690.

Thalén F. 2018. PhyloPyPruner: tree-based orthology inference for phylogenomics with new methods for identifying and excluding contamination. *Lund University Student Papers*. 8963554.

Waterhouse R.M., Seppey M., Simão F.A., Manni M., Ioannidis P., Klioutchnikov G., Kriventseva E.V., Zdobnov E.M. 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution*. 35:543–548.

Zaharias P., Pante E., Gey D., Fedosov A.E., Puillandre N. 2020. Data, time and money: evaluating the best compromise for inferring molecular phylogenies of non-model animal taxa. *Molecular Phylogenetics and Evolution*. 142:106660.

Supplementary figures.

Figure S1. Transcriptome based tree of the Neogastropoda used for the probe design. The subsets of taxa used for the ancestral states reconstruction are marked with color bars on the right, and the respective nodes – by color triangles. The newly generated transcriptomes are marked with dark-blue circles.

Figure S2. IQ-TREE-PMM tree reconstructed from the unpartitioned matrix NEO70.

Figure S3. IQ-TREE-part tree reconstructed from the partitioned matrix NEO70.

Figure S4. ASTRAL-III multispecies coalescent tree reconstructed from the gene trees of the 1817 loci of the NEO70 data set.

Figure S5. ASTEROID multispecies coalescent tree reconstructed from the gene trees of the 1817 loci of the NEO70 data set.

Figure S6. IQ-TREE-PMM tree reconstructed from the unpartitioned matrix NEO95.

Figure S7. IQ-TREE-part tree reconstructed from the partitioned matrix NEO95.

Figure S8. ASTRAL-III multispecies coalescent tree reconstructed from the gene trees of the 1817 loci of the NEO95 data set.

Figure S9. ASTEROID multispecies coalescent tree reconstructed from the gene trees of the 1817 loci of the NEO95 data set.

Figure S10. Bayesian tree (Phylobayes), reconstructed from the NEO95 matrix, from the chains 2 and 3 trees, after discarding first 25% trees as burn in.

Figure S11. IQ-TREE-PMM tree reconstructed from the unpartitioned matrix NEO95_500.

Figure S12. IQ-TREE-part tree reconstructed from the partitioned matrix NEO95_500.

Figure S13. ASTRAL-III multispecies coalescent tree reconstructed from the gene trees of the 1817 loci of the NEO95_500 data set.

Figure S14. ASTEROID multispecies coalescent tree reconstructed from the gene trees of the 1817 loci of the NEO95_500 data set.

Figure S15. Bayesian tree (Phylobayes), reconstructed from the NEO95_500 matrix, from the chains 2 and 4 trees, after discarding first 25% trees as burn in.

Figure S16. PCA scatterplot showing topological congruence of the 20 reconstructed trees based on the analysis of the clade presence/absence data matrix. Data point marked in green correspond to ASTRAL trees, in yellow – ML analysis of the complete nucleotide matrices M2, M3 and M2&3, blue – IQ-TREE PMM trees, dark red – ML and BI analyses of the 1st and 2nd codon position nucleotides of M2, M3 and M2&3.

Figure S17. Phylogenetic signal (ΔGLS) supporting alternative topologies in three contradictory nodes. Top row: distribution of ΔGLS values in the 731 loci of the NEO95 matrix under GAMMAPROTLG4X model; pink bars representing loci supporting major topology (recovered in IQ-PMM analysis), blue supporting the alternative topology (from constrained in IQ-PMM analysis) – both respective topologies shown; grey zones mark 10% of loci with highest ΔGLS values supporting one or another topology. Three bottom rows: statistics of Compositional heterogeneity (second row), Evolutionary rate (third row), and Saturation at third codon position (bottom row) for loci in three groups: 10% loci with strongest ΔGLS support for the

alternative topology (blue), 80% of loci with weak ΔGLS values, regardless of which topology they support (grey), 10% loci with strongest ΔGLS support for the main topology (pink). Red triangles mark values mentioned in the text.

ANNEXE 2 :

Tableau des 1728 échantillons sélectionnés pour réaliser les phylogénies par la méthode de capture d'exons. Le tableau est à retrouver sur le lien suivant : https://github.com/Hyperdiverseproject/Exon_capture

ANNEXE 3 :

Figures supplémentaires des 29 arbres phylogénétiques produits lors de cette thèse : 14 arbres pour le jeu de données néogastéropodes et 15 pour le jeu de données Raphitomidae : https://github.com/Hyperdiverseproject/Exon_capture

Phylogénie des néogastéropodes : développements méthodologiques et évaluation du succès de l'approche par capture d'exons.

Les néogastéropodes (Mollusca, Neogastropoda) constituent un groupe de prédateurs marins hyperdiversifié, qui inclut plus de 15000 espèces décrites. Malgré des efforts significatifs afin de résoudre les relations phylogénétiques entre les super-familles et familles, celles-ci restent largement irrésolues et basées sur un échantillonnage limité. L'enjeu de la thèse était donc de produire une nouvelle phylogénie, basée sur un plus grand nombre de marqueurs moléculaires, et sur un échantillonnage au rang générique le plus exhaustif possible. Dans un premier chapitre introductif, je présente une phylogénie moléculaire basée sur les génomes mitochondriaux, mais qui restent encore limitée, constat nous ayant conduit à la mise en place d'une approche de capture d'exons. Dans un deuxième chapitre, je présente la méthode de capture d'exons, et en particulier la stratégie de design de nouvelles sondes de capture appliquée, ainsi que le protocole d'échantillonnage mis en place, incluant environ 800 des 1200 genres considérés comme valides actuellement, ainsi qu'un grand nombre de groupes externes. J'ai inclus également un échantillonnage plus dense (plusieurs espèces par genres) pour la famille des Raphitomidae, de façon à tester le pouvoir résolutif de nos marqueurs à différentes niveaux taxonomiques. Le chapitre suivant est consacré à l'évaluation de l'impact de la distance phylogénétique sur la capacité des sondes à capturer les exons cibles, à l'aide d'un jeu de données adaptés. Enfin le quatrième et dernier chapitre présente les phylogénies moléculaires des néogastéropodes et des Raphitomidae, et leur impact potentiel sur la classification. Les résultats obtenus seront ensuite utilisés dans le cadre du projet HYPERDIVERSE pour étudier la dynamique de diversification du groupe et analyser les processus qui pourraient être à l'origine du succès évolutif du groupe.

Mots-clés : Neogastropoda, Phylogénie moléculaire, Capture d'exons, Hyperdiversification.

Phylogeny of the neogastropods: methodological developments and evaluation of the success of the exon capture approach.

The neogastropods (Mollusca, Neogastropoda) constitute a hyperdiversified group of marine predators, which includes more than 15,000 described species. Despite recent significant efforts to resolve the phylogenetic relationships between superfamilies and families, these remain largely unresolved and based on limited sampling. The objective of the thesis was therefore to produce a new phylogeny, based on a greater number of molecular markers, and on sampling at the generic rank that is as exhaustive as possible. In the introductory chapter, I present a molecular phylogeny based on mitochondrial genomes, but which remains limited, an observation that led us to the implementation of an exon capture approach. In the second chapter, I present the exon capture method, and in particular the applied design strategy for new capture probes, and the sampling protocol implemented, including approximately 800 of the 1,200 genera currently considered as valid, as well as a large number of external groups. I included denser sampling (several species per genera) for the Raphitomidae family, in order to test the resolving power of our markers at different taxonomic levels. The following chapter is devoted to evaluating the impact of phylogenetic distance on the ability of the probes to capture the target exons, using a suitable dataset. The fourth and final chapter presents the molecular phylogenies of neogastropods and Raphitomidae, and their potential impact on the classification. The results obtained will be used within the framework of the HYPERDIVERSE project to study the dynamics of diversification of the group and to analyze the processes which could be at the origin of the evolutionary success of this group.

Keywords: Neogastropoda, Molecular phylogeny, Exon capture, Hyperdiversification.