Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

École doctorale Sciences de la vie et de la santé | ED 414 Université de Strasbourg

Doctoral School of Life and Health Sciences

ICube – UMR 7357

THÈSE **présentée par / DISSERTATION presented by :**

Christelle RUTZ

soutenue le /defended on : 18 décembre 2024

pour obtenir le grade de **/** to obtain the grade of **:**

Docteur de l'Université de Strasbourg / Strasbourg University Doctor

Discipline/Spécialité **/** Discipline/Specialty : Sciences de la vie et de la santé - génétique

Assembly of the giant genome of the noble crayfish and genome evolution of Decapoda

THÈSE dirigée par / DISSERTATION supervisor :

Dr. LECOMPTE Odile	Professor, University of Strasbourg
Dr. THEISSINGER Kathrin	Research Director, University of Giessen

RAPPORTEURS :

Dr. PANAUD Olivier	Professor, University of Perpignan Via Domitia
Dr. FLOT Jean-François	Professor, Université Libre de Bruxelles

AUTRES MEMBRES DU JURY / OTHER MEMBERS OF THE JURY :

Dr. FRIEDRICH Anne	Associate Professor, University of Strasbourg
Dr. RIVALS Eric	CNRS Research Director, University of Montpellier
Dr. PAULS Steffen	Professor, University of Giessen

A ma maman

Remerciements

Je tiens, en premier lieu, à exprimer ma sincère reconnaissance aux Dr. Anne Friedrich, Dr. Eric Rival, Pr. Jean-François Flot, Pr. Olivier Panaud, et Pr. Steffen Pauls, pour avoir accepté de participer à mon jury de thèse et pour l'honneur qu'ils m'ont fait d'examiner mon travail.

Les années de recherches qui ont mené à l'aboutissement de ce manuscrit ont été riches en rebondissements, tant professionnels que personnels. Tout au long de cette aventure intense, j'ai été accompagnée et soutenue par de nombreuses personnes, sans qui ce travail n'aurait jamais vu le jour dans sa forme actuelle. Je vous en remercie tous profondément.

Je voudrais d'abord remercier toute l'équipe du CSTB, composée aussi bien de ceux qui nous ont quittés pour de nouveaux horizons que de ceux qui viennent tout juste d'arriver. Cette équipe est peuplée de personnalités colorées, qui ont su faire de cet endroit bien plus qu'un simple lieu de travail. Merci pour votre passion pour les nanars et pour les soirées jeux, et pour l'ambiance chaleureuse que vous avez créée. Merci Laëtitia et Claudine pour votre bienveillance et votre soutien. Merci Arnaud, véritable encyclopédie des solutions informatiques, de m'avoir sauvée tant de fois ! Merci à Olivier pour ton énergie débordante et ta voix douce et inaudible (c'est ironique, bien entendu). Merci Romain pour m'avoir prêté ton petit canard. Merci Julie pour ta bienveillance au sein de l'équipe, pour ton aide et pour avoir corrigé mon anglais. Merci Hiba et Amani, mes voisines de bureau, pour avoir égaillé mes journées. Merci Yannis, t'avoir vu sur la fin de ta thèse est également une des raisons qui m'ont motivée à me lancer dans la mienne, et merci d'être revenu (même si ce n'est pas pour moi bien sûr), parce que ton aide m'a été bien précieuse pour ces dernières semaines. Merci également aux Nicolas, aux Corentins, à Dorine, Kirsley, Audrey, Tam'si et Julien, d'avoir été là durant ces dernières années.
Un grand merci également à mon équipe GEODE, partagée entre la France et l'Allemagne. Merci Caterina pour avoir tant de fois corrigé toutes mes fautes d'orthographe et de grammaire et d'avoir été si intransigeante pour la rédaction de mon papier. Ça m'a beaucoup fait progresser. Merci Luka d'avoir toujours eu quelque chose à redire ou à rajouter afin de toujours faire progresser les choses. Merci à Lena d'avoir recommencé encore et encore toutes ces manips dans le désespoir de pouvoir enfin réussir ce séquençage. Plus largement, merci à tous les trois pour votre soutien, vos encouragements, et les bons moments partagés.
J'aimerais à présent exprimer ma profonde gratitude à mes deux directrices de thèse, Odile et Kathrin. Tout au long de ce parcours, votre bienveillance et votre engagement ont fait toute la différence. Vous avez toujours répondu présentes pour me rassurer et me redonner confiance lorsque les doutes s'installaient. Même lorsque les choses étaient loin d'être faciles, vous ne m'avez jamais laissée tomber, et je vous en serai éternellement reconnaissante. Votre aide, votre soutien indéfectible, votre patience et votre compréhension m'ont permis de sortir la tête hors de l'eau dans les moments les plus difficiles. Sans vous, je ne serais pas arrivée là où je suis aujourd'hui. Merci, du fond du cœur, pour tout ce que vous avez fait pour moi.

Sur un plan plus personnel j'aimerais remercier mes amis venant des quatre coins du monde, avec vous, je me sens prête pour le reste, *Parabellum*. Vous n'êtes que des cinglés, mais vous avez joué un rôle particulier dans le maintien de ma santé mentale, lava y'all.
A mes amis de master Delphine, Ana, Fadwa, Victor, et Franck, qui ont eu la même idée folle de faire une thèse, presque tous déjà docteur(e)s au moment où j'écris ces mots. Je suis

tellement honorée d'avoir été à vos côtés dans ce voyage incroyable et inoubliable. Je chéris profondément l'amitié que nous avons construite au fil de ces années, une amitié née de nos défis communs, de nos joies et de moments plus difficiles. Ce parcours qu'est la thèse est extraordinaire en soi, mais vous l'êtes encore bien davantage. Merci pour votre inspiration et votre soutien, sans lesquels ce chemin n'aurait pas eu la même saveur.
Les personnes qui vont suivre ont toutes joué un rôle incroyablement important tout au long de ma vie, vous êtes mes piliers dans cette vie et vous l'avez été pour cette thèse. Je ne saurais comment exprimer à quel point je vous aime et vous suis reconnaissante de faire partie de vos vies et que vous fassiez partie de la mienne. Merci Morgane, ma plus ancienne amie, ma sœur. Merci Margot et Lola, mes merveilleuses filleules. Merci Ludo, mon ami, le père de mes filleules, le compagnon de ma meilleure ami (oui Ludo, dans cet ordre-là). Merci Nathalie, Henri, Nicolas et Ben. Merci Delphine pour toutes nos aventures. Merci Noëlle et Frédérique. Merci Caro et Pascal pour votre amitié et pour toute votre aide et votre accueil. Merci Lulu, on sait tous les deux qu'il fera beau demain. Merci Melih, le plus grand cœur qui existe sur terre.

Pour terminer, j'aimerais remercier toute ma famille. Je sais qu'un grand nombre d'entre vous sera présent le jour J pour voir cette aventure se finaliser. Vous avez toujours été d'un soutien indéfectible et une aide émotionnelle incommensurable, en particulier durant les derniers mois de cette thèse. J'espère vous rentre fiers. Aussi fière que je le suis de faire partie de votre famille. Une pensée particulière pour mon grand-père, qui n'aura pas pu voir la fin de cette aventure mais qui je le sais a toujours été convaincu que j'y arriverais.
Merci infiniment à mes parents que j'aime de tout mon cœur, pour votre amour inconditionnel et votre soutien indéfectible. Vous m'avez toujours encouragée à aller plus loin, à croire en moi et en ce que je faisais, et cela a compté plus que vous ne le saurez jamais. Papa, merci pour ta fierté sans faille, pour toutes les fois où tu as vanté mes mérites auprès de ceux que tu rencontrais, et pour ta curiosité sincère, même quand mes explications te laissaient perplexe, et ce même après la vingtième tentative. Maman, il m'est difficile de trouver les mots pour exprimer à quel point tu me manques. Tu m'as donné tant de force au fil des ans, tant de réconfort dans les moments où je doutais de moi, et le courage nécessaire pour persévérer. J'ai aussi terminé cette thèse pour toi. Je regrette tellement que tu ne sois plus là pour voir ce que j'ai accompli, pour voir la fierté que j'imagine dans tes yeux. J'espère que tu sais à quel point je t'aime et à quel point ton soutiens et ton amour ont compté pour moi, non seulement pour achever cette thèse, mais aussi pour tout ce qui fait de moi ce que je suis. Tu me manques chaque jour.
Sur une note plus légère pour finir ces remerciements, merci à mes chiens, Hypnos et Dysis, pour avoir mangé certaines feuilles de notes, mais surtout pour toutes les bêtises qui m'ont fait sourire particulièrement dans les moments de stress.

Merci à tous. Du fond du cœur.

Summary

List of figures

List of tables

List of annexes

List of abbreviations

BGE: Biodiversity Genomics Europe

CCS: Circular Consensus Sequences

CLR: Continuous Long Reads

COST: European Cooperation in Science and Technology

DNA: deoxyribonucleic acid

EBP: Earth BioGenome Project

ERGA: European Reference Genome Atlas

ESD: Environmental Sex Determination

G10K: Genome 10K Project

G-BiKE: genomic biodiversity knowledge for resilient ecosystems

GO: Gene Onthology

GSD: Genetic Sex Determination

HiFi: High Fidelity

mtDNA: mitochondrial deoxyribonucleic acid

NGS: Next Generation Sequencing

ONT: Oxford Nanopore Technology

OI: OrthoInspector

PacBio: Pacific Bioscience Inc.

RE: Repetitive Elements

RNA: ribonucleic Acid

SatDNA: Satellite DeoxyriboNucleic Acid

TE: Transposable elements

TGS: Third Generation Sequencing

WGS: Whole Genome Sequencing

INTRODUCTION

Chapter 1 – Crayfish, keystone species in aquatic ecosystems

A keystone species, as defined by Robert Paine, is a singular species with a high trophic status whose activities exert a disproportionately significant influence on the patterns of species occurrence, distribution, and density (Paine, 1969). To understand why crayfish are considered as keystone species, in this chapter, I will first describe the general classification of Crustacea, Decapoda and crayfish and their evolutionary position. Subsequently, I will delve into the main characteristics of crayfish, their vital role within their environment and their commercial use. I will then focus on European crayfish species and the challenges and threats they are currently facing. In particular, I will discuss the devastating impact of the crayfish plague, which stands as the primary threat to European crayfish and has caused massive population loss. To conclude this chapter, I will present how genomics can help species conservation efforts.

1.1 General presentation

Freshwater crayfish - one animal, several names:

crayfish, crawdad, crawfish, yabby, freshwater lobster, or mudbug.

1.1.1 Unveiling the diversity of Crustacea

The crustacean subphylum belongs to the arthropod phylum which also includes hexapods, myriapods (centipedes and millipedes), and chelicerates (spiders, mites, scorpions) subphyla. Crustaceans encompass a wide range of organisms, with over 67 000 described species (Zhang, 2011), representing only a fraction of the estimated undiscovered species. Crustaceans demonstrate remarkable adaptability and inhabit various marine and freshwater environments, with some species even adapted to live on land (VanHook and Patel, 2008). They employ various feeding strategies, including filter-feeder, scavenging, grazing, hunting prey, or living as parasites.

The anatomy of crustaceans can be similar to other arthropods with the ventral nerve cord, joined limbs, compound eyes, exoskeletons, and segmented body plans consisting of functional units known as tagmata (VanHook and Patel, 2008). The general body plan of crustaceans typically includes three primary tagmata corresponding to the head, the thorax

(called pereon in crustacea species), and the abdomen (called pleon) along with a tail called a telson, although there are variations (Averof and Patel, 1997; Zrzavý and Štys, 1997). Protected by a hard chitinous exoskeleton, crustaceans have relatively soft-bodied structures that they shed to be able to grow (Nagasawa, 2012). Moreover, they exhibit remarkable diversity in the morphology of their appendages, both between species and within the same organism (Waltling and Thiel, 2013). These limbs are specialized for various functions such as walking, eating, grooming, swimming, or mating.

While some crustaceans have successfully adapted to terrestrial habitats, most still rely on water for breeding, except for terrestrial isopods (VanHook and Patel, 2008). Reproductive strategies vary greatly, including sexual and asexual lifestyles (VanHook and Patel, 2008). Most crustacean species have separate male and female sexes, but hermaphroditism and parthenogenesis also sometimes exist. Maternal care is uncommon, but the young may live on the mother's body for a few days after hatching in certain species. A unique group of crustaceans, the pistol shrimp of the genus *Synalpheus* that dwell in sponges, exhibit a eusocial lifestyle like bees or ants with a single reproductive female (queen) and the remaining individuals working to maintain the colony (VanHook and Patel, 2008).

The relationships between major classes of crustaceans remains a topic of debate, as well as the relationship between crustaceans and hexapods, and their phylogenetic tree is yet to be fully resolved. The Pancrustacea hypothesis suggests that crustaceans and hexapods (insects) are sister groups (Zrzavý and Štys, 1997). However, advancements in genomics and developmental biology are increasing our understanding of the evolutionary history of crustaceans (Regier et al., 2010; von Reumont et al., 2012). Non-model organisms are gaining popularity in research, and genome information is becoming increasingly important as molecular biology expands its focus beyond model organisms. With recent advances, it has been proposed that crustacean is paraphyletic, with the hexapods nested within a larger pancrustacean clade (Regier et al., 2010; von Reumont et al., 2012). Pancrustaceans would then encompass crustacean and hexapod species (Figure 1-1). Pancrustaceans can be then classified into ten different classes including Malacostraca where Decapoda can be found (Schwentner et al., 2017).

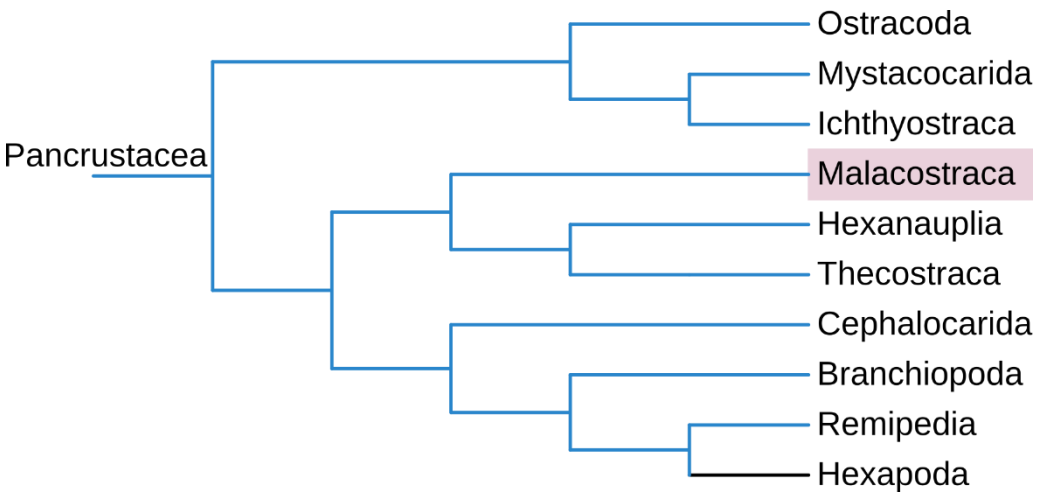

Figure 1-1: Pancrustacea phylogenetic tree. Blue branch: Crustacea, black branch: Hexapoda, Purple: phylogenetical position of crayfish in the tree.

1.1.2 Exploring the world of Decapoda

The order of decapods is remarkably diverse, comprising over 15 000 living species that includes prawns, shrimps, lobsters, crayfish, and crabs (De Grave et al., 2009, 2023). Decapods diverged from the other crustaceans 455 million years ago, during the late Ordovician period (Schram et al., 1978; Porter et al., 2005; Bracken Grissom et al., 2009; Schram, 2009). They are distinguished by ten prominent walking legs, the first pair of legs often developed as claws (Mariappan et al., 2000). However, the actual number of appendages is significantly greater and encompasses antennae, mouthparts, and abdominal appendages (Waltling and Thiel, 2013). Each set of walking appendages originates from an individual body segment, yet these segments converge into a unified unit in the anterior section of the body known as the cephalothorax (Figure 1-2). The body of a crayfish is divided into two primary tagmata along its central axis: the cephalothorax and the abdomen (Poore, 2004). The cephalothorax is formed by the fusion of the head and the thoracic segment. On the contrary to crustaceans that have a chitinous exoskeleton, the cephalothorax in Decapoda is covered by a compact hard calcium carbonate shield (VanHook and Patel, 2008). The five pairs of large legs, called pereopods, are also labelled as "walking legs," although the first pair of pereopods is also known as chelipeds or first chelipeds and is often the most massive (Figure 1-2; Mariappan et al., 2000; Creed, 2009). These claws at the end of the chelipeds (present on only walking legs 1, and 2 in shrimps) aid in walking by providing support against the substrate. Claws are strong, often with teeth and a sharp hooked end spine, and are also used for grasping and

manipulating objects. The second and third pairs of walking legs can also have claws, but narrower and weaker compared to the first pair (except for shrimps where the second pair is the most massive). These legs are involved in walking, picking up small food items and detritus, and helping the crayfish climb surfaces and vegetation. The abdomen is segmented and of similar size to the cephalothorax. It consists of six segments connected by movable joints and a soft membrane. The abdomen has protective plates on the upper and lateral sides, while the lower side is soft and can present pleopods used for swimming and carrying the eggs (Figure 1-2).

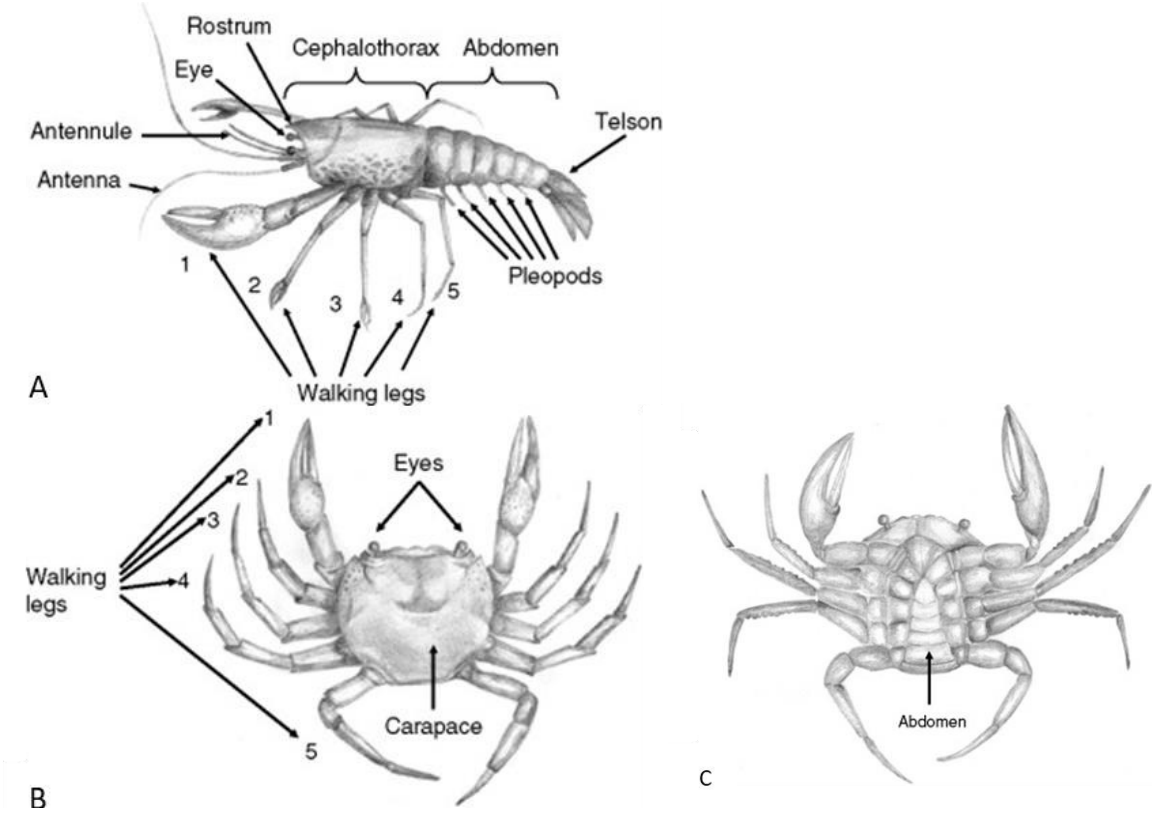

Figure 1-2: Morphology of generalised crustaceans. A. Major body regions and appendages of a generalised crayfish. B. Major body regions of a generalised crab. C. Ventral view of crab showing abdomen. Drawings by Robert Creed (Creed, 2009).

While many decapod species are known for their scavenging behaviour, they exhibit a range of feeding habits including omnivory, herbivory, carnivory and detritivory (Momot, 1995; Briones-Fourzán and Hendrickx, 2022). This diversity in feeding strategies allows decapods to be found in a wide range of habitats. They can be found in marine, semiterrestrial, and freshwater environments (De Grave et al., 2009). Marine decapods, such as crabs and lobsters, are well-known inhabitants of coastal regions and oceanic depths (Young and Elliott, 2020). They have evolved specific adaptations to thrive in saltwater environments, including the

ability to regulate their body's osmotic balance (Henry et al., 2012). Semiterrestrial decapods, like land crabs, differ from other decapods in their ability to transition between marine and terrestrial habitats. These species often migrate between land and sea for various reasons, such as mating, feeding, and avoiding predation (Watson-Zink, 2021). They have adaptations that allow them to withstand desiccation and survive in the challenging conditions of intertidal zones. Freshwater decapods, including crayfish and freshwater crabs, have successfully colonized rivers, streams, lakes, and other freshwater bodies (Kawai and Cumberlidge, 2016). They have adapted to the specific challenges of freshwater habitats, such as fluctuations in water quality, temperature, and oxygen levels (Anger, 2016).

The order of decapods can be further classified into two suborders (Wolfe et al., 2019): Dendrobranchiata, commonly known as prawns, and Pleocyemata (Figure 1-3). The latter encompasses a group formed by Stenopodidea (boxer shrimp), Caridea (swimming shrimps) and Procarididea, and a crawling/walking clade called Reptantia. Reptantia can be divided into two distinct groups. The first one comprises Achelata (spiny lobsters), Astacidea (true lobsters and crayfish), Polychelida (benthic crustaceans), and Glypheidea with only two remaining living species. The second group encompasses Axiidea (mud shrimp, ghost shrimp, or burrowing shrimp), Gebiidea (mud lobsters and mud shrimp), and the Meiura clade formed by Anomura (hermit crabs), and Brachyura (short-tailed crabs).

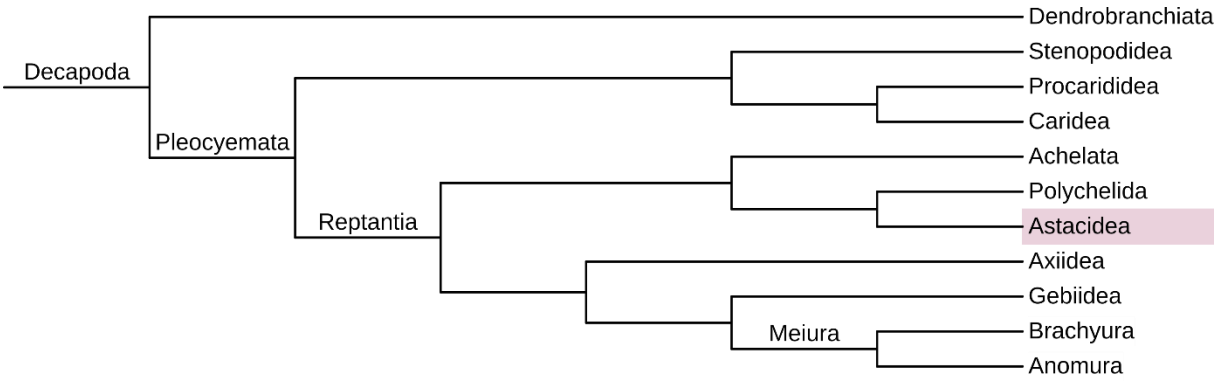

Figure 1-3: Decapoda phylogenetic tree. Purple: phylogenetic position of crayfish in the tree.

1.1.3 Geographical distribution and phylogeny of crayfish

Freshwater crayfish include over 650 identified species distributed worldwide (Crandall and De Grave, 2017). The origin of crayfish can be traced back to the end of the Permian period, approximately 250 million years ago, or even deeper into the Permian period (Breinholt et al., 2009; Porter et al., 2005). Along with lobsters, they all belong to the infraorder Astacidea. The common ancestor of crayfish and lobsters is thought to have inhabited the waters of the Paleo-Tethys Ocean by the end of the Palaeozoic era (K A Crandall et al., 2000; Sinclair et al., 2004). The Astacidea infraorder is classified into four superfamilies: Nephropoidea, Enoplometopoidea, Astacoidea (families Astacidae and Cambaridae), and Parastacoidea (Parastacidae) (Figure 1-4; De Grave et al., 2009). While the two first superfamilies encompass sea lobsters, the remaining two correspond to freshwater crayfish. However, evidence supports the monophyletic origin of crayfish, indicating that Astacoidea and Parastacoidea may constitute a single superfamily (K A Crandall et al., 2000; Cukerzis, 1987). This superfamily would thus comprise three families: Astacidae, mainly comprising European crayfish; Cambaridae, predominantly representing North American crayfish; and Parastacidae, from the Southern hemisphere. Nevertheless, two geographically isolated genera challenge the above-mentioned taxonomic division of higher crayfish: *Cambaroides* (Cambaridae) from the far east of Asia and *Pacifastacus* (Astacidae) from western North America do not align with the diversity centres of their respective genera, eastern North America and Europe, respectively (Figure 1-5; Kozák et al., 2015). Furthermore, *Cambaroides* and *Pacifastacus* phylogenetic relationship with other crayfish in the Northern hemisphere remains unclear. However, a recent study defines the Cambaroididae family as a new sister family from Astacoidea that includes *Cambaroides* species from east Asia that would have diverged early from other Astacoidea species (Audo et al., 2023).

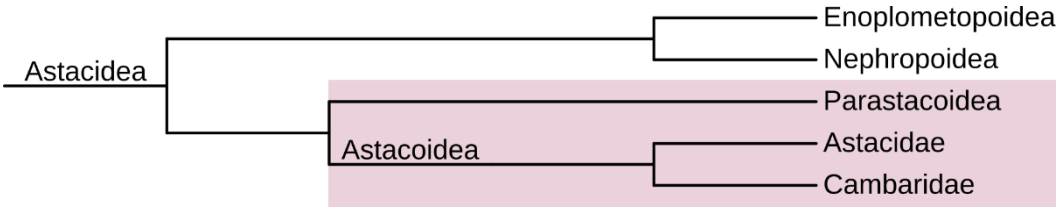

Figure 1-4: Astacidea phylogenetic tree. Purple: phylogenetic position of crayfish in the tree.

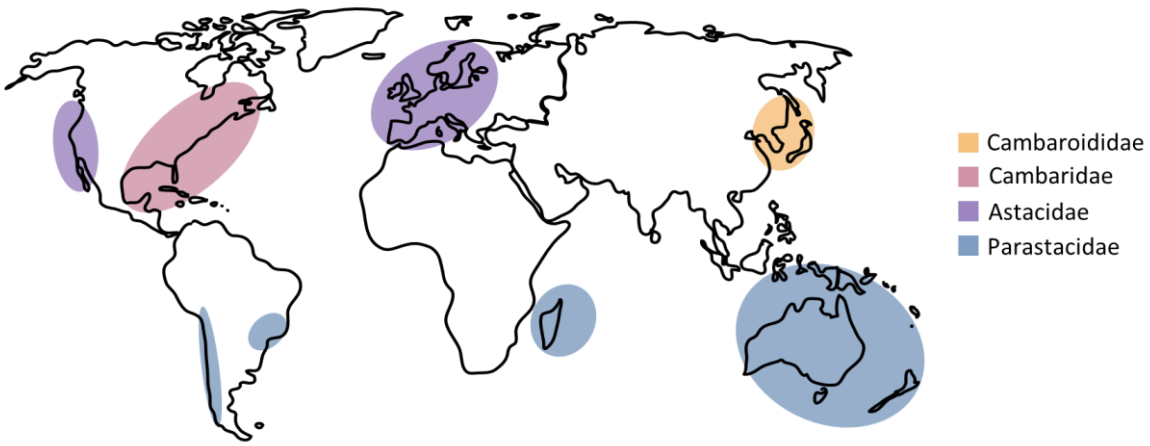

Figure 1-5: Scheme of the world distribution of crayfish. Cambaroididae correspond to Cambaroides species from far east Asia that are for the moment generally affiliated to Cambaridae species from east North America but recently described as a new family. Modified figure from Kozák et al., 2015.

Crayfish species from the Astacidae, Cambaridae and Cambaroididae families can be found in the Northern hemisphere in Europe, North America and Asia, while in the Southern hemisphere, Parastacidae species are found in South America, Africa and Australia (Figure 1-5; Kozák et al., 2015; Souty-Grosset et al., 2006). In Europe, native species belong exclusively to the Astacidae family. Astacidea encompasses 31 species, distributed within Europe and North America, in the genus *Astacus*, *Pontastacus* (still often considered a member of *Astacus*), *Austropotamobius* and *Atlantoastacus* (still often considered a member of *Austropotamobius*). However, of the 31 Astacidea species, only five species are commonly accepted to be present in Europe, one from the genus *Astacus*, two from the genus *Pontastacus*, and two from the genus *Austropotamobius* (Kozák et al., 2015; Souty-Grosset et al., 2006). While the noble crayfish, *Astacus astacus* (Linnaeus, 1758), is the most widespread and emblematic species in Europe, some subspecies are only endemic to a small region. The newly described idle crayfish, *Austropotamobius biharensis* (Pârvulescu, 2019), deriving from the species complex of the stone crayfish, *Austropotamobius torrentium* (von Paula Schrank, 1803), is a great example of an endemic species, as it is restricted to the rivers in the Apuseni Mountains in Romania (Pârvulescu, 2019).

North America has a rich diversity of crayfish, divided into Cambaridae and Astacidae families. Cambaridae is the dominating family, representing 70% of global freshwater crayfish and over

95% of Northern Hemisphere species (Crandall and Buhay, 2008; De Grave et al., 2009). Estimates suggest up to 422 species in North America, with *Procambarus* being the most species-rich genus, and *Pacifastacus* being exclusive to the western coast. Notable species, including the red swamp crayfish, *Procambarus clarkii* (Girard, 1852), and the signal crayfish, *Pasifastacus leniusculus* (Dana, 1852) became invasive species in various regions (Maciaszek et al., 2022; Souty-Grosset et al., 2016). In Asia, there are two distinct groups of crayfish, one in the east and the other in the west, with a large "crayfish-free" area across the continent (Kozák et al., 2015). East Asian crayfish belong to the independent subfamily Cambaroidinae within Cambaridae. The genus *Cambaroides*, also identified as Cambaroididae (Audo et al., 2023), is the oldest living crayfish lineage in the Northern Hemisphere, with simpler structures suggesting an ancient appearance (Braband et al., 2006). While four *Cambaroides* species are widely accepted, discussions continue about additional species. In western Asia however, crayfish belonging to the family Astacidae is likely the result of a historically late spread of European species into Asia (Kozák et al., 2015).

Crayfish from the Southern hemisphere belong to the Parastacoidea superfamily, specifically the Parastacidae family with 12 described crayfish species in the genera *Parastacus*, *Samastacus*, and *Firilastaces* (Keith A. Crandall et al., 2000; Hobbs, 1989; Holthuis, 1952). In South America eight *Parastacus* species (Hobbs, 1989) are known. While native crayfish species can't be found in the African continent itself, seven *Astacoides* species live on the island of Madagascar (Boyko et al., 2005; Hobbs, 1987). Their distribution is limited to the Southern half of the island, and they can be found at high altitudes, up to 2,000 meters above sea level. Recently, the parthenogenetic marbled crayfish, *Procambarus virginalis* (Lyko, 2017), has been introduced to Madagascar, posing a potential threat to the native species (Jones et al., 2008; Andriantsoa et al., 2019). In mainland Africa, red swamp crayfish (*P. clarkii*), noble crayfish (*A. astacus*), and spiny-cheek (*Faxonius limosus*, Rafinesque, 1817) crayfish were introduced in the late 20^{e} century (Hobbs, 1989). Australia and its adjacent islands are a major diversity centre for crayfish. The family Parastacidae is well-represented with 49 known species from the genus *Euastacus* and a few from the other genus of the family Parastacidae (De Grave et al., 2009). The Tasmanian giant crayfish, *Astacopsis gouldi* (Clark, 1936), is the largest crayfish species in the world (Hamr, 1992). New Zealand is home to the genus

Paranephrops with two species, while Papua New Guinea has the genus *Cherax* with eight species (Hobbs, 1974; Holthuis, 1982).

1.1.4 Importance of crayfish in aquatic ecosystems

Crayfish exhibit adaptability to a range of habitats, from flowing waters (lotic) to standing waters (lentic), but their existence relies on specific features of the aquatic ecosystem, such as water quality and absence of pollutants (Kozák et al., 2015; Reynolds et al., 2013; Souty-Grosset et al., 2006). Crayfish are indicator species for high water quality as they flourish in pristine biotopes. Some species exhibit tolerance to moderate water quality, which can suffice their survival. Crayfish display also diverse feeding behaviours and diets, as studied across various species (Kozák et al., 2015; Reynolds et al., 2013; Souty-Grosset et al., 2006). Factors like benthic invertebrate biomass and diversity shape their niche width and access to animal food sources. Crayfish exhibit opportunistic feeding, adjusting their consumption to available resources (Reynolds, 2011). They contribute to decomposition and recycling by breaking down dead plant material and enhancing water cleanliness by preying on vulnerable organisms and carcasses (Usio and Townsend, 2004). Their predation regulates populations of specific animals, particularly gastropods, and controls the population of habitat-dependent organisms by consuming plants (Dorn and Wojdak, 2004; Statzner et al., 2003). Broadly omnivorous crayfish are more prevalent than specialized feeders, significantly impacting ecosystem structure and function through selective consumption of vegetation and invertebrates (Reynolds et al., 2013). Moreover, interactions between crayfish and other organisms encompass predation on fish, affecting fish communities and salmonid recruitment, while also serving as a vital food resource within the food web.

Due to these ecological contributions, crayfish take the role of keystone species, playing an essential role in defining and maintaining the balance of the entire ecosystem. A keystone species holds unparalleled importance, as its presence is crucial for the ecosystem to exist in its current state (Cottee-Jones and Whittaker, 2012; Paine, 1995). If keystone species were to be removed, the ecosystem would undergo significant changes or even face extinction. Keystone species have low functional redundancy, meaning that if they were to disappear, no other species could effectively take their place in the ecosystem. Consequently, the ecosystem would undergo radical transformations, potentially allowing new and invasive species to

occupy the habitat. Keystone species can encompass a wide range of organisms, including microbes and plants, and their importance is not necessarily correlated with their size or abundance within an ecosystem. However, the majority of keystone species are animals that exert a significant influence on food webs, with variable impact depending on habitats.

Beyond their role as keystone species, crayfish have the role of ecosystem engineers, shaping and influencing the surrounding environment (Creed and Reed, 2004; Statzner et al., 2003). They achieve this by modifying and creating habitats through their biological activities or by physically altering the biotic and abiotic factors in their environment. More specifically, crayfish belong to allogenic engineers, which physically change their environment from one state to another. Crayfish are known for their burrowing activities, which involve excavating and creating complex underground structures within the substrate of rivers, streams, and lakes (Kozák et al., 2015; Reynolds et al., 2013; Souty-Grosset et al., 2006). These burrows serve multiple functions. They provide shelter and protection for crayfish themselves, as well as for other aquatic organisms seeking refuge from predators. Additionally, these burrows can improve water filtration and nutrient cycling within the ecosystem by creating pathways for water flow and increasing sediment turnover. The construction of burrows by crayfish can also influence the sediment composition and distribution, impacting the physical characteristics of the habitat (Statzner et al., 2003). By selectively choosing and moving specific types of substrates while digging, crayfish can affect the availability of various microhabitats for other organisms, influencing the diversity and abundance of species within the ecosystem. This intricate ecosystem engineering also mirrors the transformative power exhibited by invasive species, which can drastically modify existing environments without natural constraints, ultimately hindering the growth of native ecosystems (Nishijima et al., 2017).

1.1.5 Commercial crayfish exploitation

In addition to their vital importance to the ecosystem, crayfish have also significant commercial value and are exploited for various purposes, primarily driven by human demand and economic opportunities (Kouba et al., 2014; McClain, 2020; Japo Jussila et al., 2021). The commercial purpose includes aquaculture, where crayfish are cultivated in controlled environments such as ponds or tanks, to meet the growing demand for crayfish products on the market. Aquaculture allows for efficient production, enabling farmers to control the

crayfish population, optimize growth conditions, and harvest them at specific sizes for different market segments. Crayfish are widely consumed as a delicacy in many parts of the world. They are an important source of animal protein and are sought after for their unique flavour and texture. The food industry counts on crayfish as a valuable food product. Crayfish are also popular bait for recreational fishing. Many anglers use live crayfish as bait to attract predatory fish species, such as bass and catfish. Many species of crayfish are also sought after in the pet trade with colourful crayfish species being popular among aquarium hobbyists.

The introduction of non-native crayfish species for commercial purposes can also lead to the displacement of native crayfish populations (Bláha et al., 2022). Invasive crayfish species can outcompete and even prey upon native crayfish, leading to a loss of biodiversity and ecological disruption (Holdich et al., 2009). Aquaculture practices and trade can facilitate the spread of diseases and parasites among different crayfish populations by overexploitation (Longshaw, 2011). Moreover, non-native species can carry new diseases that can spread to wild populations by contaminated water and have devastating effects (Holdich et al., 2009).

To mitigate these implications, sustainable management practices, strict regulations, and public awareness about responsible crayfish exploitation are crucial. Conservation efforts and responsible aquaculture practices can ensure the long-term viability of crayfish populations while meeting the demands of the commercial sector.

1.2 European crayfish: threatened species

The Crayfish Tale – A short & entertaining educational film

(Maguire et al., 2022)

1.2.1 Biogeography of native European species

Regarding the biogeographic origin of crayfish, Europe could have been either an area of original distribution, or their “old home” into which they returned from Asia during the

Jurassic and Cretaceous periods (Kozák et al., 2015). The Eurasia continent was at that time represented by isolated or temporarily merged blocks, which could have influenced the distribution of crayfish. In Europe, we can find one native species from the genus *Astacus*, two from the genus *Pontastacus*, two from the genus *Autropotamobius* (Figure 1-6; Souty-Grosset et al., 2006; Holdich et al., 2009; Kozák et al., 2015).

Among native European crayfish, the noble crayfish (*A. astacus*), is the most widespread (Figure 1-6 A). It is among the longest-living crayfish freshwater invertebrates with a lifetime of more than 20 years (Table 1-1). Noble crayfish are found in open waters in 39 territories of Europe (Kozák et al., 2015). They prefer habitats with vegetation-covered banks and stable substrates, such as alder and willow trees, for shelter. A shelter is mostly represented by a simple shallow burrow. They avoid muddy bottoms as their habitat, but they use such places for foraging (Souty-Grosset et al., 2006).

The narrow-clawed crayfish, *Pontastacus leptodactylus* (Eschscholtz, 1823), is a complex species with more than ten different forms described as distinct species, though not universally acknowledged (Figure 1-6 B; Kozák et al., 2015). It is a large crayfish, with males reaching up to 30 cm in length, but more generally 15 cm (Table 1-1). It can be found in deeper and shallow lakes, smaller brooks, ponds, quarries, and river pools (Souty-Grosset et al., 2006). The narrow-clawed crayfish is more tolerant to organic pollution and lower concentrations of dissolved oxygen than the noble crayfish and is also well adapted to elevate salinity and tolerate turbid and muddy habitats (Souty-Grosset et al., 2006).

The thick-clawed crayfish, *Astacus pachypus* (Rathke, 1837), is one of the range-restricted native crayfish species in Europe (Figure 1-6 C). It inhabits both fresh and brackish waters, preferring rocky substrates with macrophytes and a sufficient food supply (Souty-Grosset et al., 2006). There are indications of possible hybridization with narrow-clawed crayfish (*P. leptodactylus*) (Souty-Grosset et al., 2006).

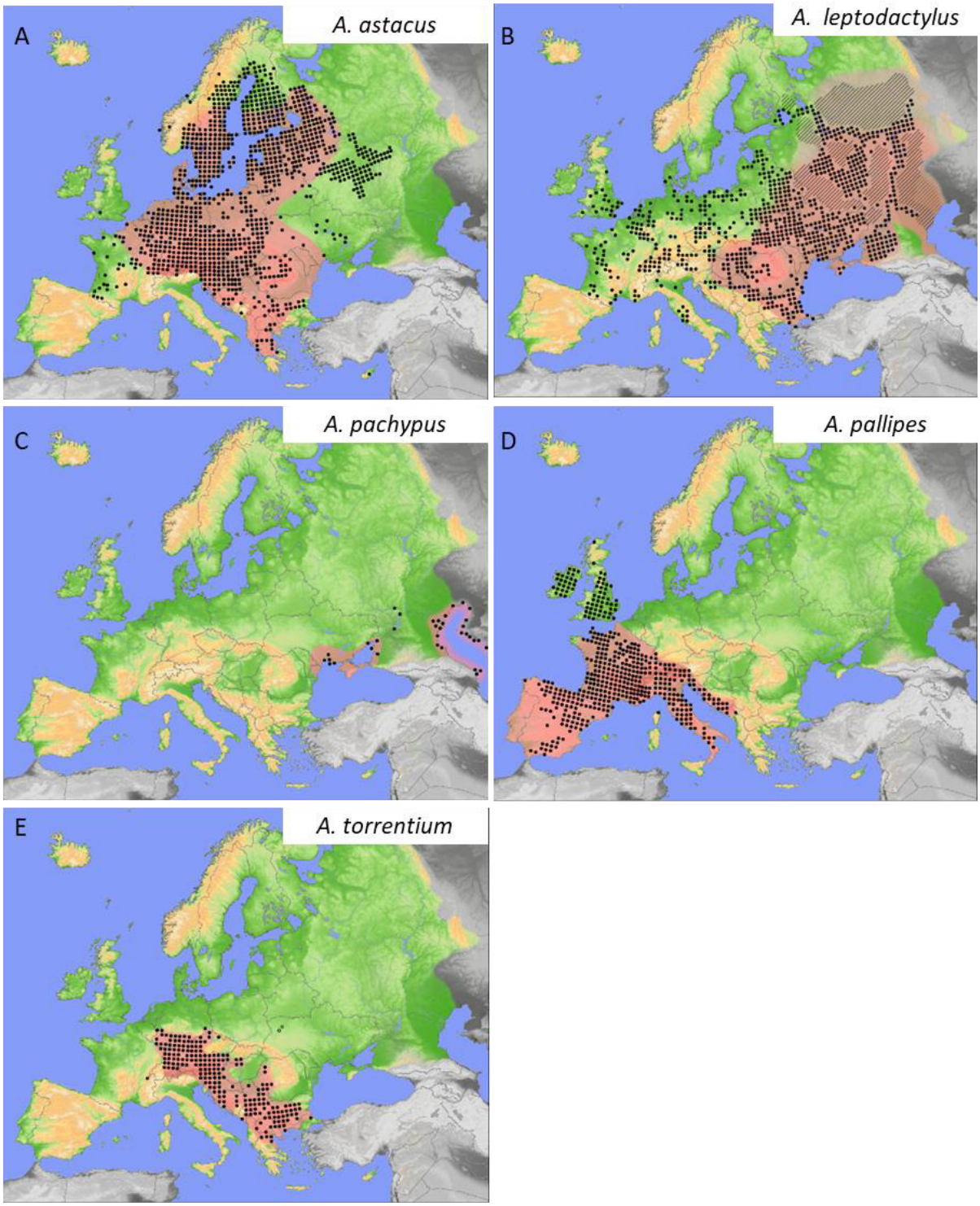

Figure 1-6: Native crayfish species wild distribution in Europe. Presumed native range is highlighted in red. The confirmed occurrence is presented in the common European Chorological Grid Reference System (CGRS, i.e., approx. 50 X 50 km grid). The hatched area covers regions where the species is considered widespread but information about specific localities is missing. A. Noble crayfish, *Astacus astacus*. B. Narrow-clawed crayfish, *(Pont)Astacus leptodactylus*. C. Thick-clawed crayfish, *(Pont)Astacus pachypus*. D. White-clawed crayfish, *Austropotamobius pallipes*. E. Stone crayfish, *Austropotamobius torrentium*. Figures from Kouba et al., 2014.

Table 1-1: Body size and lifetime of native European crayfish.

	largest size (cm)	lifetime (years)
Noble crayfish	< 15	> 20
Narrow clawed crayfish	< 30	> 10
Thick-clawed crayfish	10 to 12	NA
White-clawed crayfish	< 12	> 10
Stone crayfish	8 to 10	> 10

The white-clawed crayfish, *Austropotamobius pallipes* (Lereboullet, 1858), is considered a species complex composed of two species (Figure 1-6 D; Kozák et al., 2015). Its distribution is partly due to human-mediated introductions (Souty-Grosset et al., 2006; Kouba et al., 2014). It can be found in a wide range of habitats, including fast-flowing brooks, small rivers, slow-flowing rivers, lakes, ponds, and flooded quarries (Souty-Grosset et al., 2006). It prefers rocky substrates with shelters, but it can tolerate areas with muddy bottoms for feeding. The species is relatively tolerant to various environmental conditions, such as low dissolved oxygen, eutrophication, and varying salinity (Souty-Grosset et al., 2006).

The stone crayfish, *A. torrentium*, is a species complex, with three commonly mentioned subspecies (*A. torrentium dalmatinus*, *danubicus*, and *macedonis*) (Figure 1-6 E; Kozák et al., 2015). It is found in brooks and small rivers in mountain and submontane regions, usually with a rocky substrate and woody riparian vegetation (Souty-Grosset et al., 2006). Stone crayfish can tolerate relatively fast currents. Oxygen demands of stone crayfish are higher than those of the noble crayfish, however, the species can resist relatively severe organic water pollution (Souty-Grosset et al., 2006). An endemic population derived from the stone crayfish, *A. biharensis*, also called idle crayfish, was recently described in Apuseni Mountains in Romania (Pârvulescu, 2019).

1.2.2 Challenges for European crayfish species

The fate of European freshwater crayfish is challenged by the introduction of non-native crayfish species (Table 1-2, Figure 1-7; Japo Jussila et al., 2021; Kozák et al., 2015; Souty-Grosset et al., 2006). A non-native species added to an ecosystem is considered invasive if it

is established in the habitat and causes economic or environmental harm (Jeschke et al., 2014). Invasive species generally exhibit a more plastic lifestyle, higher adaptability, and faster reproductive rates, allowing them to compete effectively with native species (Kozák et al., 2015). Their aggressive behaviour and higher activity levels give them an advantage in competing for resources and hiding places. Invasive crayfish also have migration abilities that allow them to quickly colonize new habitats, deplete resources and force native species out.

Human influence on the environment and the introduction of non-native species have been identified as the most significant factors supporting the distribution of non-native species (Japo Jussila et al., 2021). More than the introduction of non-native species, human activities such as land-use change, pollution, stocking non-native species, and altering water management, further facilitate the spread of non-native crayfish species threatening European crayfish species.

Crayfish can also host a variety of parasites and pathogens, while also participating in commensal or mutualistic relationships with certain organisms (Kozák et al., 2015). Some organisms are exclusively associated with crayfish and rely on them as essential hosts. These relationships can have varied effects on crayfish, depending on environmental conditions and their immune status. Parasites of crayfish include epibionts, such as zebra mussels, bryozoans, protozoans, and insect larvae, which use crayfish as a surface for attachment and food filtering. Additionally, crayfish can be affected by viruses, bacteria, fungi, microsporidia, and oomycetes. Viral infections can cause significant mortality, while bacteria and fungi may lead to lesions or deteriorated health, particularly when crayfish are weakened or injured. One of the most destructive diseases of European crayfish is the crayfish plague, caused by the oomycete *Aphanomyces astaci*, carried by North American crayfish species imported into Europe.

Table 1-2: Non-native crayfish species in Europe.

Common name	Scientific name	Invasive	date of first introduction	purpose of introduction	Corresponding Figure	Problems caused by its presence
Signal crayfish	*Pacifastacus leniusculus*	Yes	1950-1960	Commercial	Figure 1-7 A	Carrier of crayfish plague. Extensive burrowing activities causing considerable damage to rivers and lake margin. High aggressivity and adaptability.
Red swamp crayfish	*Procambarus clarkii*	Yes	1973	Bait and Pet	Figure 1-7 B	Burrowing activities impact agriculture by damaging water plants and riverbanks. Causes important changes in food webs structures. Accumulate heavy metals and pollutants in body and transmits them to higher trophic level. Carrier of crayfish plague.
Yabby crayfish	*Cherax destructor*	Potentially	1983	Pet and commercial	Figure 1-7 D	Rapid reproduction. Extensive burrowing activity. High adaptability.
Redclaw crayfish	*Cherax quadricarinatus*	Potentially	NA	Pet and commercial	Figure 1-7 D	Rapid growth.
Calico crayfish	*Faxonius immunis/ Orconectes immunis*	Potentially	1997	Pet	Figure 1-7 D	Possible carrier of crayfish plague. Burrowing activities impact agriculture by damaging water plants and riverbanks.
Kentucky river crayfish	*Faxonius juvelinis/ Orconectes juvelinis*	Potentially	NA	NA	Figure 1-7 D	Rapid reproduction and growth. High adaptability. Carrier of crayfish plague.
Spiny-cheek crayfish	*Faxonius limosus/ Orconectes limosus*	Yes	1890	Bait and Pet	Figure 1-7 C	Carrier of crayfish plague. High adaptability. Displace native species.
Virile crayfish	*Faxonius virilis/ Orconectes virilis*	No	1897	Reintroduction	Figure 1-7 D	Burrowing activities impact riverbanks. High aggressivity and adaptability. Causes important changes in food web structures.
White River crayfish	*Procambarus cf. acutus*	Potentially	1973	NA	Figure 1-7 E	Adaptability to cold temperature. Rapid growth. High aggressivity.
Florida crayfish	*Procambarus alieni*	Potentially	1973	Pet	Figure 1-7 E	High reproductive rate and adaptability.
Marbled crayfish	*Procambarus fallax f. virginalis*	Yes	1990s	Pet	Figure 1-7 E	Clonally reproductive and so exhibits high reproductive potential. Causes important changes in food web structures. Carrier of crayfish plague.

(Kouba et al., 2014; Kozák et al., 2015; Souty-Grosset et al., 2006)

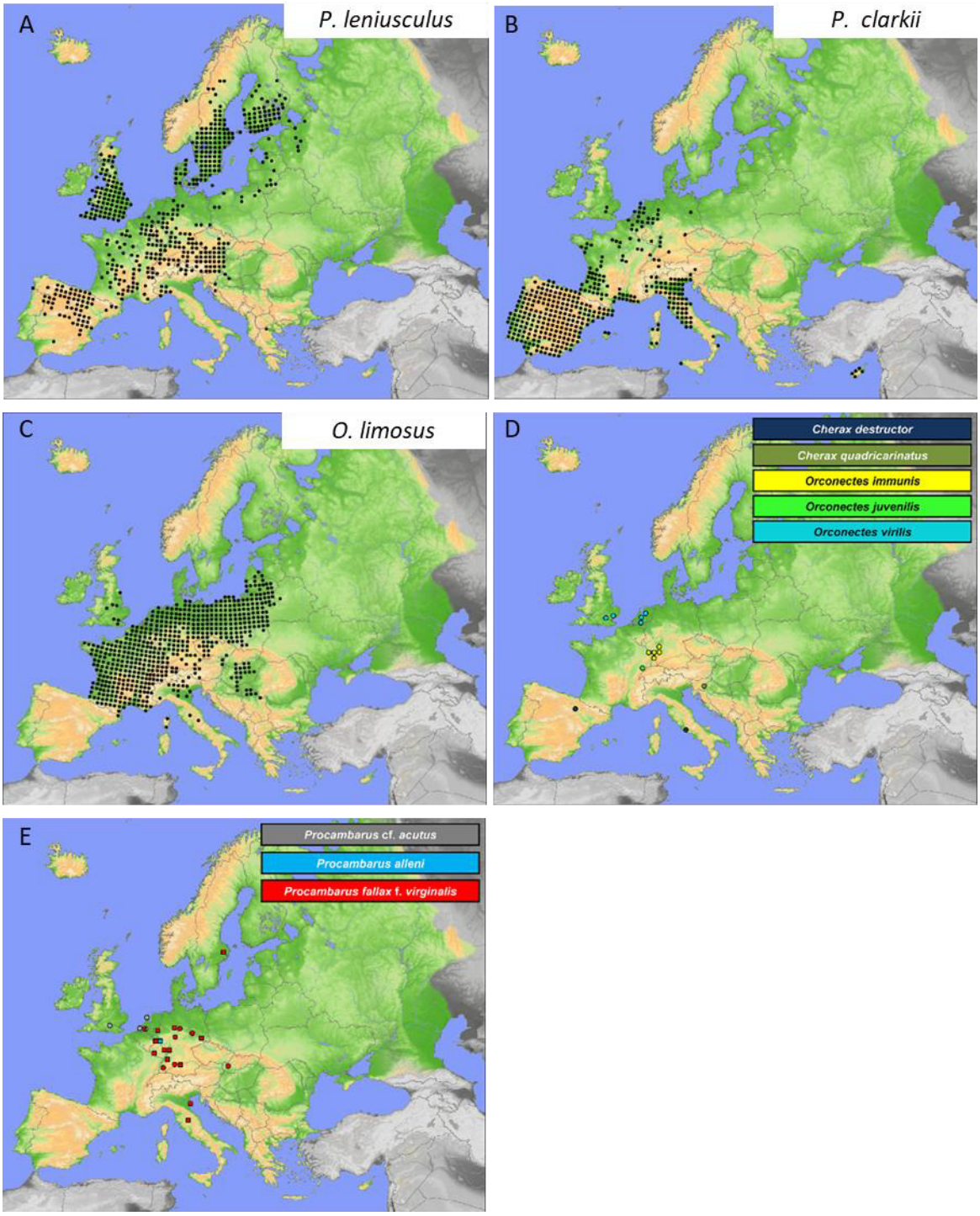

Figure 1-7: Non-native crayfish species wild distribution in Europe. Figures from Kouba et al., 2014.

1.2.3 The crayfish plague

The crayfish plague disease, with its devastating mass mortalities of native crayfish species, dates back to the mid-19th century and *Ap. astaci* is ranked among the world's top 100 invasive organisms (Lowe et al., 2004). While primarily afflicting crayfish, recent research indicates that its parasitic impact extends to freshwater crabs and shrimps (Schrimpf, 2014; Svoboda et al., 2014a, 2014b). This pathogen, responsible for substantial crayfish population losses, is believed to be spread by the introduction of chronically infected North American crayfish species (Martín-Torrijos et al., 2021). The spread of the crayfish plague epizootics closely followed the movements of the commercial crayfish trade. *Aphanomyces astaci*, the causative agent of the disease, was identified in 1903, though its original introduction to Europe remains unclear (Alderman, 1996). The disease's menace persisted through subsequent introductions of North American crayfish species. *Aphanomyces astaci* belongs to the Oomycota phylum (clade Stramenopiles), sharing characteristics with true fungi and thriving in substrates to extract nutrients (Söderhäll and Cerenius, 1999). These organisms grow as aseptate hyphae and produce zoospores, with crayfish cuticles providing a conducive environment for *Ap. astaci* growth. While certain *Aphanomyces* species engage in sexual reproduction, others, like *Ap. astaci*, tend to rely on asexual reproduction facilitated by swimming zoospores released from sporangia on crayfish bodies, ensuring efficient transmission among aquatic creatures (Cerenius et al., 1988). These zoospores are short-lived, surviving only in wet conditions and serving as the sole infectious stage of the pathogen (Johnson et al., 2002). Upon attachment to the host's cuticle, the spore triggers germination and hyphal penetration, with the host's early immune response potentially influencing the establishment of the pathogen (Souty-Grosset et al., 2006). The spore then penetrates the host's tissues, acquiring nutrients. The crayfish's immune system activates the prophenoloxidase activating system upon infection or injury, prompting melanin production (Söderhäll and Cerenius, 1999). While North American crayfish combat the pathogen by encapsulating it with melanin, European species struggle to reach comparable immune levels, underscoring the intricate interplay between the pathogen and the host's immune response (Figure 1-8; Cerenius et al., 2003).

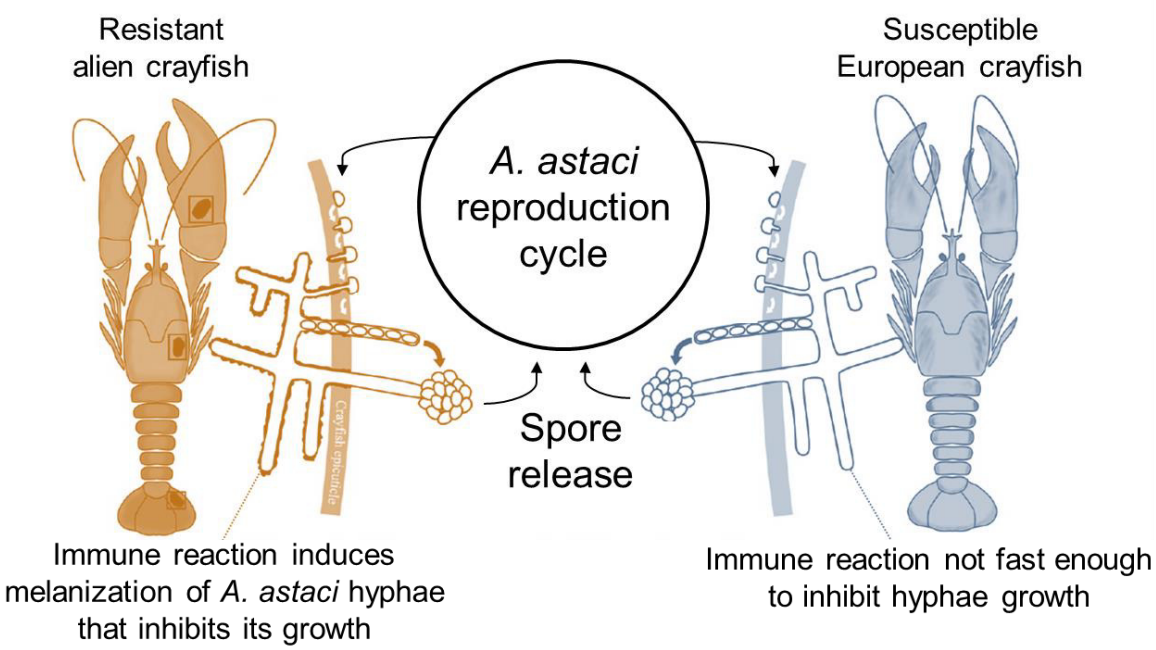

Figure 1-8: *Aphanomyces astaci* in resistant and susceptible species. In North American crayfish species, the pathogen is often found in the melanized areas of the cuticle whereas in European and other susceptible species, melanisation does not usually occur. Adapted from Rezinciuc et al., 2015.

Aphanomyces astaci can be categorized into five distinct haplogroups based on random amplified polymorphic DNA (RAPD-PCR) (Table 1-3; Huang et al., 1994; Diéguez-Uribeondo et al., 1995; Kozubíková et al., 2011). Haplogroups present variable virulence, with haplogroup A being considered lowly virulent and haplogroup B as highly virulent. Based on mitochondrial markers, haplogroup C can be grouped with haplogroup A and Haplogroup D can be divided into two groups (D1 and D2) (Makkonen et al., 2018).

Table 1-3: *Aphanomyces astaci* haplogroups and original host.

Haplogroup	Original host
A	unknown
B	*P. leniusculus*
C	*P. leniusculus*
D	*P. clarkii*
E	*F. limosus*

Despite ongoing efforts, effective treatments for crayfish plague remain elusive, and eradicating the pathogen from North American crayfish populations presents a formidable challenge. The ramifications of the crayfish plague are dire, causing widespread mortalities in affected regions and threatening indigenous crayfish populations across Europe (Alderman, 1996; Japo Jussila et al., 2021). Comprehension of the disease's transmission, virulence and variability is pivotal to the formulation of robust management strategies aimed at safeguarding crayfish populations from this catastrophic affliction.

1.2.4 Endangered European populations

European crayfish fauna has been devastated by habitat degradation, water pollution, and outbreaks of crayfish plague for the past 150 years. To economically compensate for this decline, alien North American crayfish, such as signal crayfish (*P. leniusculus*), have been deliberately released into European water courses destabilising the environment and leading to significant transformations within aquatic ecosystems. Due to inadequate population management practices across Europe, the mass mortality of native crayfish has intensified over the years. Nowadays, native European noble crayfish (*A. astacus*) and white-clawed crayfish (*A. pallipes*), face significant conservation concerns, leading to their respective categorisations as vulnerable and endangered species within the IUCN Red List Index of Threatened Species (Edsman et al., 2010; Füreder et al., 2010).

The noble crayfish (*A. astacus*) is also listed as critically endangered in the 2010 Swedish Red List. Additionally, it is listed on the Danish Red List as "at risk from eradication due to crayfish plague" (Edsman et al., 2010). Over the past 150 years, the noble crayfish has experienced a global staggering population decline of over 95% (Skurdal and Taugbøl, 2002). In the past 22 years, Sweden and Norway have shown particularly high declines, reaching approximately 78% and 61%, respectively. Similar decline rates are evident in various other countries, with a global estimate of population loss ranging from 50% to 70%. However, in certain regions, such as Finland, successful re-stocking initiatives suggest a potentially lower actual decline rate of 40% to 50% (Japo Jussila et al., 2021). In France, noble crayfish populations are nearing extinction in Lorraine and Morvan, although restocking efforts have been attempted. This species is also present in other regions of France, including Alsace. In Germany, the noble crayfish, which was once widespread and abundant in northwest Germany, has seen a decline

in the range of 56% over 22 years. From 155 sites reported in 1920 (after the introduction of the crayfish plague), 12 sites remained in 1990 (Schulz, 2000).

The white-clawed crayfish, (*A. pallipes*) has been assessed as endangered, with a decline of 50% to 80% based on presence/absence data available for England, France and Italy in the last 10 years. Notably, in the South Tyrol region of Italy, the population decline is estimated to be as high as 99.5% over 10 years. It has been suggested that if the current trend in the decline of this species persists, it could face possible extinction in Britain within 30 years (Füreder et al., 2010). It is important to highlight that although this species may appear abundant in certain areas, the degree of genetic variability may actually be low. In various populations across Croatia, France, Italy, Spain, and Portugal, low intra-population genetic variability has been observed (Gouin et al., 2006; Bertocchi et al., 2008; Diéguez-Uribeondo et al., 2008). In some cases, such as in the basin of the River Sieve in Italy, the absence of heterozygotes and high levels of inbreeding have been noted (Bertocchi et al., 2008). In France, it is reported declining significantly with subpopulations that disappeared in 14 departments, and declines occurring in 26 of the 92 departments (Souty-Grosset and Reynolds, 2009; Vigneux, 1997). However, it is widespread in France, known to occur in most departments (Souty-Grosset et al., 2006; Füreder et al., 2010). Restocking initiatives have been carried out in waterways affected by the plague. Populations of white-clawed were first discovered in Germany in 1989 and its population numbers are currently declining (Füreder et al., 2010).

The critical situation of native European crayfish species underscores the urgency of implementing effective disease management and prevention strategies, as well as combating the proliferation of alien crayfish species. These efforts are essential to protect both the biodiversity and overall functionality of aquatic ecosystems, which are currently experiencing substantial crayfish population losses. It becomes increasingly clear that comprehending the behavioural and molecular differences between alien and native species as well as within native species is paramount. This situation shows the importance of adopting a conservative genomics approach in addressing these challenges.

1.2.5 Conservation genomics for crayfish

Biodiversity corresponds to all variability that can be found in the three commonly accepted levels of diversity: genetic diversity, species diversity and ecosystem diversity (Verma, 2017). Intraspecific genetic diversity, which refers to the genetic variation within a single species, is an important basis for adaptation to global changes. Within a species, it provides the basic substrate for evolution and is crucial for understanding how a species can adapt to changes in the environment. Biodiversity conservation is crucial for maintaining the stability and sustainability of ecological systems.

Historically, marine decapod species have received greater attention in scientific research and conservation efforts, owing to the vastness and ecological significance of our oceans. Their prominent role in marine ecosystems and economic importance has led to a considerable number of scientific studies and conservation initiatives. It is however crucial to not overlook the significance of freshwater decapod species such as crayfish. These often-neglected inhabitants of rivers, lakes, and streams play vital roles in maintaining freshwater ecosystems' health and balance. The conservation of freshwater crayfish is essential to preserve the overall biodiversity, ensure ecosystem stability, and sustain the countless communities that depend on freshwater resources. As we strive to protect and understand the delicate intricacies of our planet's aquatic life, dedicating efforts to studying and safeguarding both marine and freshwater decapods becomes indispensable in creating a sustainable future for all aquatic environments. Conserving crayfish biodiversity amidst the pressing global mass extinction crisis requires a strategic integration of genomics studies. Genomic insights hold immense potential for addressing specific challenges in crayfish conservation, including species identification, biodiversity monitoring, ecosystem protection, and restoring genetic diversity in endangered populations (Theissinger et al., 2023). Efforts are underway to standardize protocols for detecting genetic diversity and incorporating genomic knowledge into conservation planning.

Employing advanced high-throughput genomic sequencing technologies applications in biodiversity research and conservation efforts include for example DNA barcoding/metabarcoding, reduced representation DNA sequencing, gene expression analysis (RNA-Seq), epigenomics, and whole-genome sequencing (WGS) (Sarwat and Yamdagni, 2016;

Theissinger et al., 2023). Each of these applications offers distinct advantages and limitations, answering diverse aspects of crayfish conservation. DNA barcoding and metabarcoding are efficient for species identification and biodiversity monitoring. However, the use of mitochondrial DNA can lead to an overestimation of the sample divergence and intra-specific divergence can bias the interpretation, and rapid evolutionary rate of mtDNA (Akhan et al., 2014; Hurt et al., 2022; Lovrenčić et al., 2022, 2020). For example, in order to study the population of the noble crayfish (*A. astacus*) across Europe, microsatellite markers were used to demonstrate that North European populations have higher genetic variation than Central European highlighting the importance of using nearby sources of noble crayfish populations for restocking programs (Gross et al., 2013; Schrimpf et al., 2014, 2017). The reduced representation DNA sequencing provides genome-wide data for non-model species and can provide estimates for genetic diversity, inbreeding, and phylogenetic relationships, while only a small portion of the genome is analysed. Gene expression data, or transcriptomics, offer insights into functional variation and rapid responses to environmental changes (Du et al., 2016; Zhong et al., 2021; Zhou et al., 2022). For example, it allowed Du et al., 2016 to study the red swamp crayfish (*P. clarkii*) intestines challenged by White Spot Syndrome Virus (WSSV). They identified mechanisms involved in the anti-WSSV immune response in crayfish that can help solve viral disease problems in crayfish breeding. A disadvantage of transcriptomic analysis is the rapid degradation of RNA compared to DNA. Moreover, RNA is sensitive to tissue types, sex, age, and life stage variations, and only expressed genes at the very moment are analysed. This is why transcriptomes do not represent the entire pool of genes and regulatory elements present in an organism. WGS data, in contrast, offer unparalleled power in discovering the full gene content for analysing various evolutionary processes. This has been exemplified by the assembly of the clonal marbled crayfish, *P. virginalis*, that allowed to identify that the third copy of the genome resulted from an autopolyploid gamete during mating of two deceitful crayfish (*Procambarus fallax,* Hagen, 1870) (Gutekunst et al., 2018). The clonal reproduction of the marbled crayfish led to a genetically uniform and close resemblance to the original stock founded in Germany in 1995. More generally, WGS is crucial for many genomic approaches to achieve accurate results. Genome assemblies provide a crucial framework for understanding and protecting biodiversity (See 2.3).

By integrating these genomic tools into crayfish conservation research, we can enhance our understanding of genetic diversity and evolution, helping to bend the curve in the anthropogenic biodiversity crisis by supporting effective conservation strategies. Genomic data built upon reference genomes can significantly enhance conservation efforts by providing insights into adaptation, genetic diversity, and population viability. Biodiversity conservation must explicitly consider the genomic diversity of a species to preserve its evolutionary potential and enable adaptive responses to environmental change. This is why promoting reference genome-based approaches in crayfish conservation research is essential for optimising conservation strategies. Knowledge transfer between the research community, conservation practitioners and society is crucial for effective conservation efforts. Reference genomes across the tree of life will serve as a solid foundation for biodiversity assessments, conservation, and restoration, similar to the impact of the Human Genome Project in biomedical sciences (Formenti et al., 2022).

Nowadays, several European initiatives, such as the European Reference Genome Atlas (ERGA) (Mazzoni et al., 2023), Biodiversity Genomics Europe (BGE) and Genomic Biodiversity Knowledge for resilient Ecosystems (G-BiKE) (Heuertz et al., 2023), are dedicated to advancing genomics applications for European species and ecosystems. ERGA aims to generate high-quality reference genomes representing European eukaryotic biodiversity to support conservation efforts for endangered species and key species for agriculture, forestry, and fisheries. G-BiKE, a network funded by the COST program, and BGE seek to establish genomic data as a standard tool for monitoring and managing wild and *ex-situ* populations of plants and animals. All initiatives aim to integrate genetic diversity monitoring into EU policy and planning on biodiversity conservation, including the missions of the European Green Deal and the Biodiversity Strategy for 2030. The joint efforts of initiatives such as ERGA, BGE and G-BiKE are expected to advocate for the incorporation of genomic data into European biodiversity protection programs.

Chapter 2 – Advances and promises of genomics

A new field of science emerged in 1987 called genomics, signifying a pivotal moment in the biological sciences (World Health Organization, 2020; Green, 2023). This multidisciplinary field is dedicated to exploring an organism's entire DNA, known as its genome. Genomics aims to unravelling the intricate genetic code that governs the organism's structure, functionality, and behavioural characteristics. This ambitious undertaking involves not only the identification and characterisation of all genes and functional elements within the genome, but also an exploration of their intricate interactions (Davies, 2002). While remarkable progress has been made in genomics, reading an entire genome directly still remains elusive to some model organisms.

In this chapter, I will commence by examining various genome sequencing platforms, followed by an exploration of diverse assembly strategies and methods. Subsequently, I will delve into different approaches for genome scaffolding and discuss methods for evaluating genome assemblies. Resultant assemblies then undergo decryption to unveil their functions and roles through annotation, where we will survey various annotation strategies. Finally, examples of applications of these genomes across a spectrum of scientific endeavours will be presented.

2.1 Next generation sequencing revolution

"Genes are like the story, and DNA is the language
that the story is written in" – Sam Kean

2.1.1 Short reads

The first sequencing technique was developed in 1975 and later evolved into the Sanger sequencing method in 1977. This method involved incorporating radiolabelled chain-terminating dideoxynucleotides by DNA polymerase and visualising them through electrophoresis gel (Sanger et al., 1977; Sanger and Coulson, 1975). Similarly, a chemical modification-based DNA sequencing method was introduced in 1977 (Maxam and Gilbert, 1977). Both techniques were time-consuming and lacked automation.

Efforts were made to automate these techniques in the following years, and in 1987, the Sanger sequencing method was successfully automated using fluorescent dyes instead of radioactive molecules, along with computer-based data acquisition (Hood et al., 1987). This innovation revolutionised sequencing and paved the way for what we now know as next-generation sequencing (NGS) (Mardis, 2008; Ansorge, 2009; van Dijk et al., 2014). The first NGS technologies (the second generation), which emerged in 2005 with Roche 454, are summarised in Figure 2-1. NGS sequencing methods rely on solid support containing micro channels or wells where the sequencing reactions occur. For these second-generation technologies, DNA amplification is necessary prior to sequencing. This is achieved through emulsion PCR in physically separated water-in-oil droplets or bridge PCR on a flow cell, which is a glass slide with one or more channels. Various sequencing approaches are viable, including single-end sequences that capture DNA from one end, paired-end sequences that encompass both ends of a fragment, and mate-pairs sequences resembles paired-end but with a longer insert size.

In 1998, the Solexa company developed a sequencing-by-synthesis method that utilised fluorescent dyes (Balasubramanian, 2015). In 2007, Illumina acquired Solexa and became the dominant player in the NGS technology market. Their method is based on reversible dye-terminators, allowing the identification of individual nucleotides as they are added to DNA strands (Metzker, 2010). The sequencing process consists of three main steps: amplification by bridge PCR, sequencing, and analysis (Figure 2-2). Starting with purified fragmented DNA, adaptors are added before loading the DNA strand onto a flowcell with nanowells. The DNA strand is then attached to primers on the flowcell surface and replicated by binding to complementary primers using unlabelled nucleotides and DNA polymerase to generate a complementary strand. The double strand bridge is then broken into single strand DNAs, and the cycle is repeated to form small clusters with the same fragments. The sequencing starts by adding primers and fluorescently labelled terminator nucleotides. The DNA polymerase binds to the primer and incorporates a labelled nucleotide, resulting in the termination of polymerisation. Non-incorporated molecules are washed away, a camera detects signals emitted by each cluster and a computer identifies the added base based on fluorescent tags wavelengths. A chemical deblocking step removes the fluorescent dye and the terminating group attached to the nucleotide, and this process is iterated for a fixed number of cycles,

enabling the simultaneous sequencing of thousands of genomic regions through massively parallel sequencing.

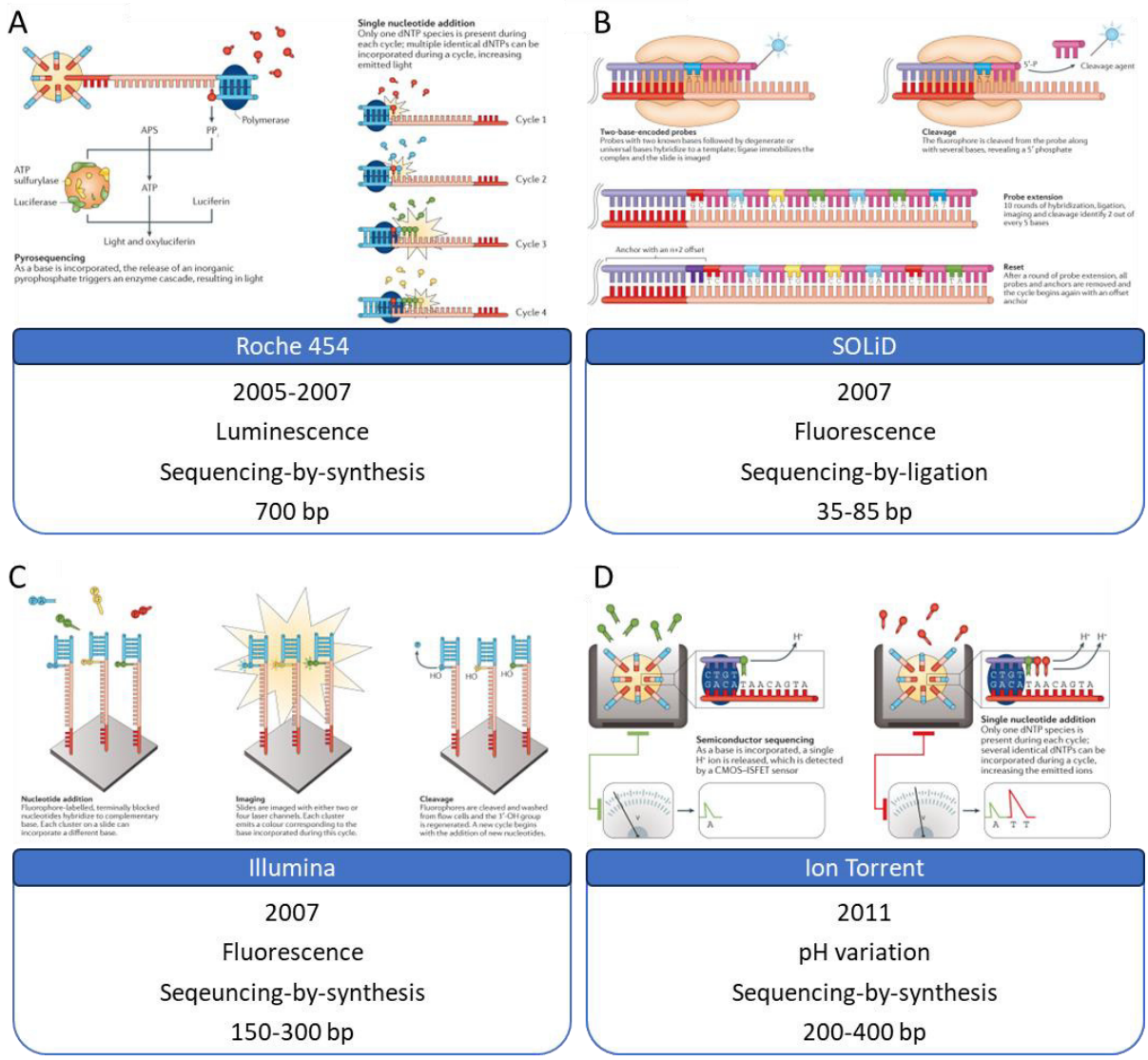

Figure 2-1. New generation sequencing (NGS) technologies. For each technology, year of release, type of detection, type of synthesis, and read length are provided. A. Roche 454. Wells are subjected to a sequential flow of unlabelled nucleotides of a single type allowing for the synthesis of a complementary DNA strand by DNA polymerase. Light emission proportional to number of incorporated nucleotides is monitored, resulting from ATP converted from ejected pyrophosphates. B. SOLiD. A mixture of fluorescent probes with two interrogation bases compete to bind to a primer. After hybridization and ligation, the colour is detected, the bound octamer is cleaved, and the cycle is repeated. D. Illumina. After incorporation of a coloured labelled terminator nucleotide by a polymerase, the fluorescence is detected, and the fluorophore removed to restart the cycle. E. Ion Torrent. The complementary strand is sequenced by adding a sequential flow of single-type nucleotides. Released protons during incorporation are detected by ion sensors. Figures adapted from Goodwin et al., 2016.

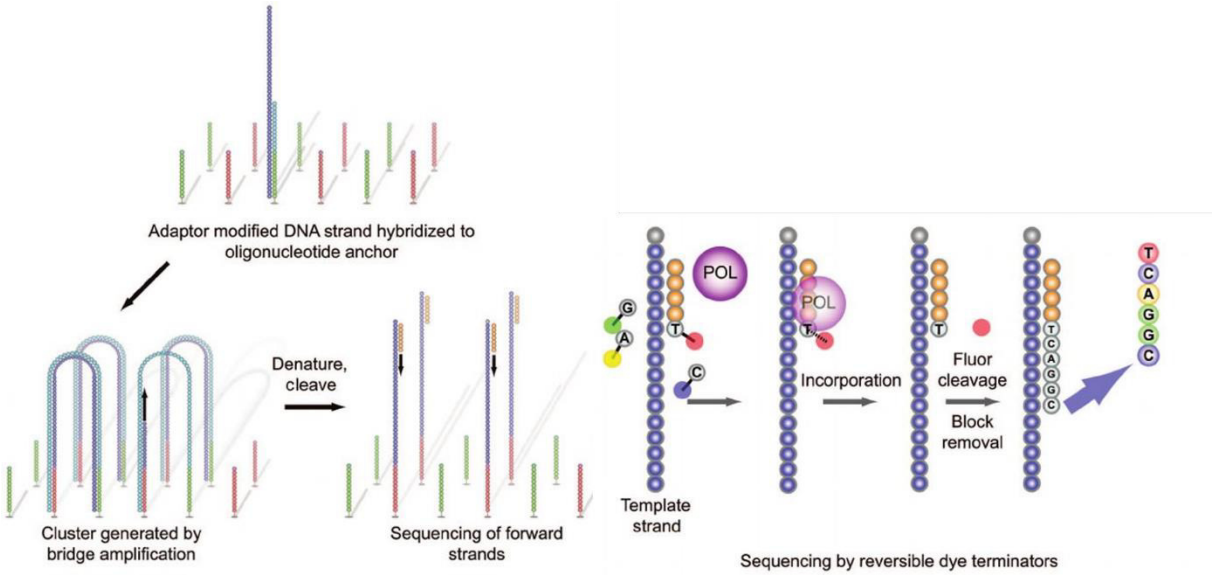

Figure 2-2: Illumina sequencing. DNA strands on the flow cell's surface replicate by binding to complementary primers, forming clusters, and unlabelled nucleotides. The bridge is then broken into single strand DNAs. Sequencing begins with primers and labelled reversible terminator nucleotides. DNA polymerase incorporates labelled nucleotides, terminating polymerization. A camera detects signals emitted by clusters. A deblocking step removes fluorescent dye and terminators. Figure from Voelkerding et al., 2009.

2.1.2 Long reads

Third-generation sequencing can be defined as the emergence of single-molecule real-time (SMRT) sequencing technologies that eliminate the need for DNA amplification. These third-generation sequencing technologies offer a distinct advantage by producing significantly longer reads compared to traditional NGS methods. Pacific Biosciences, Inc. (PacBio) stands out as a pioneer in third-generation sequencing. In 2010, PacBio introduced a groundbreaking method utilising zero-mode waveguides (ZMW), marking a significant leap forward in DNA sequencing techniques (Levene et al., 2003; Eid et al., 2009). ZMW technology involves tiny "nanoholes" containing a single DNA polymerase molecule. The circularised DNA molecule to be sequenced is loaded into the ZMW and forms a complex with the polymerase. During sequencing, the DNA polymerase synthesises a complementary strand in the presence of fluorescently labelled nucleotides. At the incorporation of these nucleotides, a light pulse excites the fluorophore that emits a signal, which is instantly captured and analysed by highly sensitive detectors positioned beneath the ZMW. The fluorophore is then removed as part of the natural incorporation of the base into the new DNA strand and released into the buffer. PacBio reads can be sequenced following two methods: Continuous Long Read (CLR) and

Circular Consensus Sequencing (CCS) (Figure 2-3). CLR sequencing will produce longer reads than CCS, however, the error rate is much higher.

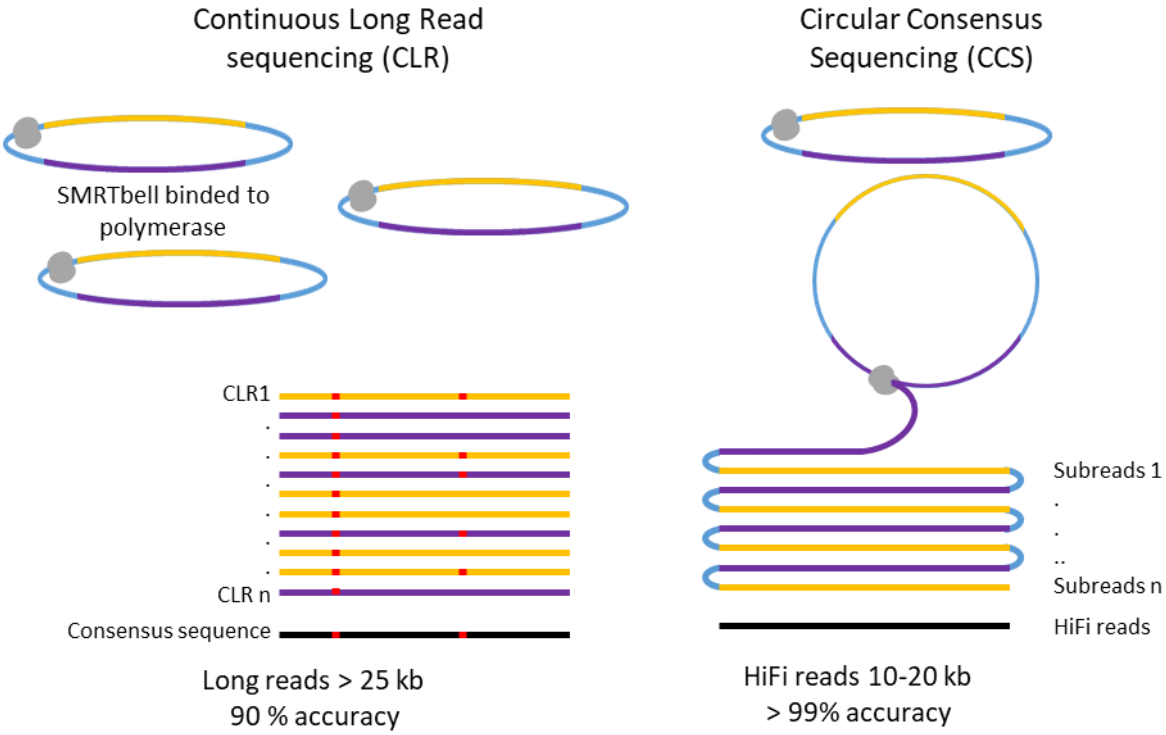

Figure 2-3: PacBio sequencing. DNA (yellow for forward strand, purple for reverse strand) is fragmented and ligated to hairpin adapters (light blue) to form the SMRTbell template. The SMRTbell is bound by a DNA polymerase immobilized on the bottom of the zero-mode waveguides. Fluorescently labelled nucleotides are added to begin the sequencing reaction. Incorporation of nucleotides emits a signal detected by a camera. CLR are generated by sequencing a SMRTbell template containing large inserts and only one or a few passes can be made because of the large DNA insert size. HiFi contains smaller inserts allowing multiple passes around the SMRTbell template.

Another notable third-generation sequencing technology was developed by Oxford Nanopore Technologies (ONT) in 2012 (Check Hayden, 2012). ONT employs nanopores embedded in electrically resistant membranes to perform sequencing (Figure 2-4; Lu et al., 2016). It measures disruptions in electrical current as individual bases pass through the nanopore, allowing for the determination of specific single molecule sequences. In ONT sequencing, a hairpin structure is ligated to double-stranded DNA, which is then linearised when passing through the pore, enabling the system to read both strands in a continuous sequence.

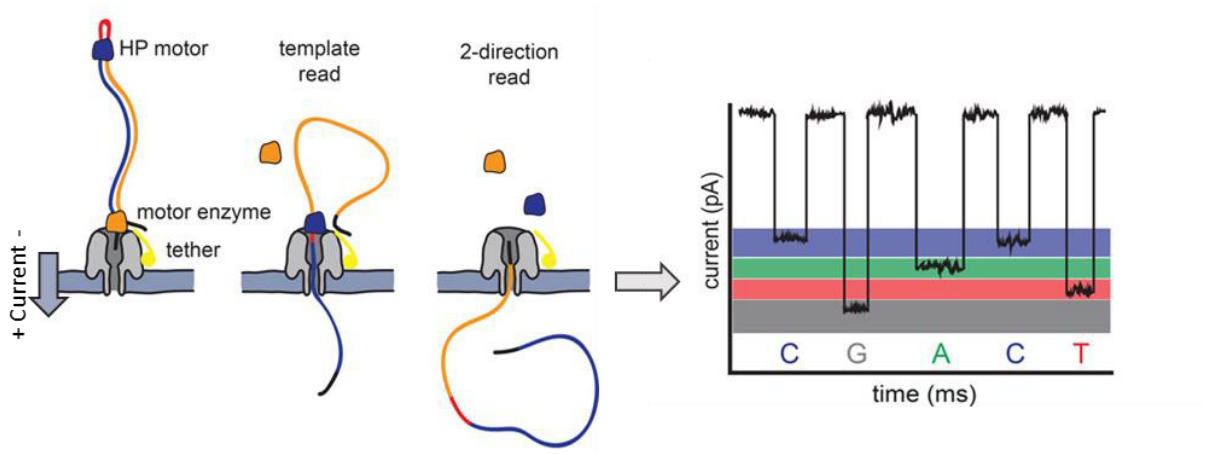

Figure 2-4: ONT sequencing. DNA templates are ligated by an adaptor with a motor enzyme, and a hairpin oligo that is bound by the HP motor protein. Bases are identified by changes in current induced by the nucleotides when passes through the pore. Figure adapted from Reuter et al., 2015.

2.1.3 Challenges of each technology

In the realm of DNA sequencing technologies, Illumina stands out for its remarkable accuracy. The foundation of this precision lies in the detection of fluorescent dyes emanating from entire clusters of the same DNA fragment being sequenced simultaneously. This collective emission generates a consensus dye signal, a key factor contributing to Illumina's high accuracy. Nonetheless, to enhance sequencing speed, Illumina transitioned from the original 4-dye method (one for each nucleotide) to a 2-dye method. In the 2-channel (2-dye) method, two complementary nucleotides are discernible through distinct colours, one through the combination of both colours, and the last through an absence of fluorescent detection, which introduces potential bias (Stoler and Nekrutenko, 2021). The absence of fluorescence can indicate either the incorporation of a G base or the complete absence of base incorporation. Additionally, the relatively limited read length (150 - 300 bp) inherent to Illumina technology poses challenges when resolving intricate genomic regions, repetitive elements, and structural variations. The use of paired-end or mate-pair reads can bring an indication of the length of repeated regions or the distance between contigs thanks to the insert size, however, it does not solve repeated regions of longer size than the inserts.

Long-read sequencing technologies provide a solution to the challenges posed by repetitive sequences by generating significantly longer reads (10 -> 25 kb for PacBio, 15 -> 100 kb for ONT) when compared to Illumina sequencing (150 - 300 bp). Nevertheless, these technologies,

based on single molecules, face the challenge of error accumulation, which can reach more than 10% error rate (Ardui et al., 2018; Jain et al., 2017). While Nanopore technology prioritised miniaturisation, rendering it portable and handheld for field applications, PacBio concentrated on error reduction. In the PacBio approach, Continuous Long Reads (CLR) are characterised by long reads that, comparable to Nanopore, present a high error rate. However, the circularisation of DNA sequences permits multiple passes for slight strand reduction, albeit still significantly longer than Illumina reads (Wenger et al., 2019). Through the comparison of multiple sequences, errors in individual reads are rectified, creating Circular Consensus Sequences (CCS), and yielding highly accurate consensus reads known as High Fidelity (HiFi) reads.

Sequencing small organisms with limited DNA poses a challenge in meeting the required DNA quantities for sequencing platforms. Although achievable with ONT (Arakawa, 2023; Freedman et al., 2016), there is no established standardized protocol. In contrast, PacBio provides protocols specifically designed for ultra-low input DNA sequencing (Procedure & Checklist - Preparing HiFi SMRTbell Libraries from Ultra-Low DNA Input, 2021). PacBio incorporates a PCR amplification step before conventional sequencing processing. Furthermore, PacBio's latest platform, Revio, introduces advanced automation, elevating accuracy by optimised algorithms, and output by increased number of ZMW, marking yet another milestone in the ever-evolving landscape of DNA sequencing technologies (PacBio Revio | Long-read sequencing at scale).

Various DNA sequencing methods, from early Sanger sequencing to the latest third-generation technologies like PacBio and Oxford Nanopore, have greatly advanced our ability to read DNA sequences. Once the raw sequencing data is obtained, the next challenge lies in genome assembly, the intricate process of reconstructing the complete genetic map from these fragmented sequences.

2.2 From reads to annotated genome

Genome assembly: the art of trying to make one big thing from millions of very small things - Keith Bradnam

Genome assembly is a critical process in genomics that involves piecing together the puzzle of an organism's DNA (Wajid and Serpedin, 2016; Dominguez Del Angel et al., 2018; Jung et al., 2020; Ekblom and Wolf, 2014). To facilitate genome assembly, closely related genomes can be used as reference, such as that of a related species or a previous version of the same organism, to aid in the assembly process of the target species (Figure 2-5). This kind of assembly is called reference-based or assisted assembly (Gnerre et al., 2009; Lischer and Shimizu, 2017). By aligning the new sequencing data to the reference genome, matching regions can be identified, and the relative positions of sequences can be determined to assemble the reads. Reads can also be aligned to a reference genome and grouped by regions of similarity to the genome. These groups of reads can then be assembled separately, but because they are ordered by regions of similarity to the reference it will simplify the assembly. While reference-based assembly is efficient and can produce high-quality results, it may not capture long range variations or unique features in the target genome, making it most suitable for closely related organisms or when a well-characterised reference genome is available.

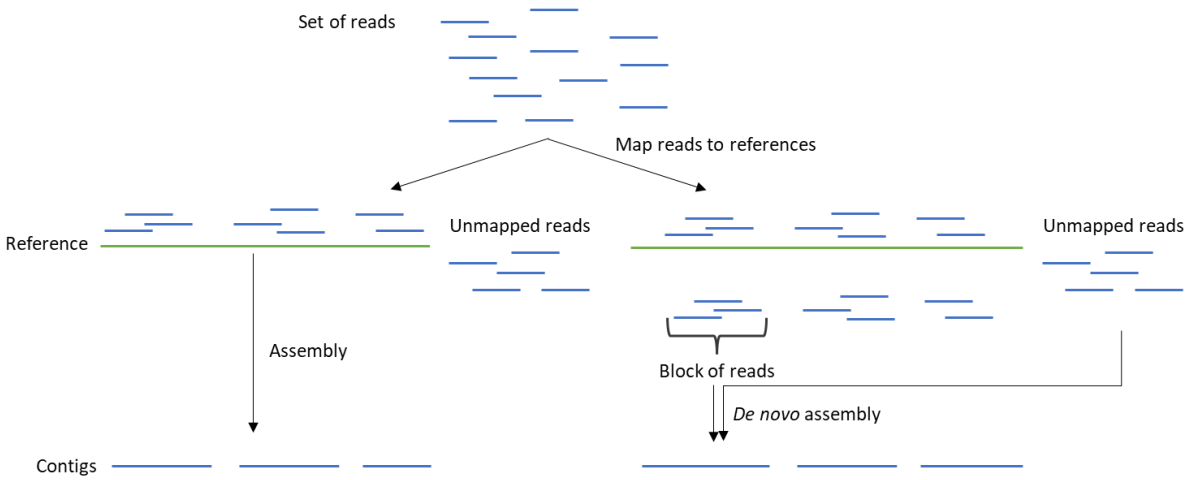

Figure 2-5: Reference-based assembly. Reads are mapped to a reference genome. On the left, reads are directly assembled in contigs based on alignments. On the right, unmapped reads, and mapped reads grouped by blocks, are de novo assembled in contigs.

In the context of non-model organisms, reference genomes of closely related species are often not available. In such cases, *de novo* genome assembly becomes a necessity (Q. Chen et al.,

2017; Sohn and Nam, 2018). *De novo* assembly is the process of constructing a genome from scratch when no reference genomes are available and can involve the use of various assembly methods (Simpson and Pop, 2015).

Greedy assemblers

The earliest approaches involved the application of greedy algorithm assemblers such as SEQAID (Peltola et al., 1984) and VCAKE (Jeck et al., 2007). This strategy follows a step-by-step procedure, prioritising the amalgamation of reads based on their overlapping quality, whether in terms of length or a more intricate quality metric (Figure 2-6). To identify these local optima, the process incorporates pairwise distance calculations and clusters reads with significant overlaps. During this iterative process, contigs grow by either adding new reads or merging with previously constructed contigs, extending the assembly by adding each individual read. The iteration continues until a minimum quality threshold is achieved. However, overlapping regions that conflict with existing contigs are excluded from consideration. Although the greedy strategy was once widely employed, even playing a pivotal role in the Human Genome Project, it exhibits limitations, particularly when dealing with large read sets and regions containing repetitive sequences (Bang-Jensen et al., 2004). Consequently, in the realm of contemporary *de novo* sequence assembly, these techniques have fallen out of favour, and were replaced by graph-based algorithms such as Overlap Layout Consensus (OLC) approach and de Bruijn-graph.

Figure 2-6: Greedy assembly. Pairwise distances are calculated between reads. Contigs grow by adding new reads with significant overlap to previously constructed contigs and merging them, extending the assembly by adding each read. Reads with overlapping regions that conflict with existing contigs are excluded from consideration.

Overlap Layout Consensus (OLC) assemblers

In the OLC approach, used for example by the assembler Canu (Koren et al., 2017), each read is represented as a vertex in the graph and edges connect pairs of vertices if their reads overlap (Figure 2-7). The assembly process encompasses three key stages: overlap, layout, and consensus. During the overlap stage, overlapping read pairs are identified such as in the greedy approach. However, to accelerate the process, an index mapping k-mers to reads is constructed, with k-mers being a segment of the reads of a length of k. This approach significantly reduces the search space by rapidly identifying potential overlaps. In the layout stage, the assembly graph is constructed, and the reads are ordered and oriented to generate unitigs, i.e., groups of reads that can be assembled unambiguously with minimal risk of misassembly. Finally, a consensus sequence is computed from ordered and oriented reads, resulting in a set of contigs. The OLC approach has shown significant success, particularly with long-read sequencing technologies such as PacBio and ONT, which produce reads with longer and more informative overlaps. However, OLC methods have been challenged by the computational burden posed by extensive short-read data. Additionally, the resulting overlap graph could become impractically large with high-depth data and numerous spurious overlaps. Consequently, the de Bruijn graph approach is often adopted for assembling high-throughput short-reads.

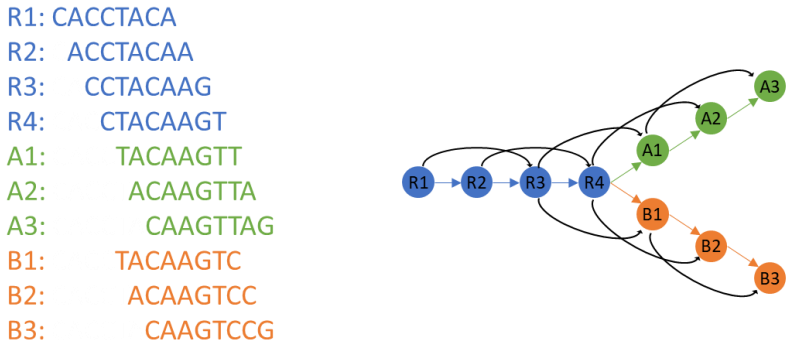

Figure 2-7: Overlap Layout Consensus assembly. Overlapping read pairs are identified and an index of mapping k-mers is constructed. Layout stage correspond to the graph construction, reads are ordered and oriented to generate unitigs. A consensus sequence is computed from ordered and oriented reads, resulting in a set of contigs.

De Bruijn graph assemblers

De Bruijn graph-based assembly methods employ a strategy where sequences are dissected into k-mers, forming a graph by linking overlapping k-mers (Figure 2-8). This method is used

by numerous tools such as ABySS (Simpson et al., 2009; Jackman et al., 2017) and SOAPdenovo (Luo et al., 2012). After extraction, k-mers are added as graph vertices, and overlapping k-mers are connected by edges. The assembly task is navigating a path through the graph, ensuring each edge is traversed precisely once. This approach is widely used in short-read sequencing and can efficiently handle large datasets. However, the clarity of the graph is often muddled by sequencing errors and sampling biases. Low-frequency k-mers, often corresponding to sequencing errors can be removed from the graph for clarity. The existence of repetitive sequences introduces multiple possible pathways, with only one being the correct solution, rendering it impossible to definitively trace a single path through all edges. Consequently, most assemblers aim to construct contigs that capture the unambiguous, linear segments of the graph. The de Bruijn graph approach offers a significant computational advantage over other assembly strategies, because the read overlaps are implicitly represented in the graph's structure. Nonetheless, managing the de Bruijn graph within stringent memory constraints became a significant challenge. Several methods have been tried to solve this problem, such as using a hash table of k-mers to represent the graph or implementing Bloom filters to reduce the memory storage of k-mers.

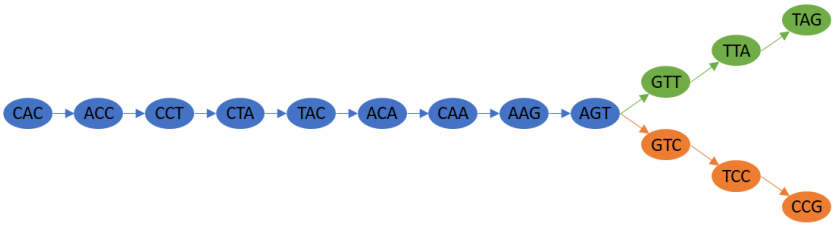

Figure 2-8: De Bruijn graph assembly. K-mers are added as graph vertices and overlapping k-mers are connected by edges. Navigating a path through the graph, ensuring each edge is traversed precisely once produces the assembly.

Hierarchical assemblers

Hierarchical assembly represents a distinct approach, tailored specifically to accommodate long but less accurate sequencing reads (Figure 2-9; Chin et al., 2013). This methodology, used by HGAP (Chin et al., 2013), begins by identifying the longest available reads. Subsequently, shorter reads are aligned to these lengthy counterparts, leading to the generation of highly accurate consensus sequences, often referred to as pre-assembled or error-corrected reads. During this process, low-coverage regions within the longest reads are trimmed. The resulting pre-assembled reads are then aligned with one another to create contigs.

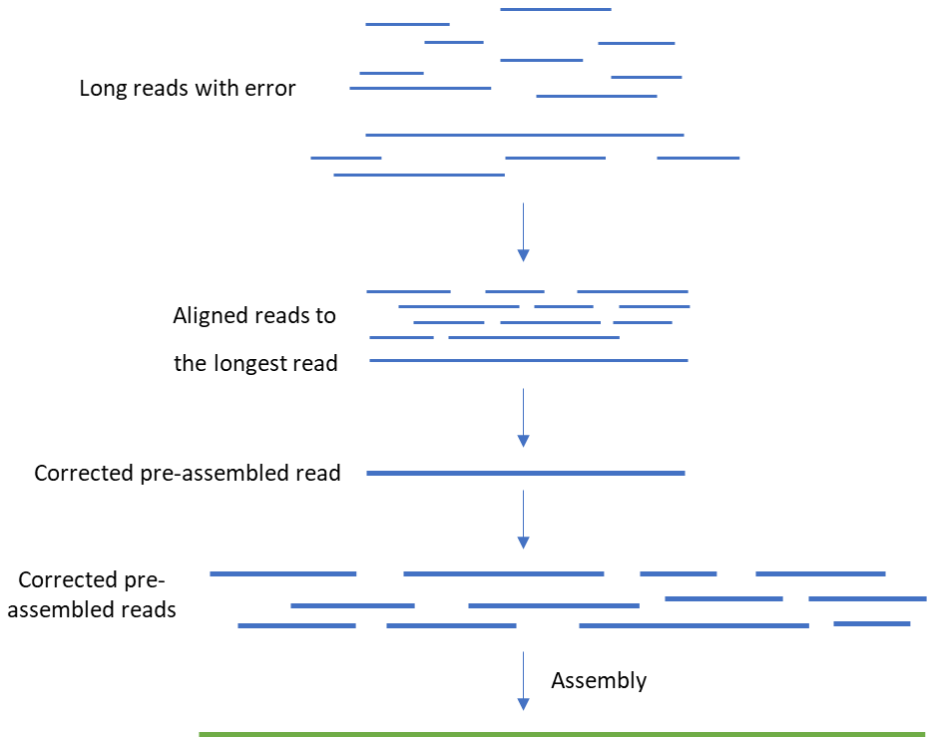

Figure 2-9: Hierarchical assembly. Longest reads are selected, and other reads are mapped and assembled into an accurate pre-assembled reads. Pre-assembled reads are then used for the genome assembly.

Hybrid assemblers

Hybrid assembly methods, such as used by MaSuRCA (Zimin et al., 2013) and DBG2OLC (Ye et al., 2016), have emerged as a powerful approach that leverages the strengths of both short-read and long-read sequencing technologies. By combining data from these two sources, hybrid assembly effectively mitigates the limitations inherent to each technology, resulting in a comprehensive and accurate genome assembly. This approach is particularly invaluable when dealing with complex genomes rich in repetitive elements. There are several strategies for hybrid assembly (Figure 2-10). One common approach involves using short-read data to correct long reads (ONT and PacBio CLR), enhancing their assembly accuracy by eliminating sequencing errors. This correction process is crucial in preventing misassemblies, typically utilising OLC approach for assembling the long reads. Alternatively, hybrid assembly may begin with the assembly of short reads using a de Bruijn graph-based method. Subsequently, these contigs are reorganised and oriented by aligning them to long reads. This alignment can be achieved through OLC methods or mapping, with long reads serving the dual purpose of filling gaps between short-read contigs and organise and ordered them. In another variant of hybrid assembly, short reads are initially assembled and then employed to correct long reads,

subsequently facilitating the assembly of the corrected long reads. These diverse hybrid assembly methods highlight the flexibility and effectiveness of this approach in generating high-quality genome assemblies.

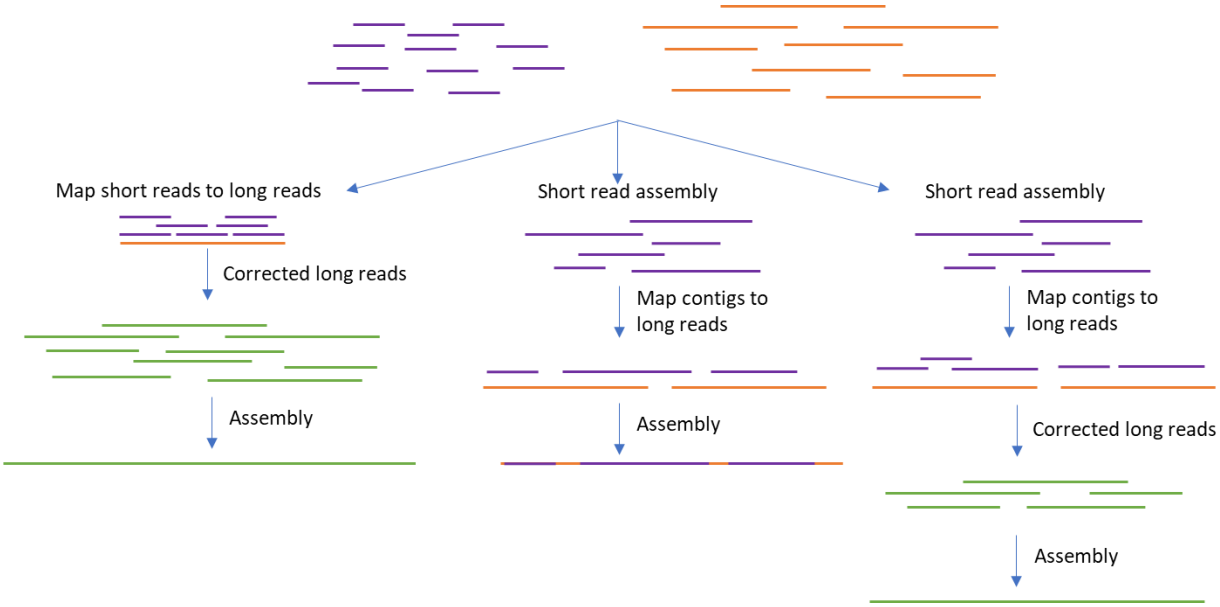

Figure 2-10: Hybrid assemblies. Short reads can be used to correct long reads that are then assembled (left). Short reads can be assembled in contigs, and organised and oriented using long reads, also used to fill gaps (middle). Short reads can be assembled, then contigs are used to correct long reads that are then assembled (right).

2.2.1 Improving assembly contiguity

To enhance the continuity of a genome assembly, a pivotal step involves genome scaffolding, which can be accomplished through various techniques (Q. Chen et al., 2017; Dominguez Del Angel et al., 2018; Ekblom and Wolf, 2014; Jung et al., 2020). The order, relative orientation and distance between two contigs can improve the quality of the assembly, even if nucleotide bases between contigs are not known and produce scaffolds.

In cases where a reference genome is available, one approach is to align the initially assembled contigs with the reference genome (Figure 2-11). This alignment serves the purpose of organising and assembling contigs effectively. Alternatively, the use of paired or mate-pair reads with known long insert sizes, can be employed to bridge the gaps between contigs (Ekblom and Wolf, 2014). These paired reads are aligned to the contigs, allowing for the determination of both the distance and orientation between them. In a similar vein as in hybrid assembly methods, long-read sequencing can also be aligned to establish connections

and sequence orders among contigs. This results in more contiguous assemblies and offers the capability to fill gaps between unitigs with repetitive contigs.

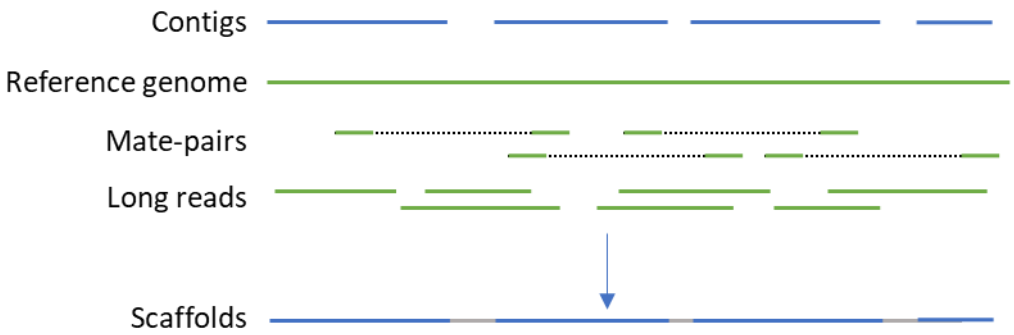

Figure 2-11: Scaffolding methods. Reference genome, mate-pair reads, and long reads can be used to align contigs. Contigs are organised and ordered allowing the identification of relative distances between contigs to form scaffolds.

Furthermore, advancements in long-range scaffolding technologies, such as optical mapping and Hi-C, have emerged to provide information at the chromosomal level, contributing significantly to the refinement of genome scaffolding. Optical mapping is a high-resolution imaging technique employed to construct a physical representation of a genome (Figure 2-12; Zhou et al., 2007). The process involves exposing DNA fragments to enzymes that target specific sites, leading to digestion of one strand. The digested strand is then repaired using fluorescently marked nucleotides, with each site referred to as label. There are two methods for DNA labelling: NLRS (Nicking, Label, Repair, and Stain), which employs restriction enzymes, and DLS (Direct Label and Stain), utilising the DLE-1 enzyme for more uniform marking without DNA cleavage. The labelled restriction sites on genomic DNA molecules are loaded into nano-channels and linearised over multiple cycles. A camera captures images of these nano-channels after fluorochrome activation, which are then processed to derive information on label positions and distances between them. These processed molecules are assembled to generate an optical map known as the whole genome map. *In silico* digestion of the assembly is performed using the same enzyme, and the positions of labels are compared between the optical map and the digested assembly. This comparison facilitates the organisation and orientation of contigs. This is especially valuable when dealing with complex genomic regions like centromeres, which contain numerous repetitive elements. Additionally, it enables the estimation of gap sizes between contigs, ultimately culminating in the generation of scaffolds. The applications of optical mapping are diverse, including validation and enhancement of *de*

novo genome assembly, analysis of structural variations such as translocations, and copy number variations (CNVs).

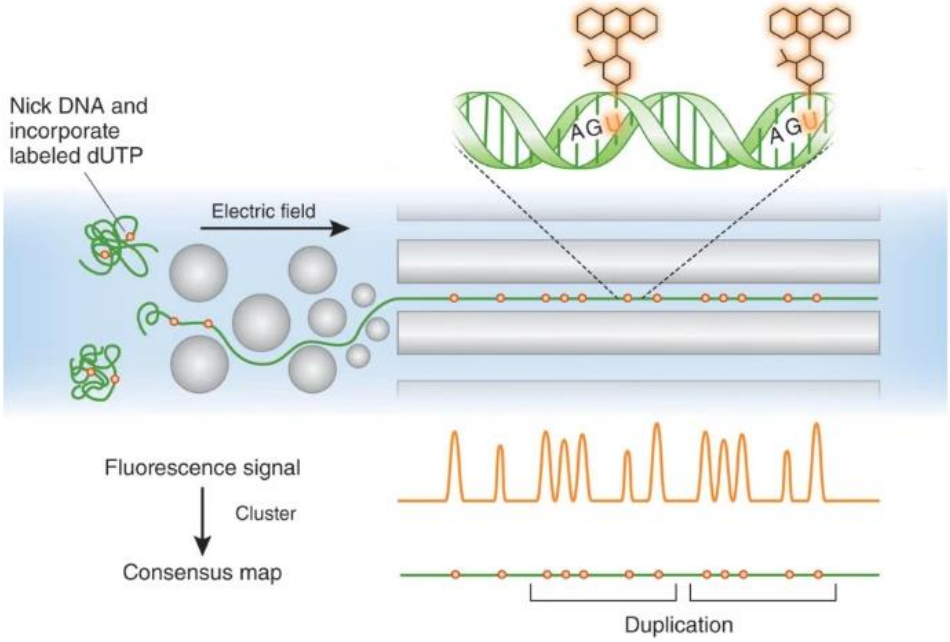

Figure 2-12: Optical mapping. DNA molecules are digested by enzymes on one strand and repaired by a polymerase to incorporate fluorescently labelled nucleotides. Labelled molecules are driven through tiny obstacles into channels using an electric field. A camera captures images of these nano-channels after fluorochrome activation, which are then processed to derive information on label positions and distances between them. Figure from Michaeli and Ebenstein, 2012.

Hi-C is a technique for capturing the conformation of chromosomes, providing valuable insights into the three-dimensional architecture of the genome (Figure 2-13; Lieberman-Aiden et al., 2009). This process exploits chromosome compaction, which brings physically distant DNA regions into proximity, even though they may be separated in the linear sequence. In scaffolding, Hi-C serves a dual purpose, enabling the identification of physical interactions between contigs and placing them within their accurate spatial context. During Hi-C sequencing, interactions between two DNA regions create contact points, with distances ranging from a few nucleotides to tens of kilobases (Liu et al., 2021). Various commercial protocols are available for conducting Hi-C experiments, all based on a common set of reactions. The first step involves fixing the chromatin conformation within the nucleus while maintaining these contacts. Subsequently, DNA is fragmented using either restriction enzymes or DNAse. The contact points are then bridged by adapters carrying biotin. After protein removal, biotin-bearing fragments are purified using streptavidin beads, and the DNA is

prepared for sequencing. Although chromosome-scale assemblies are often achieved, challenges arise in specific regions like centromeres and large tandem repeats. The initial alignment of Hi-C reads with the assembly and the subsequent removal of reads aligning to different positions can make organising these complex regions difficult. Additionally, estimating and setting the sizes of gaps between contigs can be arbitrary.

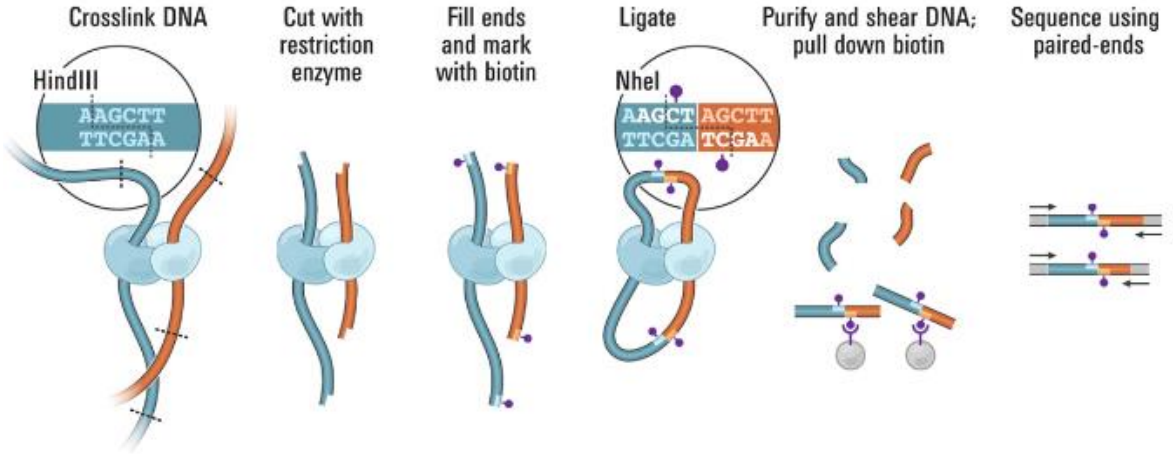

Figure 2-13: Hi-C sequencing. Spatially adjacent DNA fragments (dark blue and orange) are fixed using formaldehyde and linked by proteins (light blue). DNA is fragmented using restriction enzymes or DNAse. The contact points are then bridged by adapters carrying biotin (purple dots). After protein removal, the biotin-bearing fragments are purified, and the DNA is prepared for paired-end sequencing. Figure from Lieberman-Aiden et al., 2009.

2.2.2 Assembly evaluation

Assembly evaluation is a crucial step in the genome assembly process, serving as a means to assess the quality and accuracy of the reconstructed genome. Evaluating an assembly involves comparing it to a reference genome, if available, or utilising various metrics and statistics to measure factors such as contiguity, correctness, and completeness. Key evaluation metrics encompass N50/N90, which corresponds to the minimum scaffold or contig length at or above which 50%/90% of the assembly lies, thus offering insights into the assembly's contiguity (Nature Biotechnology, 2018). The L50 corresponds to the smallest number of contigs that allows to cover 50% of the assembly. Another critical measure is NG50, indicating the scaffold length that covers 50% of the genome's total length. The evaluation also considers coverage, deeming an assembly accurate when at least 90% of the bases exhibit a 5X read coverage. Furthermore, the total length of the assembly is compared to the estimated genome size (taking into account the number and length of gaps) to evaluate the assembly completeness. In addition, tools like BUSCO (Manni et al., 2021) scrutinise the presence and completeness of

conserved genes within the assembly, to assess genome completeness. Aligning reads to the assembly enables the assessment of assembly consistency. This alignment process allows for the examination of read coverage across the assembly, facilitating the detection of regions with either low (possible assembly error) or high (repeated region) coverage that may be of interest. Additionally, it aids in the identification of misassemblies through the analysis of unaligned ends of reads. In sum, assembly evaluation determines the reliability of the assembled genome, guiding subsequent refinement and validation efforts, and ultimately ensuring the accuracy of downstream genomics analyses and interpretations.

2.2.3 Annotation

Genome annotation is a fundamental aspect of genomics, involving the decoding of genetic information within an assembly to pinpoint genes, functional elements, and other important features. It comprises two key components: structural annotation and functional annotation (Brent, 2005; Dominguez Del Angel et al., 2018; Jung et al., 2020; Yandell and Ence, 2012). The structural annotation refers to the process of locating functional elements within a genome. The annotation process starts by identifying repetitive elements (discussed in Chapter 3) and non-coding RNA genes (e.g., tRNA, rRNA, microRNAs), which play vital roles in cellular processes, including gene regulation. Predictions of eukaryotic genes involves identifying their structures: exons, introns and splice sites. For coding genes, it requires localising coding sequences (CDS), start and stop codons, and untranslated regions (UTRs) (Figure 2-14). Furthermore, a comprehensive annotation involves identification of promoters, enhancers, and other regulatory elements that control gene transcription.

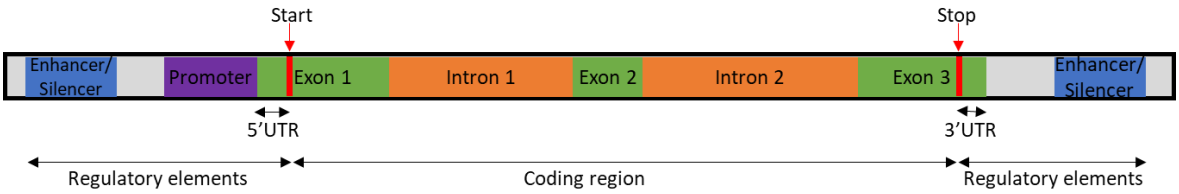

Figure 2-14: Eukaryotic gene structure. Enhancer and silencer (blue), promoter (purple) and untranslated regions (UTR, green) form regulatory elements. Coding regions are bounded by start and stop codons.

The prediction of protein-coding genes is crucial in genome annotation. It follows three primary approaches: *ab initio*, evidence-based, and combiners. The *ab initio* method relies on genomic sequence data, involving the development of specific statistical models and software

parameters tailored to individual genomes that need to be trained and optimised. This approach is labour-intensive but is valuable in predicting genes without available homolog sequences.

In contrast, the evidence-based approach leverages similarities with other sequences, such as transcripts and polypeptides, as a source of information. A wealth of available protein sequences in databases, like RefSeq (O'Leary et al., 2016), and UniProt (The UniProt Consortium, 2023), makes this method universally applicable. It offers strong clues about gene presence and location but may not provide precise gene structure details. Polypeptide sequences, being more conserved than nucleotide sequences, can even be aligned across distantly related species. However, they may not accurately depict the gene's exact structure. Transcripts, while offering accurate structural information, are less comprehensive and occasionally contain inaccuracies due to incomplete mRNA processing. Moreover, not all genes have associated transcript information.

Combiners represent a widely favoured gene prediction approach, integrating aspects of both *ab initio* and evidence-based methods. They often start with an *ab initio* prediction and then incorporate evidence-based information. Different combiners have varying approaches; some aim to select the most suitable model or form a consensus from input data, which may include predictions from *ab initio* tools. Others take a more integrated approach, allowing the evidence data to modify the *ab initio* prediction. The latter approach prioritises consistency and permits one type of information to override the other if it leads to a more coherent prediction.

To assess predicted genes, the Annotation Edit Distance score (AED), which compares predicted genes with supporting evidence, such as transcript sequences, can be used. A low AED score signifies a more precise prediction. When over 90% of genes exhibit AED scores below 0.5, it indicates a well-annotated genome.

After gene prediction, the subsequent phase involves functional annotation, which encompasses the assignment of biological and biochemical roles to the identified genes and components. This entails the characterisation of gene functions within biological processes,

including metabolism, signalling, and development, by scrutinising protein domains and functional patterns. Functional annotation requires the use of a controlled and standardised vocabulary, commonly known as ontology, to name predicted functional features. The Gene Onthology (GO) (Ashburner et al., 2000; The Gene Ontology Consortium et al., 2023) is the most comprehensive and widely used vocabulary for this purpose. It classifies genes into three categories based on functional properties: molecular function, biological process, and cellular components. Genome annotation requires the annotation of each protein domain, which are structural and functional units within proteins. These domains can fold independently and often have specific roles in cellular processes. Computational methods, such as Pfam (Mistry et al., 2021) and InterPro (Paysan-Lafosse et al., 2023), are frequently employed for the identification and annotation of protein domains through comparative analysis with already annotated proteins.

Comparative genomics assumes a critical role in genome annotation by comparing the newly sequenced genome with existing annotated genomes. This comparison helps pinpoint conserved genes, gene families, and evolutionary connections. Predicted protein functions can be computationally deduced by assessing the similarity between the target sequence and sequences within public repositories like UniProt (The UniProt Consortium, 2023), RefSeq (O'Leary et al., 2016), and Ensembl (Martin et al., 2023). While these repositories house and curate annotated genomes, caution is advised when assigning functions solely based on sequence similarity, as shared domains between evolutionarily independent sequences may misleadingly suggest homology. Annotation quality is enhanced when a significant portion of recognisable domains is found within predicted proteins, surpassing 50% of the predicted proteins. Therefore, when possible, prioritising orthologous sequences for annotation over merely similar ones is recommended. With the increasing abundance of sequences in public repositories, diverse searches can be conducted, and their outcomes can be amalgamated to establish a consensus annotation. Once proteins are annotated, pathway databases, such as KEGG (Kanehisa and Goto, 2000) and STRING Database (Szklarczyk et al., 2023), provide insights into the interconnected biochemical pathways in which these genes participate, providing a holistic view of their roles within biological systems.

Evaluating the completeness of genome annotation is also advisable, accomplished by inspecting the presence of conserved protein families using BUSCO (Manni et al., 2021). Based on these results, one can determine whether further refinement of gene prediction is necessary before reevaluating the annotation.

2.3 Exploitation of Whole Genome Sequencing

Only time and money stand between us and knowing the composition of every gene in the human genome – Francis Crick

2.3.1 Functional applications

A well-assembled and annotated genome stands as a reference, serving as the fundamental cornerstone upon which our comprehension of an organism's genetic constitution and functional diversity is constructed. It provides an extensive panorama of an organism's genetic composition, encompassing not only protein-coding genes but also functional RNA molecules. Furthermore, it affords insights into genome organisation, elucidating the intricate arrangement of genes, regulatory elements, and repetitive sequences. This understanding is pivotal for deciphering how genome organisation influences an organism's biology and adaptation.

The integration of reference genomes with transcriptomics unveils the panorama of genomic diversity and allows exploration of variations in gene expression across diverse individuals, tissue types, and under varying conditions (Alvarez et al., 2015; Conesa et al., 2016). This approach not only provides a holistic outlook on actively expressed genes but also delves into the nuances of their expression, thereby shedding light on an organism's developmental processes, fundamental biological functions, responses to stimuli, adaptability, and disease resistance (Campbell et al., 2018; Green II and Kronforst, 2019). This integration unveils the mechanisms through which genes shape an organism's traits and responses to its environment, thus presenting a comprehensive perspective on cellular functions and interactions. This holistic view supports advancements in drug discovery, biomarker identification, and species conservation.

The amalgamation of genomic and epigenomic data provides a deeper understanding of how epigenetic modifications, such as DNA methylation and histone modifications, modulate gene expression (Rey et al., 2016, 2020). This synergy unravels the intricate interplay between genetics and epigenetics, offering profound insights into developmental processes, disease mechanisms, and cellular responses.

When integrated with proteomics and metabolomics, a reference genome extends its utility to unravel the complex molecular mechanisms underlying biological processes (Wanichthanarak et al., 2015). Proteomics enables comprehensive profiling of an organism's entire set of expressed proteins, elucidating their functions and interactions. In contrast, metabolomics delves into the realm of small molecules involved in metabolic pathways, shedding light on the intricate processes governing metabolism.

The seamless integration of these 'omics' approaches with genome assembly provides a holistic understanding of cellular functions (Fu et al., 2021; Wang and Zhang, 2014; Zhang and Kuster, 2019). This enables researchers to decipher how genes, proteins, and metabolites collectively contribute to an organism's traits, responses to its environment, and overall biological complexity, supporting biodiversity conservation strategies.

2.3.2 Evolution and biodiversity

Reference genomes are indispensable tools in the study of evolution and biodiversity, offering profound insights into genetic diversity and evolutionary processes (Theissinger et al., 2023). One of their primary applications lies in tracking the evolution of gene repertoires across species, achieved through comparative genomics. This comparative analysis of reference genomes from diverse species unravels crucial details about evolutionary relationships, adaptations, and speciation events. By identifying genetic gains and losses, scientists discern the genetic foundations of specific traits and adaptations. Comparing entire genomes enhances our understanding of evolutionary relationships and enables the resolution of complex evolutionary scenarios. Scrutinising genome organisations among different species facilitates the analysis of synteny, revealing significant differences such as gene inversions, duplications, gains, or losses. These comparisons offer insights into the age of genome modifications and contribute to the construction of phylogenetic trees that trace the intricate

evolution of species (Ahrenfeldt et al., 2017; Sakoparnig et al., 2021). Such phylogenetic trees provide a comprehensive overview of the evolutionary tapestry of life forms, leading to the creation of comprehensive phylogenetic trees as exemplified in projects like the Tree of Life project (Tree of Life Web Project).

A genome assembly's significance extends to the identification and classification of unknown species, particularly valuable for biodiversity studies. Reference genomes obtained from various environmental samples, such as soil or ocean water, are used to identify and catalogue the genetic diversity contributing to our understanding of the Earth's biodiversity richness. Assemblies illuminate the evolutionary history of extinct species and their interactions with modern counterparts through the sequencing and assembly of ancient DNA (Muffato et al., 2023). Environmental genomics, an essential field supported by whole genome assembly, allows the exploration of biodiversity within diverse ecosystems.

Reference genomes also aid in delineating the pan-genome (genes present in at least one genome) and the core-genome (genes present in all genomes) of a species or a group of related organisms. This approach unveils genetic diversity within a taxonomic group, shedding light on shared and unique genetic features (Secomandi et al., 2023; Wang et al., 2023). By comparing genomes, it unveils critical insights into the differentiation of gene organisation and pinpoint mutations, insertions, deletions, and their associations with phenotypic traits. This genomic detective work extends to unravelling the genetic underpinnings of disease resistance, shedding light on the genes or pathways responsible for survival (Gong et al., 2023; Zhang et al., 2017).

2.3.3 Intraspecific comparisons

In the field of population genomics, a reference genome serves as an essential tool for delving into genetic diversity within species. A reference genome for a given species significantly facilitates intraspecific comparisons by providing a foundational template for assembling genomes from diverse populations (see 2.2). These intraspecific genome comparisons involve the meticulous scrutiny of individuals within the same species, unveiling a wealth of information. It can reveal genetic differentiation, disparities in gene organisation, mutations, insertions, deletions, and various genomic variations across distinct populations. These

variations often underlie phenotypic traits, offering invaluable insights into the evolutionary adaptations within a species. Pioneering projects like the 1000 Human Genome Project (1000 Genomes Project Consortium et al., 2010) and the 1002 Yeast Genome Project (Peter et al., 2018) have vividly demonstrated the power of intraspecific comparisons in deciphering genetic diversity. Notably, such comparisons can pinpoint specific genes or pathways, empowering the development of more effective disease management strategies using Genome-Wide Association Studies (GWAS) (Uffelmann et al., 2021). GWAS serves as a vital tool for identifying candidate genes associated with distinct phenotypes and adaptations. It contributes significantly to our understanding of pathogenicity, disease resistance, and the development of disease-resistant varieties in both plants and animals (Sánchez-Roncancio et al., 2022).

When dealing with matters related to species conservation, reference genomes become invaluable resources for estimating relative allele and genotype fitness by screening Single Nucleotide Polymorphisms (SNPs), particularly when experimental approaches pose logistical challenges (von Thaden et al., 2020). These reference genomes play a pivotal role in shaping conservation programs, aiding in decisions related to captive breeding, translocation initiatives, and the preservation of biodiversity (Theissinger et al., 2023). Preservation of biodiversity and sequencing of all known species is the main aim of the Earth BioGenome Project (EBP) (Lewin et al., 2022) leading to project such as ERGA (European Reference Genome Atlas) (Mazzoni et al., 2023) dedicated to advancing European biodiversity knowledge, and G10K (Genome 10K Project) (Koepfli et al., 2015) for vertebrates. Sequencing technological advancements have proven instrumental in facilitating genome assembly projects for non-model organisms. Non-model organisms present unique challenges, primarily stemming from the absence of closely related reference genomes that typically aid in genome assembly. The noble crayfish serves as an illustrative example in this regard. Despite its ecological and economic significance, crayfish remains relatively understudied, with only four available genomes to date. Furthermore, when examining the broader taxonomic group of Decapoda, which encompasses crayfish among various other species, the lack of genomic information becomes even more apparent, with only 25 genomes publicly available. Apart from their immediate conservation relevance, Decapoda genomes hold intrinsic scientific

interest due to their distinctive composition. These genomes possess unique attributes that warrant exploration.

Chapter 3 – The crayfish genome, an enigma

Decapoda is a diverse order of crustaceans and encompasses a wide range of species, including the fascinating and economically significant group of crayfish. As seen in Chapter 1, these freshwater arthropods have captured the attention of researchers due to their ecological importance and economic value. However, they are still little explored at genomic level, despite the intriguing characteristics of their genomes and the contribution that reference genomes could make to the protection of threatened crayfish species (see Chapter 2).

In this chapter, I explore and compare the genome organization, gene content and chromosome numbers of presently available decapod genomes. I will then present repetitive elements classification, highlighting their roles in the genomic landscape. Emphasizing their significance in arthropod species, I unravel the contributions of these repetitive elements to the genomic complexity inherent of large genomes. I will then present the challenges in deciphering reference genomes of giant non-model organisms, addressing sequencing technologies and assembly strategies.

3.1 Crayfish and Decapoda genomes

It doesn't count unless it is in a public database (Richards, 2019)

3.1.1 Genome organisation

Crayfish exhibit a remarkable genomic diversity, making them intriguing subjects for genome organization studies with large and variable size. For example, the genome size in the European noble crayfish (*A. astacus*) is estimated to be around 17 Gb and the idle crayfish (*A. biharensis*) around 12 Gb (Theissinger et al., unpublished results). More generally, the estimated genome sizes in the infraorder Astacidea (crayfish and true lobsters) can vary from 3 to 19 Gb (Figure 3-1; Gregory, 2023). Based on data from the Animal Genome Size database, genome size variations are also present among other decapod species (Gregory, 2023). As example, Caridea species (true shrimp) can have a genome size of up to 40 Gb (Figure 3-1). Several factors contribute to genome size variation. Genome size has already been proven to be correlated to animal size in crustacean species in general, but not in decapods (Hessen and

Persson, 2009; Petersen et al., 2019; Sproul et al., 2022; Wu and Lu, 2019). On the other hand, eco-physiological and life-history traits, such as habitat or breathing organs used, have been proven to correlate with genome size in various decapods (Ryan Gregory, 2002; Iannucci et al., 2022). Lastly, the presence of repetitive sequences, such as transposable elements and non-coding DNA, can significantly contribute to genome size (Yuan et al., 2017; Tang et al., 2021).

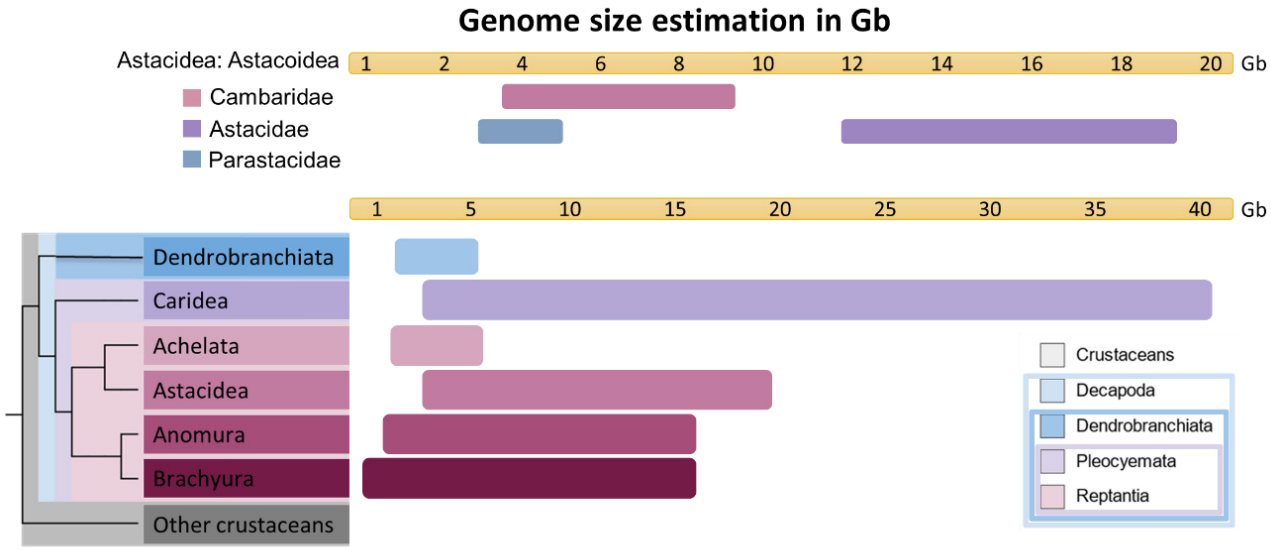

Figure 3-1: Crayfish and Decapoda genome size estimation. Based on the Animal Genome Size Database (Gregory, 2023).

At a larger scale, decapods also present a varying number of chromosomes (reviewed in González-Tizón et al., 2013; Lécher et al., 1995). In the case of Astacidea, there is notable diversity even at the genus level. For instance, the genus *Faxonius* is known to have a varying number of chromosomes, typically ranging from 90 to 250. The largest number of chromosomes is found in the signal crayfish with a 2n of 376 (Figure 3-2; Crandall and De Grave, 2017; Niiyama, 1962). The noble crayfish (*A. astacus*) has a diploid chromosome number of 176 (Figure 3-2; Mlinarec et al., 2011). At the decapod level, Dendrobranchiata, Caridea, and Brachyura exhibit lower chromosome numbers than Astacidea, Anomura, and some Achelata (Figure 3-2). Decapods, including crayfish, generally possess medium to small-sized chromosomes. Ploidy, which refers to the number of sets of chromosomes in an organism's cells, is another intriguing aspect of crayfish genomics. Crayfish exhibit a variation in ploidy, that can have profound effects on crayfish biology, ecology, and evolution. The

triploid marbled crayfish, *P. virginalis*, is sterile and has a clonal reproduction in contrast to the diploid deceitful crayfish (*P. fallax*) from which it originates.

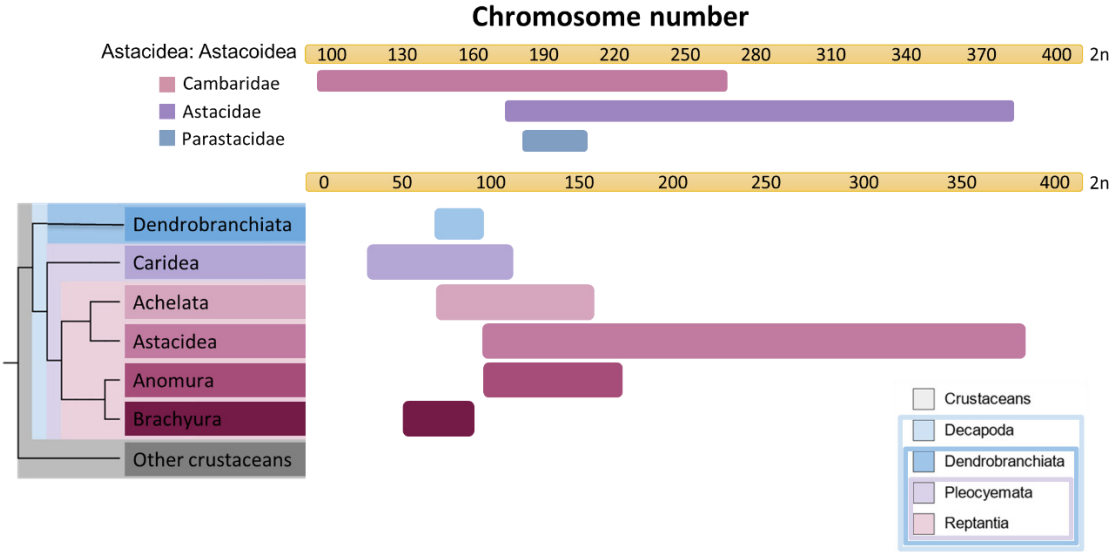

Figure 3-2: Crayfish and Decapoda chromosome number.

Studying the genome organisation of decapods, particularly crayfish, has far-reaching implications. Genomic data is used to unravel the genetic basis of important traits, such as disease resistance, growth rates, and tolerance to environmental stressors (Carroll et al., 2022; Fernandez-Gutierrez and Gutierrez-Gonzalez, 2021; Meyerson et al., 2020). This knowledge can inform breeding programs for aquaculture and conservation efforts for endangered crayfish species, as outlined in section 1.2.5 Conservation genomics for crayfish.

3.1.2 Available genomic sequences

The Decapoda order is home to a diverse range of species, with an estimated count of nearly 15 000 extant species. Currently, only 44 Decapoda species genomes are publicly accessible, and out of these genomes, only 17 have accompanying annotations (Table 3-1). Genomes within the Decapoda order tend to exhibit substantial sizes (Figure 3-1). Yet, the majority of the available assemblies have been constructed utilising short reads and only partially covers the genome (Table 3-1). In the most recently published genomes, the data for estimating genome sizes is generally lacking, but it is reasonable to presume that the assembly sizes might underestimate the genome size. For instance, within the *Panulirus* genus (spiny lobsters), the

twelve available assemblies range from 0.936 to 1.926 Gb, while the two available estimates of genome size reach 3.2 Gb for *Panulirus ornatus* and 5.5 Gb for *Panulirus argus,* the latter having no assembly (Jimenez et al., 2010).

Table 3-1. Available Decapoda genomes.

Order/ Suborder	Organism Name	Assembly accession	Estimate size in Gb (ref)	Ass size (Gb)	Level	Tech	Contig N50 (kb)	Scaffold N50 (kb)	Release date	Annot
Dendrobanchiata	*Penaeus japonicus*	GCF_017312705.1	2.1 (1)	1.7	Sca	SR, LR	132.8	234.9	12/02/2021	Yes
	Penaeus chinensis	GCF_019202785.1	2.6 (2)	1.5	Chr	LR	470.2	36870.7	13/07/2021	Yes
	Penaeus indicus	GCA_018983055.1	2.8 (1)	1.9	Sca	SR, LR	463.4	34408.7	21/06/2021	Yes
	Penaeus vannamei	GCF_003789085.1	2.2 (1)	1.7	Sca	SR, LR	86.9	605.6	16/11/2018	Yes
	Penaeus monodon	GCF_015228065.2	2.2 (1)	2.4	Chr	SR, LR	45.2	44862.1	05/11/2020	Yes
Pleocyemata/ Caridea	*Palaemon carinicauda*	GCA_004011675.1	6.2 (3)	6.7	Sca	SR	0.7	1.0	11/01/2019	Yes
	Pandalus platyceros	GCA_005815305.1	NA	0.1	Sca	SR	1.1	1.5	24/05/2019	No
	Caridina multidentata	GCA_002091895.1	3.2 (4)	1.9	Sca	SR	0.8	0.8	03/04/2017	No
	Macrobrachium nipponense	GCA_015110555.1	4.6 (5)	2.3	Chr	LR	244.3	93504.2	29/10/2020	Yes
	Halocaridina rubra	GCA_037179515.1	NA	1.2	Ctg	LR	81	NA	15/03/2024	Yes
Pleocyemata/ Achelata	*Panulirus ornatus*	GCA_018397875.1	3.2 (6)	1.9	Sca	SR	5.4	8.1	18/05/2021	No
	Panulirus marginatus	GCA_032361885.1	NA	1.3	Sca	SR	1.7	1.8	04/10/2023	No
	Panulirus pascuensis	GCA_032361865.1	NA	1.1	Sca	SR	1.4	1.6	04/10/2023	No
	Panulirus inflatus	GCA_032361765.1	NA	1.3	Sca	SR	2.6	2.9	04/10/2023	No
	Panulirus gracilis	GCA_032361445.1	NA	1.3	Sca	SR	2.3	2.5	04/10/2023	No
	Panulirus guttatus	GCA_032361385.1	NA	1.6	Sca	SR	3.2	3.6	04/10/2023	No
	Panulirus interruptus	GCA_032273725.1	NA	1.4	Sca	SR	2.2	2.5	02/10/2023	No
	Panulirus laevicauda	GCA_032273605.1	NA	1.4	Sca	SR	3.2	3.5	02/10/2023	No
	Panulirus longipes	GCA_032273845.1	NA	1.2	Sca	SR	1.6	1.8	02/10/2023	No
	Panulirus cygnus	GCA_032361485.1	NA	1.0	Sca	SR	1.1	1.2	04/10/2023	No
	Panulirus versicolor	GCA_032361705.1	NA	1.5	Sca	SR	3.9	4.2	04/10/2023	No

	Panulirus homarus	GCA_032361405.1	NA	1.3	Sca	SR	2.6	2.9	04/10/2023	No
Pleocyemata/ Astacidea	*Procambarus virginalis*	GCA_020271785.1	3.5 (7)	3.7	Sca	LR	12.2	144.4	04/10/2021	No
	Procambarus clarkii	GCF_020424385.1	8.5 (8)	2.7	Chr	LR	217.7	17011.5	12/10/2021	Yes
	Cherax destructor	GCA_009830355.1	4.5 (9)	3.3	Sca	SR, LR	80.9	87.2	03/01/2020	No
	Cherax quadricarinatus	GCF_026875155.1	5.0 (10)	5.2	Chr	SR, LR	146.4	45061.5	14/12/2022	Yes
	Homarus americanus	GCF_018991925.1	7.7 (11)	2.3	Sca	SR, LR	133.3	759.6	23/06/2021	Yes
	Homarus gammarus	GCA_958450375.1	4.1 (12)	1.8	Sca	SR, LR	1641.7	1823.2	13/09/2023	No
Pleocyemata/ Anomura	*Paralithodes camtschaticus*	GCA_018397895.1	7.2 (6)	3.8	Sca	SR	5.8	7.0	18/05/2021	No
	Paralithodes platypus	GCA_032716605.1	5.4 (13)	5.0	Chr	LR	334.2	55470.5	17/10/2023	No
	Birgus latro	GCA_018397915.1	6.2 (6)	3.0	Sca	SR	5.3	6.3	18/05/2021	No
	Pagurus hirsutiusculus	GCA_030323965.1	NA	0.8	Sca	SR	0.9	1.3	22/06/2023	No
	Pagurus granosimanus	GCA_030265335.1	NA	0.8	Sca	SR	0.9	1.4	13/06/2023	No
	Pagurus beringanus	GCA_031763525.1	NA	1.0	Sca	SR	1.5	2480.1	20/09/2023	No
	Pagurus longicarpus	GCA_028571265.1	4.8 (14)	0.6	Sca	SR	1.3	1.7	09/02/2023	No
	Coenobita brevimanus	GCA_032717465.1	NA	4.8	Chr	SR, LR	1754.8	42958.6	17/10/2023	No
	Petrolisthes cinctipes	GCA_033782935.1	NA	1.5	Ctg	LR	706.7	NA	17/11/2023	Yes
	Petrolisthesmani maculis	GCA_034508575.1	NA	0.9	Ctg	LR	218.9	NA	15/12/2023	Yes
Pleocyemata/ Brachyura	*Chionoecetes opilio*	GCA_016584305.1	NA	2.0	Sca	SR, LR	149.6	208.1	08/01/2021	Yes
	Portunus trituberculatus	GCF_017591435.1	2.2 (15)	1.0	Chr	SR, LR	4121.4	21793.9	30/03/2021	Yes
	Eriocheir sinensis	GCF_024679095.1	2.2 (15)	1.8	Chr	SR, LR	717.3	16975.5	16/08/2022	Yes
	Callinectes sapidus	GCA_020233015.1	2.2 (15)	1.0	Chr	SR, LR	9.3	18846.6	04/10/2021	No
	Metacarcinus magister	GCA_029783475.1	NA	0.7	Sca	SR, LR	330.1	16129.4	17/04/2023	No
	Scylla paramamosain	GCF_035594125.1	NA	1.2	Chr	LR	11400.0	23600	12/01/2024	Yes

Ass: assembly, SR: short reads, LR: long reads, Chr: chromosome, Sca: scaffold, Ctg: contig, NA: not available. (1) (Swathi et al., 2018), (2) (Meng et al., 2021), (3) (Yuan et al., 2017), (4) (Kawato et al., 2021), (5) (Jin et al., 2021), (6) (Veldsman et al., 2021), (7) (Gutekunst et al., 2018), (8) (Shi et al., 2018), (9) (Austin et al., 2022), (10) (Tan et al., 2020), (11) (Polinski et al.,

2021), (12) (Deiana et al., 1999), (13) (Tang et al., 2021), (14) (Rheinsmith et al., 1974), (15) (Liu et al., 2016).

Genomes assembled using long reads, or a combination of short and long reads, exhibit a superior level of contiguity, as substantiated by statistical analyses (Table 3-1; Hotaling et al., 2023). However, it is essential to recognize that, when taking the genome sizes into account, all available decapod assemblies suffer from fragmentation (Table 3-1). Only ten genomes manage to attain a chromosomal level of contiguity. The noticeable disparity between estimated sizes and assembly sizes (Table 3-1) can be primarily attributed to the prevalence of repetitive elements (REs), which introduce a high level of complexity into the assembly process (Pop, 2009; Tørresen et al., 2019; Treangen and Salzberg, 2012). The presence of such REs often results in assembly loops, complicating the resolution process. Consequently, assembly procedures frequently exclude these intricate regions, leading to the production of shorter contigs and scaffolds than initially anticipated. Moreover, when dealing with short repeated sequences, known as satellite DNA, assembly programs tend to truncate these sequences due to their inability to precisely determine the exact number of repetitions. This truncation further contributes to the differences between the estimated and assembled sizes (Table 3-1). As a result, the contigs and scaffolds produced may fall short of expected lengths due to these inherent challenges. The intricate landscape of REs presents a hurdle in achieving comprehensive and accurate genome assemblies in decapods.

Among decapod genomes, only four crayfish genomes are currently available, originating from the genus *Procambarus* and *Cherax*. This scarcity underscores the limited knowledge in the field of crayfish genomics, with only two of these genomes featuring published annotations. Remarkably, none of the accessible genomes represent European crayfish species. Moreover, the estimated genome size of crayfish stands among the largest of decapods, and the disparities between estimated and assembled genome sizes are conspicuous. Additionally, the genome from the family Astacidae is estimated to be notably larger than those from other families from Astacoidea superfamily (Figure 3-1). Hence, there is a compelling interest in investigating European species of Astacidea to enhance our understanding of crayfish genomes and utilise this knowledge for effective management and conservation efforts.

3.1.3 Gene content

With only ten genomes having been annotated, Decapoda remains relatively understudied in terms of protein-coding gene content. A synthesis of data from various studies indicates a general stable number of protein-coding genes, hovering around 25 000, accompanied by an average of five to six exons per gene (Chen et al., 2023; Cui et al., 2021; Huerlimann et al., 2022; Kawato et al., 2021; Lv et al., 2022; McGrath et al., 2016; Polinski et al., 2021; Ren et al., 2022; Tan et al., 2020; Tang et al., 2020a, 2020b; Uengwetwanit et al., 2021; J. Wang et al., 2022; Q. Wang et al., 2022; Xu et al., 2021; Zhang et al., 2019). This concords with the number of genes found in other crustaceans since a recent analysis of 66 genomes (Yuan et al., 2023) reported a generally constant gene count of approximately 25,000 genes per genome. This is in the middle range of the number of genes found in arthropods, which varies from 10,000 to 36,000 (Thomas et al., 2020). It is noteworthy that, despite the large genome size estimation in decapods and the identification of numerous duplicated sequences in penaeid shrimp, whole genome duplication was solely observed in *Macrobrachuim nipponense* among the decapods (Jin et al., 2021). However, the overall number of 25,000 protein-coding genes in decapods must be treated with caution, as it conceals disparities between species and within the same species, depending on the study. For instance, in *Portunus trituberculatus* the gene count was calculated at 16 796 (Tang et al., 2020b), being the lowest in crustaceans, and at 19 981 in another study (Lv et al., 2022), aligning more closely with counts observed in other decapod and crustacean species. Conversely, the gene count in *Penaeus monodon* was estimated at around 26 000 to the upper limit of 30 000, as reported by different studies (Huerlimann et al., 2022; Uengwetwanit et al., 2021).

A comparative analysis reveals a unique methylation pattern in decapods, and more generally in crustaceans, with increased methylation in introns compared to exons, contrary to the conserved methylation in exons observed in other arthropods (Lewis et al., 2020). This distinctive methylation profile was previously noted in the marbled crayfish (*P. virginalis*) (Gatzmann et al., 2018). Comparative analyses also suggest the presence of numerous lineage-specific genes in decapod species. It's noteworthy, however, that these comparisons often involve only a limited number of closely related species, typically fewer than three decapods (Cui et al., 2021; Lv et al., 2022; Q. Wang et al., 2022; Xu et al., 2021; Zhang et al., 2019). The

limited genomic annotations in decapods underscore the need for further research to unravel the intricacies of gene content conservation in this diverse crustacean order.

3.2 Repetitive elements in large genomes

Repetitive elements in DNA are like biological time capsules, revealing the history of genome evolution over millions of years. (Smit, 1999)

3.2.1 Classification and structure of repetitive elements

Decapods present large genome size (Figure 3-1), however, gene number around 25 000 suggest that repetitive elements (REs) have a determining role within decapod genomes. REs are genomic sequences characterized by their high similarity to other sequences within the same genome, and their multiple copies. There are two major categories of REs: tandem repeats, also called satellite DNA (SatDNA), and transposable elements (TEs), also known as interspersed repeats (Jurka et al., 2007).

SatDNA consists of tandemly repeated nucleotide patterns, known as repeat units, that are repeated a variable number of times (Figure 3-3; Garrido-Ramos, 2017). SatDNA can be further categorized into three subtypes based on the length of their repeat units: microsatellites (units < 10 bp), also known as simple repeats or Short Tandem Repeats (STRs), distributed across the entire chromosome; mini-satellites (units: 10-100 bp), primarily found in telomeres; and satellites (units > 100 bp), predominantly located in centromeres. Genomes may contain various families of satDNA, but typically, only one or a few predominant satellite families are present (Macas et al., 2007; Ruiz-Ruano et al., 2016).

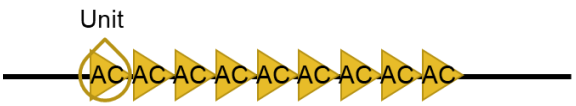

Figure 3-3: Satellite DNA. SatDNAs are composed of DNA sequences called units (triangles) tandemly repeated.

TEs, colloquially referred to as "jumping genes", can account for a substantial portion of eukaryotic genomes. TEs are considered selfish parasitic DNA elements, as they encode proteins necessary for their own replication and dissemination or depend on host machinery. TEs can exist in transcriptionally active state and propagate through the genome or remain as

inactive part of the genome. Based on their transposition mechanisms, transposons can be classified into two categories: retrotransposons (Class I) and DNA transposons (Class II) (Table 2; Kojima, 2019; Wicker et al., 2007). Class I TEs employ an RNA-mediated mechanism for transposition, often described as "copy-and-paste" (Figure 3-4 A). The presence or absence of Long Terminal Repeats (LTR) at their ends subdivides them into subclasses (Table 3-2). Each subclass may consist of various repetitive element families. In non-LTR elements, LINEs encode a reverse transcriptase, rendering them autonomous, while SINEs are non-autonomous and rely on the machinery of autonomous elements for mobility. LTR elements can exist in both autonomous and non-autonomous forms. Class II TEs transpose through a DNA-based mechanism and do not require an RNA intermediate (Table 3-2, Figure 3-4 B). They are translocated following a “cut and paste” method, leaving only residues at the first location corresponding to TIR or direct repeats. However, during replication we can observe a “copy and paste” translocation allowing the spread of the TE. They can be either autonomous, encoding a transposase or a helicase, or non-autonomous.

Table 3-2. Transposable elements classification.

Class	Features	Order	Comments
Class I	Presence of Long Terminal Repeat (LTR)	LTR	Framed by LTR, target for DNA sequence rearrangement (similar to class II subclass 2 Polintons).
		DIRS	Framed by inverted repeats.
		Penelope	Framed by a particular type of LTR (pseudo-LTRs).
	Absence of LTR	Long Interspersed Nuclear Elements (LINEs)	Encodes a reverse transcriptase, autonomous elements.
		Short Interspersed Nuclear Elements (SINES)	Non-autonomous elements that rely on the machinery of autonomous elements for mobility.

Class II Subclass 1	Short Terminal Inverted Repeat (TIR)	TIR	Transposition mechanism use cleaving sites at the TIR on both DNA strand and exploit gap repair mechanism to be translocated.
	Short direct repeats	Crypton	
Class II Subclass 2		Helitron	Transposes with rolling-circle mechanism by cutting only one strand.
		Polintons (also called Maverick)	Framed by long TIR and excise a single strand and then use extrachromosomal replication to be integrated to a new site.

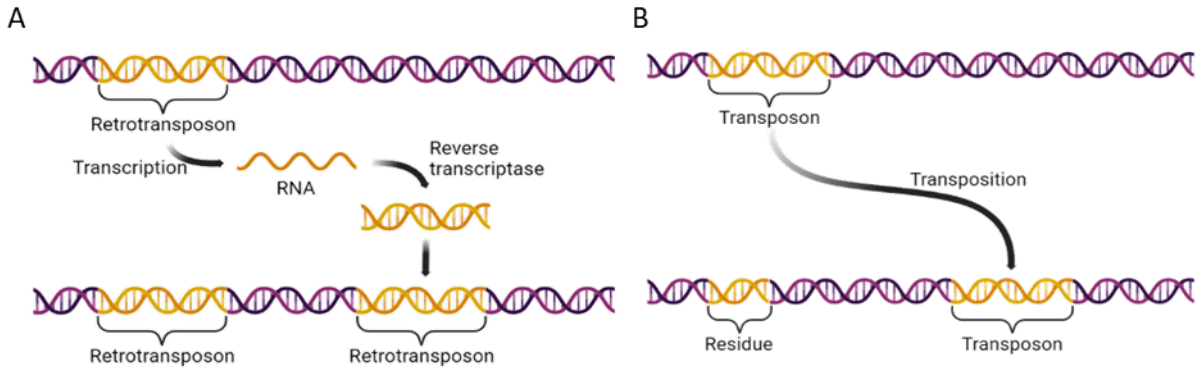

Figure 3-4: Transposable elements. A. Class I: retrotransposons. Transposition is RNA-mediated and often described as "copy-and-paste". B. Class II: DNA transposons. Transposition is DNA-based and do not require a RNA intermediate. Often described as “cut-and-paste”, duplication occurs when transposition happens during DNA replication.

TEs typically encode proteins essential for their autonomous transposition. However, over the course of genome evolution, insertions and deletions (indels) can occur within TEs, resulting in incomplete versions that can no longer encode transposition enzymes. These non-autonomous transposable elements are incapable of initiating new integration events independently. Some of these elements may lose internal sequences, giving rise to miniature versions of the original TE, such as MITEs (miniature inverted-repeat transposable elements) within Class II. Nevertheless, the original elements' boundaries are preserved, allowing the machinery of autonomous elements to recognize and transpose them (Yang et al., 2009).

3.2.2 Roles of repetitive elements

REs, once dismissed as "junk DNA," have emerged as crucial contributors to genomic structure and function (Kim et al., 2012). Recent research has shown their substantial impact on evolution, gene expression regulation, and the induction of genetic variation (Bourque et al., 2018; Liao et al., 2023; Shapiro and Sternberg, 2005). The conservation of repetitive sequences throughout evolution and their high frequency suggests vital biological functions related to genomic stability, architecture, and evolutionary dynamics. REs play a pivotal role in genome evolution, driving genetic diversity through the promotion of rearrangements, duplications, and mutations over evolutionary timescales.

SatDNAs have been demonstrated to hold specific functions in gene and genome regulation, implicated in chromosome organization, pairing, and the formation of the centromere locus and structure of the telomeres (Plohl et al., 2008, 2012). At centromeres, satDNAs play a pivotal role in chromosome segregation during cell division, forming the kinetochore, a structure essential for the proper distribution of genetic material to daughter cells. Furthermore, satDNA participates in epigenetic regulation, influencing heterochromatin establishment and modulating gene expression in response to stress (Pezer et al., 2012; Biscotti et al., 2015). The expansion or contraction of these satDNAs has significant impact on gene function. Moreover, satDNAs have been proven to drive genome plasticity and promote adaptive evolution (Yuan et al., 2021b). SatDNAs, including microsatellites and minisatellites, serve as valuable genetic markers, frequently employed in genetic studies for individual identification, population genetics, and forensic analysis.

TEs, as powerful drivers of genomic diversity, contribute to the evolution of species over time. Actively transposing of TEs within a host genome can enhance genome plasticity, leading to gene rearrangements. Recombination events between homologous regions dispersed by related TEs at distant genomic positions can result in deletions, duplications, and inversions (Bennetzen and Wang, 2014; Bourque et al., 2018; Deininger et al., 2003). TEs, when inserted into genes or coding regions, can alter gene expression, potentially causing deleterious effects, such as diseases, or neutral effects on the host (Barrón et al., 2014; Burns and Boeke, 2012; Kim et al., 2012). TEs can also influence gene expression and the regulation of nearby genes by inserting near or within genes (Figure 3-5; Cowley and Oakey, 2013). They may provide new

regulatory elements or disrupt existing ones. Additionally, TEs have been implicated in the creation of new genes and functional elements, serving as raw material for the evolution of novel genetic functions (Graveley, 2001). In response to environmental stressors, organisms may exhibit increased TE activity in their genomes (Lanciano and Mirouze, 2018). All REs, but especially TEs, contribute to the dynamic expansion or contraction of genome sizes, influencing the overall genomic content. The interplay of these repetitive elements shows their multifaceted role in shaping the complexity and adaptability of genomes across diverse organisms.

3.2.3 Repetitive elements as key components of Decapoda genomes

REs constitute a substantial portion of eukaryotic genomes, with their prevalence exemplified by their representation exceeding 45% in the human genome (Koning et al., 2011). Notably, in certain plants such as maize, REs can comprise over 90% of the genome (SanMiguel et al., 1996). In sharp contrast, human gene exons encompass only around 3% of the genome, and protein-coding sequences constitute a mere 1% (Schumann et al., 2010). Among Arthropoda, TEs are mostly studied in insects and demonstrate a wide distribution, with proportions varying from 1% to 60%, according to studies by Peterson et al., (2019) and Wu and Lu, (2019), encompassing 73 and 14 arthropod genomes, respectively. A more extensive study involving over 600 insects revealed TE proportions ranging from 1% to 80% (Sproul et al., 2022). Intriguingly, the assembly of the Antarctic krill, *Euphausia superba*, unveiled that 92% of its genome is constituted of REs, with 78% of them being TEs, implying the potential for exceptionally high RE content in arthropods (Shao et al., 2023). In Arthropoda, studies generally show the prevalence of DNA transposons, with exceptions such as in *Drosophila* species and *Daphnia pulex*, where LTR elements are more common (Petersen et al., 2019; Sproul et al., 2022). LINE elements exhibit widespread distribution across all arthropods, dominating in Hemiptera, Lepidoptera, and the crustacean *Hyallela azteca* (Petersen et al., 2019; Sproul et al., 2022). In contrast, SINEs are relatively underrepresented in arthropods (Petersen et al., 2019; Sproul et al., 2022). Despite efforts to characterize TEs in insect genomes, the diversity observed may not be representative for other arthropod species (Petersen et al., 2019; Sproul et al., 2022; Wu and Lu, 2019).

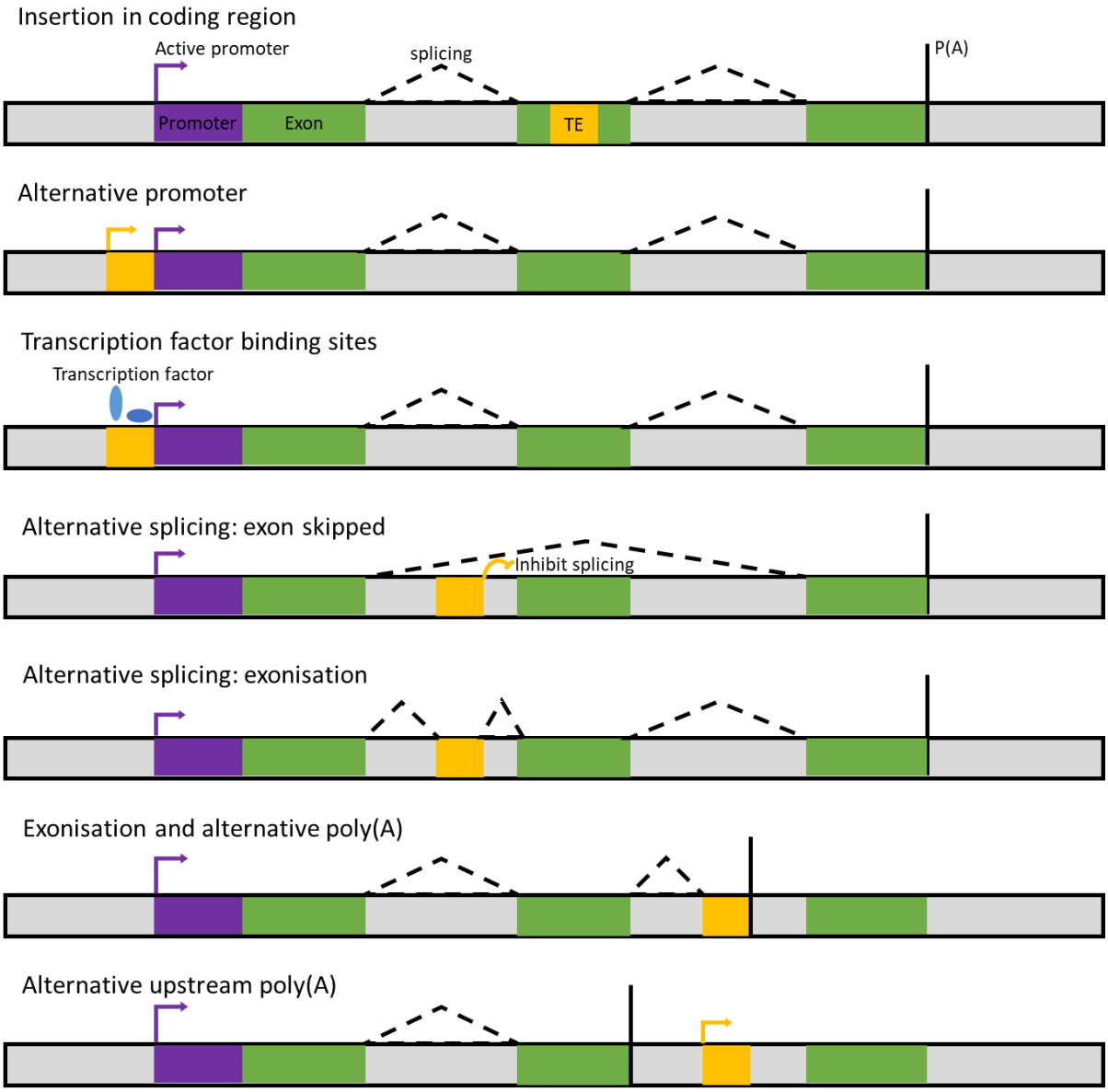

Figure 3-5: Impact of transposable elements on the transcriptome. Insertion of a TE within a coding region can disrupt it, leading to the inhibition of protein synthesis. Insertion into a promoter region can generate an alternative promoter or introduce binding sites for transcription factors, thereby giving rise to tissue- or stage-specific proteins. Inserted into an intronic region, they can induce alternative splicing. This alternative splicing may hinder the recognition of a splice acceptor site by the splicing machinery or incorporate the TE into the mature transcript. Insertion in the intronic region may prompt alternative polyadenylation (poly(A)) either through the provision of an alternative polyadenylation signal or by interfering with host gene transcription and causing upstream polyadenylation due to promoter activity. Adapted from Cowley and Oakey, 2013.

Despite the pervasive presence of REs in decapods they remain mostly underexplored. Reports on REs included in assembly publications of decapod genomes reveal a variable proportion of repetitive elements, ranging from 8% to 82% (Austin et al., 2022; Bachvaroff et al., 2021; Gutekunst et al., 2018; Jin et al., 2021; Katneni et al., 2022; Kawato et al., 2021; Liu

et al., 2022; Polinski et al., 2021; Tan et al., 2020; Tang et al., 2020b; Uengwetwanit et al., 2021; Veldsman et al., 2021; Q. Wang et al., 2021; Xu et al., 2021; Yuan et al., 2021a; Zhang et al., 2019). This variable proportion of REs could explain the marge variation in decapod genome sizes (Figure 3-1). Tan and colleagues annotated the repeatome of eight decapod species, estimating repetitive content between 27% and 50%, with LINEs being predominant in most genomes, except for *Penaeus vannamei*, which exhibited a higher abundance of DNA transposons (Tan et al., 2020). However, number of genomes in this study is limited and comparative studies on SatDNAs in decapods are lacking. Considering the large size of the noble crayfish (*A. astacus*, 17 Gb, Theissinger et al., unpublished results) and the propensity of numerus REs in decapods, we can expect not only a high proportion of TEs, but also a large proportion of satDNAs (Boštjančić et al., 2021).

3.3 Reference genomes of giant non-model organism challenges

Because research publications mostly describe successful experiments […] (Adema, 2021)

3.3.1 Choice of sequencing technologies

As the landscape of genome sequencing technologies evolves at a rapid pace, it is noteworthy that these advancements are primarily tailored for model organisms and genomes of small to medium sizes (≈3 Gb). The various advantages of each sequencing platform have already been explored in 2.1.3. In the context of European crayfish species, such as the noble crayfish (*A. astacus*) and the idle crayfish (*A. biharensis*), exclusively employing short-read sequencing is impractical due to the substantial likelihood of a significant proportion of REs, given their large genome size. Consequently, REs would remain unresolved, posing a significant challenge. Furthermore, the computational costs associated with generating the voluminous data required for the assembly of such large genomes are prohibitively high. The adoption of long-read sequencing becomes imperative, albeit at a higher cost, to effectively address the challenges posed by REs. However, the inherent lower accuracy of long reads necessitates multiple passes for error correction, i.e. HiFi reads for PacBio or corrected reads for Nanopore. The presence of REs in large genomes compromises DNA integrity, making it vulnerable to fragmentation during sequencing, resulting in sequences that fall short of anticipated lengths. The intricate nature of these repeats also introduces technical sequencing challenges, where

the polymerase (in the case of PacBio) or the membrane (for Nanopore) may lose track when encountering extended regions with identical sequences. This phenomenon has been observed in crustaceans and molluscs (Adema, 2021; Angthong et al., 2020; Athanasio et al., 2016).

To strike an optimal balance between accuracy and coverage, a prevailing strategy involves the amalgamation of both short and long reads. Long reads are particularly adept at resolving REs and generating elongated contigs, while short reads contribute to precision and ensure high coverage throughout the genome. This hybrid approach maximizes the strengths of each sequencing technology, addressing the unique complexities posed by large and repetitive genomes of organisms such as European crayfish.

3.3.2 Choice of assembly strategies

In the realm of European crayfish species, the absence of reference genomes presents a challenge, given that the closest available genomes originate from a distinct family. The adoption of an *ab initio* assembly emerges as the only viable approach.

The assembly of such large genomes demands substantial quantities of data and significant computational capacity. Complicating matters, several assemblers are optimized for small to medium-sized genomes, rendering them unsuitable for handling such voluminous datasets. To surmount this challenge, a stepwise assembly approach has been proposed for assembling large genomes such as used for the Siberian larch (Table 3-3; Kuzmin et al., 2019). To produce the assembly, only paired-end and mate-pair Illumina reads of various insert sizes were used, following a stepwise assembly using CLC Assembly Cell (https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/enterprise-ngs-solutions/clc-server-command-line-tools/). This involves partitioning the entire pool of reads into multiple batches, each independently assembled to manage the data volume. The initially assembled contigs are then consolidated and subjected to further assembly, resulting in the completion of the assembly process. This technique aims to introduce minimal bias and facilitates the assembly of large datasets using tools that are typically not optimized for such extensive genome assembly.

As established in 3.3.1, considering large genome size and propension for high number of REs in decapod species, the use of long reads in addition to short reads appear to be imperative. Three predominant strategies for *ab initio* assembly of long and short reads are widely acknowledged: (1) short-read assembly paired with long-read scaffolding, such as used for the assembly of the onion (Table 3-3; Finkers et al., 2021) using MaSuRCA (Zimin et al., 2013) and DBG2OLC (Ye et al., 2016); (2) long-read assembly followed by short-read correction that have been used for the assembly of the axolotl (Table 3-3; Nowoshilow et al., 2018) using MARVEL (Nowoshilow et al., 2018), followed by correction with Illumina paired-end reads *via* Pilon (Walker et al., 2014); (3) hybrid assembly approach as used for the assembly of the coast redwood genome (Table 3-3; Neale et al., 2022) using MaSuRCA (Zimin et al., 2013) in hybrid mode.

Table 3-3: Giant genome assembly statistics.

Common name	Siberian larch	Onion	Axolotl	Coast redwood
Scientific name	*Larix sibirica*	*Allium cepa*	*Ambystoma mexicanum*	*Sequoia semperivens*
Long reads (Gb)		116	1000	NA
Coverage of long reads		7	32	22
Short reads (Gb)	1570	769	262	3238
Coverage of short reads	92	48	7	122
Estimate size (Gb)	12	16.4	32	31.5
Assembly size (Gb)	12.3	14.9	32.4	26.5
Nb of scaffolds (thousands)	11000	92.9	125.7	393.4
Scaffold N50 (Kb)	6.4	454	3000	44944.4
REs (%)	80	72.4	64.6	70

(Keinath et al., 2015)

In the case of giant genomes, such as that of noble crayfish (*A. astacus*), long-read or hybrid assembly are preferable to address issues arising from REs. Assembling short reads initially would demand more computational resources and time due to the complexities associated with resolving REs, making the assembly of long reads a more favourable approach. In addition, the deployment of large-scale scaffolding technologies (see 2.2.1) warrants consideration. Optical mapping, for instance used for the assembly of the onion, axolotl, and coast redwood genomes, significantly enhances assembly contiguity by organizing and orienting contigs into larger scaffolds, particularly beneficial in resolving complex regions.

Simultaneously, Hi-C methodology, also used for the assembly of the coast redwood, adds spatial context between contigs, providing deeper insights into interactions within genome sequences. However, a prerequisite for these technologies is the availability of a reasonably robust genome assembly as a starting point that is a challenging step for giant and repetitive genomes.

CONTRIBUTIONS

Chapter 4 – Unveiling the Repetitive Landscape of Decapoda

4.1 Introduction

4.1.1 Selection of Decapoda genomes

The genome size of the noble crayfish, *Astacus astacus*, from which we aim to generate an assembly, is estimated to be 17 Gb (Theissinger et al., unpublished results). Within the Astacidae family, to which the noble crayfish belongs, genome sizes vary from 12 Gb to 19 Gb, while other crayfish families range from 2.4 Gb to 9.5 Gb (see Chapter 3). More generally, decapod species exhibit a wide range of genome size estimates, ranging from 1 Gb (Brachyura, crabs) to 40 Gb (Caridea, true shrimp). To address the significant variation in genome size between the Astacidae noble crayfish and other crayfish families and more broadly among decapods, our focus turned to the study of REs.

At the time of the study (last access 22 May 2022), only 22 genomes of decapods were publicly available, with estimated genome sizes ranging from 1 Gb (Brachyura, crabs) to 8.5 Gb (Astacidea, crayfish and lobsters) Gb. Most of these genomes are largely fragmented with the scaffold N50 varying from 0.7 kb to 4.1 Mb (see Chapter 3). The Busco score (Manni et al., 2021), calculating the percentage of genes present in an assembly based on a database of known genes present in a clade, varies from 0.2% to 96.6% completeness. Given the high fragmentation of these genomes and to retain as much taxonomic diversity as possible, we set the threshold at 40% completeness, except for *Caridina multidentata* with a Busco score of 25%. This lower threshold was deemed adequate for our analysis, as we focused on repetitive elements rather than gene content. Two Caridea genomes failed to meet the 25% completeness threshold and were therefore excluded from our study.

Despite the importance of repetitive elements (see Chapter 3), their study within decapod species remains limited, with only one comparative study involving eight decapod species (Tan et al., 2020). Apart from this study, RE annotation in decapods is typically provided within genome assembly publications, if such analysis has been conducted. Consequently, studies on REs often employ different protocols and tools which tend to detect different REs, making comparisons of the RE landscape challenging. Hence, the necessity arose for a new annotation of REs and the adoption of a standardized approach to their identification.

4.1.2 Repeat annotation strategy: existing programs and resources

Annotation of REs often relies on public databases of REs. Some of these databases are specific to certain clades, such as ACLAME (Leplae et al., 2004) for bacteria, archaea plasmid, and virus or APTEdb (Pedro et al., 2021), specific to plants. There is a database specific to transposable elements in arthropods called ArTEDB (Wu and Lu, 2019), that then excludes all satDNAs. There are also more general databases, such as the well-known RepBase database (Bao et al., 2015), and the Dfam database (Storer et al., 2021). For well-studied genomes, such as for model species, some tools based on sequence homology use databases of already known REs, such as RepeatMasker (Smit et al., 2013) that use the Dfam database by default, and CENSOR (Kohany et al., 2006) designed for RepBase. These tools compare genomic sequences to known repetitive elements, annotating regions of significant similarity and masking genomes for downstream analyses.

For less explored taxonomic lineages lacking representative RE sequences in databases, *ab initio* annotation tools such as RepeatModeler (Flynn et al., 2020) and RepeatExplorer (Novák et al., 2013) are employed. RepeatModeler identifies potential REs in genomic sequences, clusters them into families based on sequence similarity, aligns sequences within each family to build consensus sequences, and then classifies these consensus sequences into different categories based on their structural and sequence characteristics. RepeatModeler also provides an option for *de novo* LTRs detection through their structural characteristics, such as the presence of long terminal repeats and target site duplications, during the clustering and consensus sequence construction process. This enhances the sensitivity and accuracy of RepeatModeler in identifying LTRs. On the other hand, RepeatExplorer use raw reads and not assemblies to construct a graph based on k-mer counts, identifies densely connected regions representing clusters of similar sequences, and then annotates these clusters based on sequence composition and similarity to known repeats.

The general approach for RE annotation is to use *ab initio* tools to predict REs, such as RepeatModeler, and then annotate REs using newly identified REs as database with RepeatMasker for example. In addition, a database of known elements, such as RepBase, is generally used in addition to the newly identified RE database. We used this general approach

to annotate REs in 232 soil invertebrate genomes as part of the MetaInvert project (Supplementary 1; Collins et al., 2023).

4.1.3 Originality of the chosen approach

In the context of Decapoda genomes, I opted to create a comprehensive database of predicted repeats amalgamated *ab initio* species-specific REs identified in each species in decapods prior to annotation. This initiative was particularly crucial for decapod species, which are underrepresented in repeat studies, making homology-based identification insufficient. By constructing a database composed of each species-specific REs, reference sequences tailored to each species was generated, facilitating annotation transferability even to distantly related decapods.

For *ab initio* identification of repeats, RepeatExplorer (for satDNAs annotation only) and RepeatModeler was employed. This choice was guided by the significance of tandem repeats in decapods and the inefficiency of RepeatModeler for satellite repeats. This is because satDNAs are often underrepresented in assemblies and are so not detectable by RepeatModeler but could still be present in reads and identified by RepeatExplorer. Subsequently, the database was integrated with RepBase to augment it with well-known and annotated repeats. This extended database was then employed in conjunction with the homology-based annotation tool RepeatMasker, which also utilises the Dfam database. The resulting annotations were applied across all genomes (see methodology).

4.2 Publication

genes

MDPI

Article

Abundance and Diversification of Repetitive Elements in Decapoda Genomes

Christelle Rutz [1], Lena Bonassin [1,2,3], Arnaud Kress [1], Caterina Francesconi [2,3], Ljudevit Luka Boštjančić [1,2,3], Dorine Merlat [1], Kathrin Theissinger [2,†] and Odile Lecompte [1,*,†]

[1] Department of Computer Science, ICube, UMR 7357, University of Strasbourg, CNRS, Rue Eugène Boeckel 1, 67000 Strasbourg, France; christelle.rutz@etu.unistra.fr (C.R.); bonassin@unistra.fr (L.B.); akress@unistra.fr (A.K.); luka.bostjancic@senckenberg.de (L.L.B.); dorine.merlat@etu.unistra.fr (D.M.)
[2] LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Senckenberg Biodiversity and Climate Research Centre, Georg-Voigt-Str. 14-16, 60325 Frankfurt am Main, Germany; francesconi@uni-landau.de (C.F.); kathrin.theissinger@senckenberg.de (K.T.)
[3] Department of Molecular Ecology, Institute for Environmental Sciences, Rhineland-Palatinate Technical University Kaiserslautern Landau, Fortstr. 7, 76829 Landau, Germany
* Correspondence: odile.lecompte@unistra.fr
† These authors contributed equally to this work.

Abstract: Repetitive elements are a major component of DNA sequences due to their ability to propagate through the genome. Characterization of Metazoan repetitive profiles is improving; however, current pipelines fail to identify a significant proportion of divergent repeats in non-model organisms. The Decapoda order, for which repeat content analyses are largely lacking, is characterized by extremely variable genome sizes that suggest an important presence of repetitive elements. Here, we developed a new standardized pipeline to annotate repetitive elements in non-model organisms, which we applied to twenty Decapoda and six other Crustacea genomes. Using this new tool, we identified 10% more repetitive elements than standard pipelines. Repetitive elements were more abundant in Decapoda species than in other Crustacea, with a very large number of highly repeated satellite DNA families. Moreover, we demonstrated a high correlation between assembly size and transposable elements and different repeat dynamics between Dendrobranchiata and Reptantia. The patterns of repetitive elements largely reflect the phylogenetic relationships of Decapoda and the distinct evolutionary trajectories within Crustacea. In summary, our results highlight the impact of repetitive elements on genome evolution in Decapoda and the value of our novel annotation pipeline, which will provide a baseline for future comparative analyses.

Keywords: transposable elements; satellite DNA; Crustacea; annotation; evolution; genome size; library

Citation: Rutz, C.; Bonassin, L.; Kress, A.; Francesconi, C.; Boštjančić, L.L.; Merlat, D.; Theissinger, K.; Lecompte, O. Abundance and Diversification of Repetitive Elements in Decapoda Genomes. *Genes* **2023**, *14*, 1627. https://doi.org/10.3390/genes14081627

Academic Editor: Antonio Figueras

Received: 7 July 2023
Revised: 5 August 2023
Accepted: 12 August 2023
Published: 15 August 2023

1. Introduction

With over 15,000 living species, Decapoda represents a diverse order of Crustacea that includes lobsters, crayfish, crabs, prawns, and shrimps [1]. They are a crucial component of marine and freshwater ecosystems [2,3]. The Decapoda order originated around 455 million years ago, in the Late Ordovician, and is divided into two suborders: the Dendrobranchiata (commonly known as prawns) and the Pleocyemata. The latter encompasses Caridea (swimming shrimps) and a crawling/walking group called Reptantia that consists of Achelata (spiny lobsters), Astacidea (true lobsters and crayfish), Anomura (hermit crabs), and Brachyura (short-tailed crabs) [4].

Decapoda are characterized by highly variable genome sizes. According to the Animal Genome Size Database (https://www.genomesize.com, accessed on 17 May 2022), genome size estimates range from 2.3 Gb for *Penaeus duorarum* to 5.1 Gb for *Aristaeomorpha foliacea* in the Dendrobranchiata suborder. In Pleocyemata, particularly in the Caridea infraorder,

genome size variations are even more striking, with estimates ranging from 3.2 Gb for *Antecaridina* sp. to 40 Gb for *Sclerocrangon ferox*. Freshwater crayfish (Astacidea infraorder) also display substantial genome size variations, ranging from 2 to 6 Gb in Cambaridae and Parastacidae families. Recent genome size estimates for the noble crayfish *Astacus astacus* and the narrow-clawed crayfish *Pontastacus leptodactylus*, both representatives of the Astacidae family, reach 17 Gb (K. Theissinger, unpublished results) and 18.7 Gb [5], respectively. Decapoda also displays high variation in the number of chromosomes. The number of chromosomes in the Dendrobranchiata suborder is mainly at a 2n of 88 (reviewed in [6,7]), while this number can explode in Pleocyemata species to a 2n of 376 for the Astacidea *Pacifastacus leniusculus* [8,9].

Variations in genome sizes are usually attributed to the presence of repetitive elements (REs), which can represent the major part of the genome in some eukaryotic species [10]. A high proportion of REs can greatly complicate genome sequencing and can lead to fragmented and incomplete assemblies [11–13]. This may explain the notorious difficulties encountered in the sequencing of large Decapoda genomes, with only eight assemblies available at the chromosome level. To date, the relationship between the genome size and repeat content, and the impact of REs on genome evolution, remain poorly studied in Crustacea.

The role of REs can be diverse (reviewed in [14]). They can affect transcription and regulation at transcriptional and post-transcriptional levels. Through their ability to act as signals to locate and process information stored in coding sequences, they can influence damage repair, DNA restructuring, chromatin and nuclear organization, and cell division. REs can be classified into two types: tandem repeats (satellite DNA, satDNA) and transposable elements, TEs, also known as interspersed repeats [15].

SatDNAs consist of tandemly repeated patterns of nucleotides, called repeat units (monomers) [16]. Different satDNA families are present in the genome, with usually only one or a few predominant families [17–20]. SatDNAs can have specific roles in gene and genome regulation, such as chromosome organization, pairing, and segregation formation of the centromere locus [21,22], in epigenetic regulation of heterochromatin establishment, and modulation of gene expression in response to stress [23,24]. In Crustacea, some SatDNA transcripts can have an impact on the inter-molt stage [25]. Despite their importance, the distribution patterns, percentage, and copy number of satDNAs are not yet fully explored in Crustacea.

Transposable elements (TEs) are mobile elements known to participate in DNA replication and cause gene rearrangements that can confer new functional properties [26–29]. Deletions, duplications, and inversions can be caused by recombination events between homologous regions dispersed by related TEs at distant genomic positions. When they are inserted into genes or coding regions, TEs can alter gene expression and may produce deleterious effects, such as diseases, or neutral effects on the host [28,30–32]. Organisms living in challenging environmental conditions can have more TEs in their genome, increasing genome plasticity to respond to stress factors [33]. TEs can be divided into two classes based on their replication mechanisms: Class I elements transpose with RNA-mediated mechanisms (retrotransposons), while in Class II the transposition mode is DNA-based (DNA transposons) [34–37]. In Class I, LTR retrotransposons and Penelope-like elements are characterized by Long Terminal Repeat (LTR). DIRS are bound by direct or inverted repeats. Finally, LINEs (long interspersed nuclear elements) and SINEs (short interspersed nuclear elements) are retrotransposons that do not have terminal repeats but a polyA tail at the 3′ end. Unlike LINEs, SINEs evolved from non-coding RNA genes and are non-autonomous. Class II can be divided into two subclasses. Subclass 1 includes TIR and Crypton elements, while subclass 2 includes Helitrons and Mavericks. Apart from SINEs, most TEs encode proteins that are necessary for their transposition in an autonomous way. However, accumulation of mutations can lead to incomplete versions of TEs that no longer encode transposition enzymes. The identification of these truncated alternatives represents a particular challenge for automated annotation pipelines.

Currently, there are several pipelines available for annotation of REs. The most commonly used tools are RepeatModeler2 [38] and RepeatMasker [39]. However, a wide variety of additional tools have been developed, such as RECON [40], RepeatScout [41] and LtrHarvest/Ltr_retriever [42], REPET [43], RepeatExplorer [44] (based on paired-end reads). The availability of multiple tools highlights the lack of a standardized protocol, making it impossible to directly compare the RE composition between different genomes based solely on the literature. Moreover, current pipeline annotations of REs fail to identify a significant portion of divergent repeats in non-model organisms. To address these limitations, we designed a standardized protocol for RE annotation that encompasses both TEs and satDNAs. This pipeline was used to establish the RE landscape of twenty Decapoda and six other Crustacea, enabling an objective comparison of the Decapoda repeatomes in terms of abundance, composition, and evolutionary dynamics. Our standardized approach allowed us to assess the contribution of REs to the evolution of the enigmatic Decapoda genomes. Furthermore, we explored the possibility of using the REs as reliable phylogenetic markers for Decapoda. Lastly, this study also provides a new library of REs in Decapoda genomes that extends the existing databases and can be used for future analyses.

2. Materials and Methods

2.1. Genomic Datasets

Available assemblies for Decapoda species were downloaded from NCBI GenBank and RefSeq (last accessed 16 February 2022). Contig and scaffold N50 are useful values to estimate the contiguity of the genome by indicating the length of the shortest contig or scaffold that cover 50% of assembly. However, Decapoda genomes present variable N50 values (Table S1). The BUSCO completeness score, which can be independent of the contiguity of the genome, was also determined for each genome to assess the completeness of the assemblies (Table 1) [45]. Only the 20 genomes with a BUSCO completeness score of at least 25% were selected. Considering the low number and fragmentation status of available Decapoda genomes, a lower BUSCO score threshold than usually used was chosen to retain at least one genome in all infraorders that had genome assemblies. To obtain a broader perspective of the landscape of Decapoda REs compared to crustaceans, we added 6 non-Decapoda crustaceans (Table 1). This allowed us to see if Decapoda species have a different or similar trend in terms of the proportion of the individual repeat families, the presence/absence of RE families, and finally their evolutionary trajectories in comparison to six other Crustacea.

Table 1. Genomic datasets used in this study.

Suborder/Infraorder	Species	Assembly Access ID	Assembly Size (Mb)	BUSCO Completeness (%)	Paired-End Illumina Reads SRA Access ID	Estimate Genome Size (Mb)	Estimate Genome Size Reference
Dendrobranchiata	*Penaeus chinensis*	GCF019202785.1	1466	90.7	SRR13452153	2660	[46]
	Penaeus indicus	GCA018983055.1	1936	88.5	SRR12969543	2810	[47]
	Penaeus japonicus	GCF017312705.1	1705	96.6	DRR278744	2170	[47]
	Penaeus monodon	GCF015228065.1	2394	83.9	SRR11278066	2200	[47]
	Penaeus vannamei	GCF003789085.1	1664	84.8	SRR13661692	2270	[47]
Caridea	*Caridina multidentata*	GCA002091895.1	1949	25.2	DRR054559	3230	[48]
	Macrobrachium nipponense	GCA015104395.1	1985	41	SRR9026393	4600	[49]

Table 1. *Cont.*

Suborder/Infraorder	Species	Assembly Access ID	Assembly Size (Mb)	BUSCO Completeness (%)	Paired-End Illumina Reads SRA Access ID	Estimate Genome Size (Mb)	Estimate Genome Size Reference
Achelata	*Panulirus ornatus*	GCA018397875.1	1926	70	SSR13822589	3230	[50]
Astacidea	*Procambarus virginalis*	GCA020271785.1	3701	67	SRR12901906	3500	[51]
	Procambarus clarkii	GCF020424385.1	2735	94.3	SRR14457195	8500	[52]
	Cherax destructor	GCA009830355.1	3337	81.7	SRR10467055	4500	[53]
	Cherax quadricarinatus	GCA009761615.1	3237	69.9	SRR10484712	5000	[54]
	Homarus americanus	GCF018991925.1	2292	93	SRR12699166	7700	[55]
Anomura	*Paralithodes camtschaticus*	GCA018397895.1	3810	44.2	SRR13805857	7290	[50]
	Paralithodes platypus	GCA013283005.1	4805	71.7	SRR1145749	5490	[56]
	Birgus latro	GCA018397915.1	2959	57.7	SRR13816158	6220	[50]
Brachyura	*Chionoecetes opilio*	GCA016584305.1	2003	91	SRR11278230	1655	
	Eriocheir sinensis	GCA013436485.1	1272	92.6	SRR11971329	2230	[57]
	Portunus trituberculatus	GCF017591435.1	1005	93.5	SRR9964028	2250	[57]
	Callinectes sapidus	GCA020233015.1	998	90.4	SRR15834103	2290	[58]
Other Crustacea	*Amphibalanus Amphitrite* (Cirripedia)	GCA019059575.1	808	93.9	SRR9595623	481	[59]
	Armadillidium vulgare (Isopoda)	GCA004104545.1	1725	84.5	SRR8156178	1660	[60]
	Daphnia magna (Phyllopoda)	GCA020631705.2	161	98.6	SRR15012074	238	[61]
	Darwinula stevensoni (Podocopida)	GCA905338385.1	382	90.3	SRR8695251	437	[62]
	Eurytemora affinis (Copepoda)	GCA000591075.2	389	91	SRR2452640	616	[63]
	Hyalella Azteca (Amphipoda)	GCA000764305.4	551	93.8	SRR1556043	1050	[64]

2.2. Identification and Annotation of Repetitive Elements

2.2.1. Identification of Satellite DNA Families

For each species, a set of Illumina paired-end reads was randomly chosen in the SRA database (Table 1). Reads that mapped to the mitochondrial genome were discarded, and the remaining reads were sampled to represent 1.6% of estimated genome size. Genome size estimations were retrieved for all genomes, except for *Chionoecetes opilio* (Table 1). For this genome, all short paired-end reads corresponding to the assembly were downloaded and the genome size was estimated using KmerGenie version 1.7051 [65]. The sets of reads were then analysed using the TAREAN pipeline, Galaxy version 2.3.8.1 [66] (reads trimmed at 100 bp and default parameters) to compile each species-specific library of satellite elements.

2.2.2. Construction of a Common Library of Repetitive Elements

De novo identification of repetitive elements in each genome was performed using RepeatModeler2 version 2.0.1 [38] with the LTRStruct option and default parameters. The LTRStruct option is an LTR structural discovery pipeline that allows a better identification of LTR elements by using LTR_Harvest and LTR_retriever.

All species-specific libraries of repetitive elements identified with RepeatModeler2 were renamed according to the RepBase version 26.05 [67] nomenclature, with the repeat family, a unique number for the family to distinguish the different sequences of the repeat, the 3-letter species name, the repeat class and family, and finally the complete species name. Similar renaming was applied to species-specific libraries of high-confidence satellites identified by the TAREAN pipeline, with the addition of a 'tarean' tag after the unique number.

All species-specific libraries of high-confidence satellites and repeats identified by the TAREAN pipeline and RepeatModeler2 were combined with the Arthropoda-specific subset of RepBase26.05 to form a single library (Figure 1). This library was then split into 2 sub-libraries. The first one corresponds to the known TEs and the second one represents unknown TEs, satellites, and simple repeats.

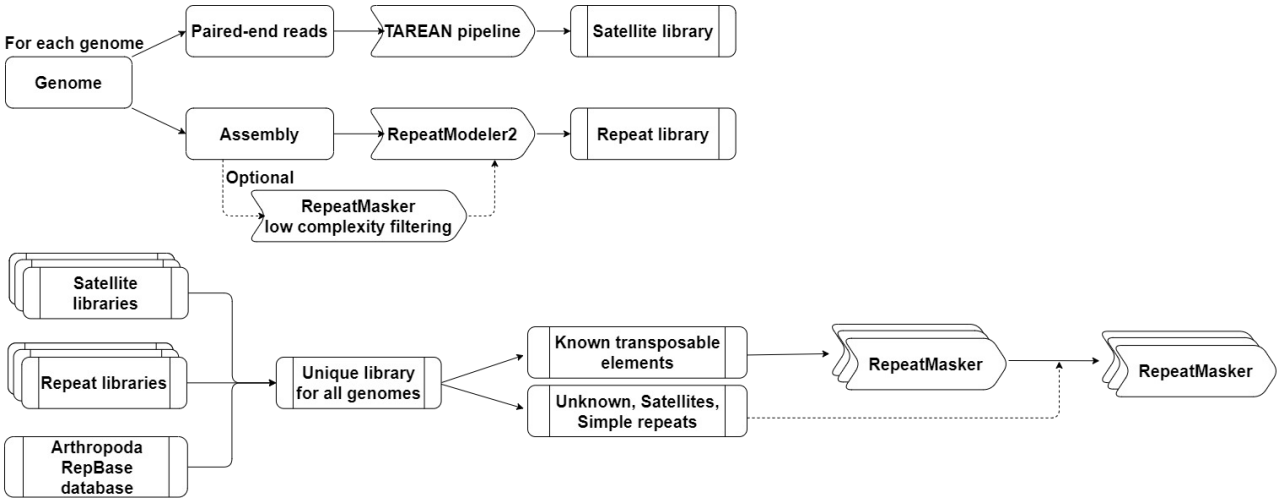

Figure 1. Standardized annotation protocol for repetitive elements developed in this study.

2.2.3. Identification of Repetitive Elements

In order to annotate repetitive elements that are present in the 26 crustacean genomes, we used RepeatMasker version 4.1.2-p1 [39] following a two-step approach (Figure 1). First, we used RepeatMasker with the library of known TEs using the options -a -gccalc -excln -s -nolow to identify and mask TEs in genomic sequences. We then performed a second run of RepeatMasker (with -a -gccalc -excln -s options) on the previously masked genomes using the second library to identify unclassified TEs, satellite DNA, and simple repeats. The ProcessRepeats and buildSummary tools of RepeatMasker were then used to combine all results and produce a detailed summary of annotations.

2.2.4. Statistical Analysis

In order to test for correlation between genome size, assembly size, repeats, or TE load (number of copies) or percentage, we used a linear regression model and the Spearman rank sum method with $\alpha = 0.005$ using R package ggplot2 with lm method. A dendrogram was produced by calculating pairwise distances between repeat profiles (the pattern of presence and absence of repetitive elements) using hclust with the Euclidean method, and the heatmap was plotted using Orange3 [68]. The sequence divergence distribution was calculated as Kimura distances (rates of transitions and transversions) using the RepeatMasker tools "calcDivergenceFromAlign.pl" and "createRepeatLandscape.pl".

3. Results and Discussion

3.1. Construction of Repetitive Elements Reference

To obtain a comprehensive view of REs in Decapoda and reduce the number of elements classified as "unknown", we developed a standardized protocol to annotate TEs and satDNAs at the genomic level (see Methods and Figure 1). This pipeline integrates the consensus sequences of the Arthropoda section of the RepBase database and the *de novo* identification of REs in all species by a combination of RepeatModeler2 and the TAREAN pipeline, in order to generate an extensive library of consensus sequences. The TAREAN pipeline was used to specifically identify satDNAs. Due to their structure and high sequence homogeneity, satDNAs are extremely difficult to assemble and are often excluded from the assembly [12]. Therefore, we searched for satDNAs in Illumina raw reads paired-end sequences using the TAREAN pipeline to construct the "Satellite libraries". Using the TAREAN pipeline, we retrieved between 0 and 43 satDNA families annotated as "High fidelity", while RepeatModeler2 identified only 0 to 4 satDNA families (Table 2).

Table 2. Number of RE libraries identified and annotated using species-specific libraries or a merged library from all species. RMo—RepeatModeler2, Tp—TAREAN pipeline.

Suborder/Infraorder	Species	Ab Initio satDNA Families Identified		Number of Families Annotated Using RMo Species-Specific and Repbase as Library for Each Species			Number of Families Annotated Using Merged Libraries of RMo and Tp Libraries for All Species and Repbase		
		RMo	Tp	All RE Families	Percentage of Unknown	satDNA Only	All RE Families	Percentage of Unknown	Satdna Only
Dendrobranchiata	*P. chinensis*	1	7	7547	12.38%	24	22,702	3.44%	56
	P. indicus	1	2	8252	7.72%	30	24,237	3.40%	57
	P. japonicus	3	5	7693	7.25%	29	22,611	3.61%	59
	P. monodon	0	4	8647	9.28%	28	25,183	3.57%	57
	P. vannamei	0	3	7621	8.85%	30	23,240	3.49%	55
Caridea	*C. multidentata*	1	6	11,104	11.93%	38	28,065	11%	74
	M. nipponense	2	0	10,455	19.68%	38	26,021	13.42%	57
Achelata	*P. ornatus*	1	6	8850	21.13%	35	25,995	8.12%	60
Astacidea	*P. virginalis*	1	31	9213	28.26%	33	26,483	9.95%	96
	P. clarkii	2	39	8838	22.52%	34	26,051	13.67%	97
	C. destructor	4	24	10,391	14.10%	40	29,970	6.88%	92
	C. quadricarinatus	1	43	10,411	14.33%	35	26,966	4.99%	96
	H. americanus	1	2	9557	24.16%	35	27,873	17.29%	61
Anomura	*P. camtschaticus*	2	19	11,431	24.95%	33	30,169	14.36%	95
	P. platypus	0	36	11,332	32.76%	34	31,798	13.27%	109
	B. latro	1	2	11,053	25.48%	37	31,207	16.30%	59
Brachyura	*C. opilio*	0	0	10,400	22.89%	29	26,561	12.26%	52
	E. sinensis	1	0	8486	20.74%	29	23,937	11.82%	49
	P. trituberculatus	0	0	7399	12.28%	20	21,070	6.42%	39
	C. sapidus	0	2	6911	13.68%	18	19,041	8.68%	31

Table 2. *Cont.*

Suborder/Infraorder	Species	Ab Initio satDNA Families Identified		Number of Families Annotated Using RMo Species-Specific and Repbase as Library for Each Species			Number of Families Annotated Using Merged Libraries of RMo and Tp Libraries for All Species and Repbase		
		RMo	Tp	All RE Families	Percentage of Unknown	satDNA Only	All RE Families	Percentage of Unknown	Satdna Only
Other Crustacea	*A. Amphitrite* (Cirripedia)	1	1	6717	27.06%	14	11,969	14.90%	22
	A. vulgare (Isopoda)	0	13	9431	17.40%	27	19,098	11.91%	47
	D. magna (Phyllopoda)	2	3	3643	17.90%	10	6805	14.63%	11
	D. stevensoni (Podocopida)	1	2	9762	25.59%	22	17,339	23.89%	38
	E. affinis (Copepoda)	1	8	6069	33.37%	32	13,334	24.15%	46
	H. Azteca (Amphipoda)	1	10	6851	16.21%	28	14,424	13.69%	46

Using our newly developed pipeline, we identified between 3643 and 11,431 families of REs in the different assemblies, including between 7.25% and 33.37% of "unknown" sequences (Table 2). Unknown elements are repetitive sequences that could not be further classified. The lowest percentage of unknown elements is observed in Dendrobranchiata species. This might be explained by the presence of the annotated TEs of the Dendrobranchiata *Penaeus vannamei* in RepBase, allowing a better identification in closely related species.

All detected REs were renamed according to the RepBase nomenclature. In fact, the RE classification by Wicker et al. (2007) [35] is widely used, but new TEs have been characterized since the establishment of the classification in 2007, resulting in conflicts in TE databases. Kojima (2019) [37] improved the classification of the RepBase database [40], but TE annotations can differ between RepBase, RepeatModeler2 database, and DFAM due to capital letters or multiple naming of the same element, for example. A manual correction of repeat names was thus applied when needed in order to obtain a clear annotation.

All libraries generated by RepeatModeler2, the TAREAN pipeline, and RepBase were merged into a single library. This extensive database contains a total of 71,601 sequences including sequences from RepBase. Among these families, known TEs represent 31,579 sequences. With this new merged library, we considerably extended the number of annotated families compared to the RepBase database of Arthropoda REs. Indeed, RepBase provides consensus sequences of 13,906 repetitive elements in Arthropoda, including 109 satDNAs. These elements are distributed in 218 Arthropoda species and in Eukaryota or Metazoa common ancestors. However, only sixteen Crustacea and six Decapoda species are represented, with 1419 and 328 sequences, respectively. Moreover, most Decapoda sequences (320) are from a single species, *P. vannamei*, as repeats from other species have not been submitted to RepBase. This shows the lack of knowledge of REs in Decapoda species in established databases. Our work also extended the number of known satDNA families in Decapoda species, with 405 consensus sequences compared to the 109 present in RepBase. The new REs identified in this study are provided in Supplementary Materials (Figure S1). Well-categorized REs have also been submitted to RepBase.

3.2. Annotation of Repetitive Elements in Decapoda Genomes

With our new extensive database, we performed two rounds of annotation using RepeatMasker. In the first round we only used known TEs in order to have a better characterization and reduce the proportion of unknown TEs, and in the second we used all the remaining REs. We identified between 6805 and 31,798 consensus RE sequences in the different assemblies (Table 2). This represents an increase of approximately 16,500 families on average in Decapoda compared to previous annotations and 6500 for the other Crustacea. Moreover, our standardized protocol successfully identified the type of REs that were previously unclassified for most species (now between 4.40% and 24.15%). This represents a considerable improvement over the results obtained with the widely used species-specific databases.

Taking into account all the satDNA families annotated in the genome with the merged library, we annotated between 11 and 109 different families (previously 10 to 40 using the species-specific strategy, Table 2). The Astacidea and Anomura infraorders have higher numbers of satDNA families, ranging from 92 to 109, except for *H. americanus* and *B. latro*. The latter two species have a number of satDNA families more similar to the other Decapoda species, with 61 and 59 satDNA consensus sequences, respectively. The large number of satDNA families detected in Astacidea and Anomura is in agreement with the 258 families detected in the crayfish *Pontastacus leptodactylus* [5]. The diversification of satDNA families in Astacidea and Anomura is remarkable compared to the observations in other species. For example, *Drosophila* species generally have less than ten different families in their genomes, and humans have nine [20,69]. However, a large number of satDNA repeats has already been found in Arthropoda, such as *Triatoma infestans* (42 families, genome size 1.4 Gb) [70], *Locusta migratoria* (62 families, genome size 6 Gb) [18], the morabine grasshoppers (129 families, genome size 5 Gb) [71], and the fish *Megaleporinus microcephalus* (164 families, assembly size 1.2 Gb) [72]. It should be noted that our results may still underestimate the real number of satDNA families, due to the fragmentation of available assemblies (Table S1). In fact, some satDNA families identified by the TAREAN pipeline in Illumina reads were not retrieved in the genome assembly. It is likely that the missing satDNAs were contained in reads that were not included in the final assembly. However, the number of satDNAs remains consistent in each infraorder.

Interestingly, the number of RE families is correlated with both estimated genome size and assembly size (Table 1) with a Spearman rank correlation test of $\rho = 0.83$, p-value = 8.925×10^{-8} and $\rho = 0.92$, p-value = 1.146×10^{-6}, respectively. The same correlation is observed with satDNA families, with Spearman rank correlation test of $\rho = 0.84$, p-value = 6.875×10^{-8} and $\rho = 0.90$, p-value = 3.83×10^{-10}, respectively. This result reveals the importance of the diversification of RE families in larger genomes.

The strategy used in this study increases the knowledge of REs in Decapoda species and provides an extended library that can be used in future studies (Figure S1). Unfortunately, there are still a large number of unknown REs in some of the annotated genomes. A manual curation of the library would be necessary but was beyond the scope of this study. We also want to mention that, due to the high presence of REs, genome assemblies are often fragmented, preventing the exhaustive annotation of TEs that can be absent from the assemblies or split into two contigs. The study of Sproul et al. (2022) of more than 600 insect species showed the influence of sequencing technology on repeat detection, with long read assemblies containing 36% more repeats than short-read assemblies and a huge impact on LTR detection [73]. This is because assemblies based on long reads are often more contiguous [74,75]. In our case, most of the genomes were assembled using long reads or a combination of long and short reads, and short-read assemblies do not stand out concerning repeat content or diversification (Table S1).

3.3. Proportion of Repetitive Elements in Decapoda Genomes

The RE proportions are variable both between and within phylogenetic clades of the analysed species. The proportion of REs in the studied Arthropoda genomes is above 40%.

Exceptions are two Decapoda species, *C. quadricarinatus*, with the lowest contig N50, and *C. multidentata*, with the lowest BUSCO score. They present 38.73% and 39.02% of repeat content, respectively (Table 1, Figure 2, and Table S1). The non-Decapoda *H. azteca* also presents fewer REs, with 26.12%, and is one of the genomes assembled with short reads only (Figure 2 and Table S1), but given the fragmented status of these genomes, these percentages may underestimate the RE proportion. Compared to the Decapoda species, which have an average of 59.7% REs in their genomes, the non-Decapoda Crustacea analysed in this study exhibit a lower proportion of REs, with an average of 46.4%. However, it is important to note that *A. vulgare* stands out among the non-Decapoda studied, as it has a remarkably high percentage of repeats (76.26%). If *A. vulgare* is excluded, the average of REs in non-Decapoda is reduced to 40.4% and the difference is significant, with Wilcoxon p-value = 0.0074. Within Decapoda species, Anomura presents an especially high percentage of REs, with on average 73.6%. Indeed, the Anomura species *P. platypus* has the highest proportion of REs among the studied species with 78.89% (Figure 2). In contrast, the genome with the lowest percentage of repeats was the non-Decapoda *H. azteca* with 26.12%. Thus, the RE proportions were highly variable among the phylogenetic clades, as was the content of RE categories.

We also observed a variability in the content of REs within suborders. Among Decapoda, Dendrobranchiata exhibited half the amount of LINEs compared to Pleocyemata, with up to 35.3% in the Astacidea *C. destructor* (Figure 2). Dendrobranchiata was characterized by a high proportion of DNA transposons, for example in *A. vulgare,* with between 13% and 18% of DNA transposons. The Anomura infraorder has the highest percentage of LTRs, with more than 16%, and the Achelata *P. ornatus* has the lowest, with 3.24%. SINE elements were rare in all genomes, ranging from 0.02% in *H. Azteca* to 2.54% in *P. trituberculatus*. DIRS elements contribute less than 1% of the repeat content in almost all genomes. The main exception was *M. nipponense*, where DIRS represented 8.84%. This species also has the highest proportion of Penelope elements, with 5.18%. The infraorder with the second highest number of Penelope elements was Astacidea, with a mean of 2.3%. Unclassified elements were less frequent in the Dendrobranchiata suborder, with around 3.5%, probably because of the better characterization of REs in this suborder in the RepBase database, with the almost exclusive presence of annotations derived from *P. vannamei*. Therefore, more divergent species present a higher proportion of unclassified elements, such as *E. affinis* with 24.15%. The content variability suggests that the different suborders of the studied crustacean species have specific major REs present in their genomes.

According to RE studies of Decapoda species included in assembly publications, the proportion of REs varies from 8% to 82% [48–51,53–55,76–84]. Tan et al. (2020) annotated the repeatome of eight decapod species and estimated repetitive content between 27% and 50%, with the majority of the genomes having more LINEs, except for *P. vannamei*, which had more DNA transposons [54]. Compared to these studies, we annotated approximately 10% more repeats with our pipeline. For the *P. virginalis* genome, 8.8% of repetitive elements were retrieved in the assembly Pvir0.4 (GenBank accession: GCA_002838885.1) and 27.52% in the study of Tan et al. (2020) [54]. However, in the assembly DKFZ_Pvir_1.0 (GenBank accession: GCA_020271785.1), the new assembly version used in this study, we annotated 57.87% of repetitive elements [51,55]. In the assembly of *P. clarkii*, Xu et al. (2021) annotated 82.42% of repeats, while in our study, we observed only 71.26% (Figure 2). For the *P. platypus* genome, we observed similar overall results to Tang et al. (2021) (Figure 2) [56]. However, the percentages of LINEs and LTRs are increased by almost 10% each, while unknown TEs were reduced to 17%. The percentage of REs in *E. sinensis* was estimated at 40.5% and 61.42% in two different studies [54,85], while here we determined that repetitive elements represent 58.93% of the genome (Figure 2). Taken together, these results show that our method provides greater or equal proportion of REs but with a better characterization.

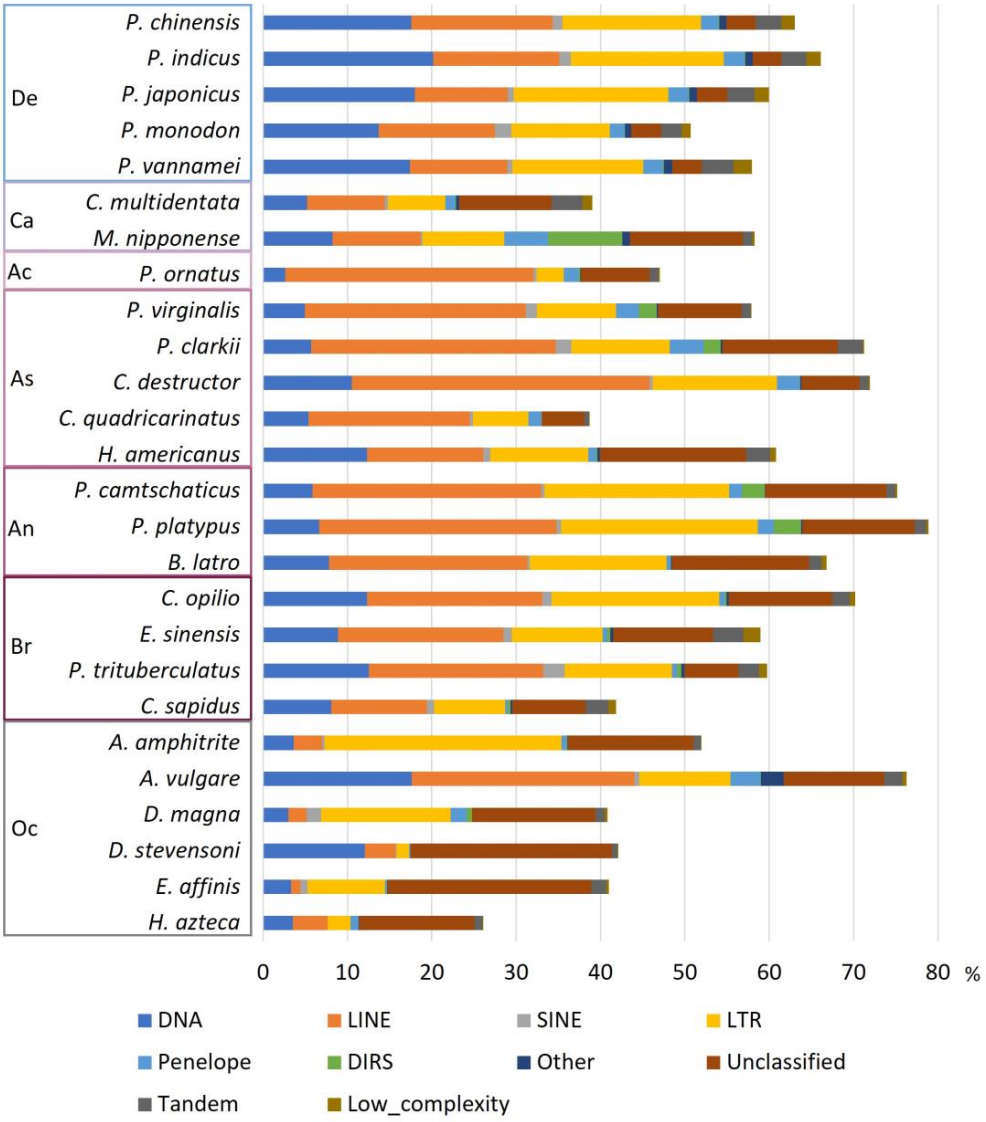

Figure 2. Proportion and content of repetitive elements in genomes. Percentage of repetitive elements in the genome by class of repetitive elements. De, Dendrobranchiata; Ca, Caridea; Ac, Achelata; As, Astacidea; An, Anomura; Br, Brachyura; Oc, other Crustacea.

The Decapoda species studied here all presented high proportions of REs, ranging from 58% to 79% (Figure 2). They are in the upper range of what is generally observed in Arthropoda. Indeed, comparative studies carried out on arthropods (mainly based on insects) report highly variable proportions of TEs, ranging from 1% to 80% [73,86,87]. We expect even higher proportions of REs with the forthcoming sequencing of giant genomes in Decapoda or other Crustacea. Recently, the assembly of the Antarctic krill (belonging to a sister order of Decapoda) demonstrated that 92% of its genome is constituted of REs, 78% of them being TEs, indicating that Arthropoda can have an extremely high proportion of REs [88]. In terms of TE landscape, Decapoda presents only a few SINE elements, as for all Arthropoda (Figure 2). Previous studies in Dendrobranchiata species reported that the most abundant groups of repeats, disregarding simple sequence repeats, were DNA transposons or LINEs, with different results depending on the bioinformatic tools used [73,86,87]. Here, we showed that DNA transposons were the major subclass in all Dendrobranchiata species, followed by LINEs (Figure 2). This is similar to what is observed in most insect species, where DNA transposons are generally the major TE group present in genomes [73,86,87]. Interestingly, our results revealed a different situation in the

studied Pleocyemata species, where LINE and LTR elements are more abundant (Figure 2). This can be compared to what is observed in some insect orders exhibiting a different TE composition: LTRs are more abundant in Diptera species, and Odonata and Orthoptera species are richer in LINE elements [73,86]. The change in the major type of REs between suborders suggests an altered strategy for genome stability maintenance and regulation of REs between suborders. Sproul et al. (2022) demonstrated that LINE-rich species lineages present many REs that are associated with protein-coding genes [73]. Such associations suggest consequences regarding phenotype evolution. The presence of a TE near a gene can lead to methylation changes. Indeed, it already has been shown that LINEs can serve as amplifiers for silencing away from the X-chromosome inactivation center, and LINEs and SINEs for gene imprinting [34,89]. The movement of a LINE, or other TE, to a new genomic locus, can thus have an impact on nearby gene expression, and ultimately reshape gene expression networks and impact genome evolution.

3.4. Correlation between Genome Size and Repetitive Elements

The 20 Decapoda species analysed in the present study have large differences in genome size estimations (1.6 Gb to 8.5 Gb). These differences were also evident in assembly sizes, although less pronounced (1 Gb to 4.8 Gb). The variability of the genome sizes raised the question of the contribution of REs to their host genome. After masking each genome, we calculated the load of REs, i.e., the number of copies of REs and TEs only, and the percentage of REs and TEs only. We then tested for a correlation between the aforementioned values and both assembly size and estimated genome size. The assembly size was positively correlated with both the load ($\rho = 0.87$, p-value $= 1.864 \times 10^{-6}$) and the percentage of TEs ($\rho = 0.6$, p-value $= 1.48 \times 10^{-3}$) (Figure 3A,B). The estimated genome size (Table 2) was positively correlated with the load of TEs ($\rho = 0.62$, p-value $= 7.114 \times 10^{-4}$), but there was no significant correlation with the percentage of TEs ($\rho = 0.47$, p-value $= 1.421 \times 10^{-2}$) (Figure 3C,D). Although the number of satDNA families was correlated with both assembly size and estimated genome size, when satDNA elements are included, the significance of the correlation between the load of REs and genome/assembly size is smaller (Figure S1). The correlations between the percentage of REs and both assembly and estimated genome size were not significant, with $\alpha = 0.005$ (Figure S1).

For the first time in Decapoda species, a strong correlation is demonstrated between assembly size and load (number of copies) of TEs. This strong positive correlation reveals the impact of the number of TEs on the size of the assembly, with larger genomes associated with a higher presence of TEs. The percentage of TEs or REs is more often analysed than the load. In our study, the percentage of TEs was less significantly correlated with genome or assembly size than the load of TEs, and REs were not correlated with genome size. As in our study, Petersen et al. (2019) [86] found a positive correlation between the percentage of TEs and assembly size in arthropods, but they also found a positive correlation between the percentage of TEs and estimate size, which was not observed in our study. Moreover, Sproul et al. (2022) [73] found a positive correlation between the proportion of REs and assembly size in insects, which was not confirmed in our study. The differences between our results and the cited studies are likely due to the difficulties in assembling REs in large genomes such as Decapoda [73,86]. During assembly, REs can be excluded from the assembly even if they are present in the genome. It is therefore expected that REs are more correlated with assembly size than the estimated size. REs can also be fragmented and included in the assembly only partially, contributing to the load of REs in the genome but not to the percentage. This could explain the higher correlation coefficient observed for the load of REs in Decapoda genomes and highlights the usefulness of studying both percentage and load of REs in fragmented assemblies. The presence of fragmented REs is particularly true for satDNAs, which are often concatenated, since the assembler cannot define how many repetitions are present if they are not entirely covered by a long read. These difficulties in assembling satDNAs are particularly pronounced when assemblies are highly fragmented, as in this study, and could explain the decrease in or absence of the significance of the tests

when including satDNAs. An improvement in genome contiguity could therefore affect inferences of correlation between REs and genome size. However, removing genomes of BUSCO score of less than 50% does not change conclusions on correlations between repeats and genome size.

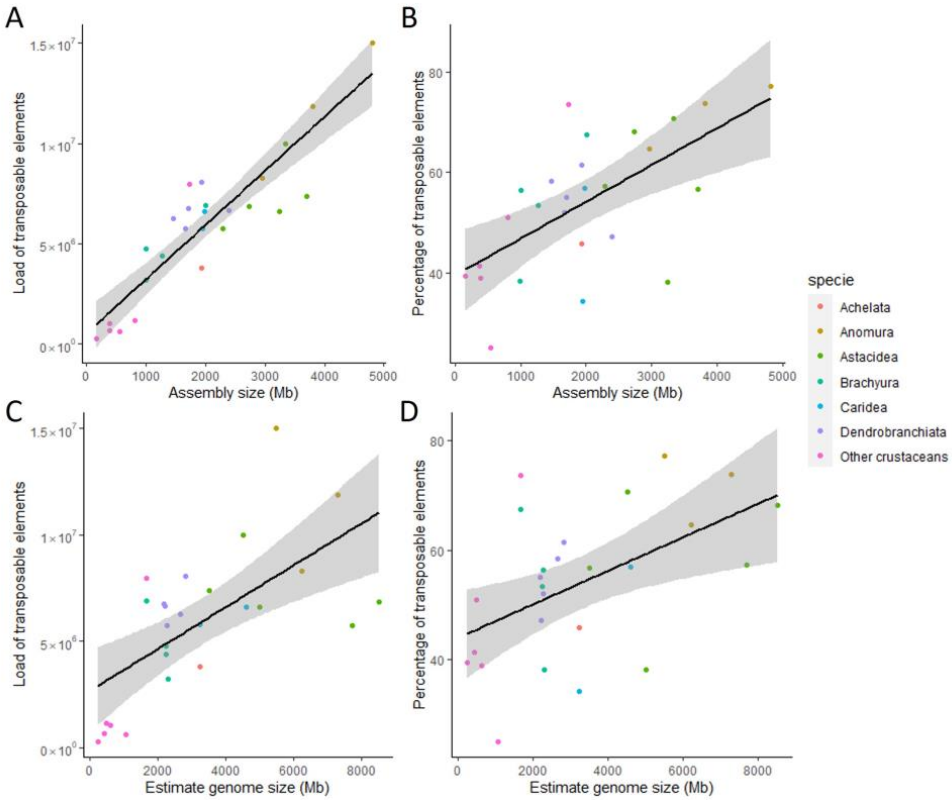

Figure 3. Correlation between genome size and TEs. Correlation plots between assembly or estimated genome size and load (number of copies) or percentage of TEs. Orders and suborders are indicated by different colours. (**A**). Correlation between assembly size and the load of TEs. Spearman rank correlation test: $\rho = 0.87$, p-value $= 1.864 \times 10^{-6}$. (**B**). Correlation between assembly size and the percentage of TEs. Spearman rank correlation test: $\rho = 0.6$, p-value $= 1.48 \times 10^{-3}$. (**C**). Correlation between estimated genome size and the load of TEs. Spearman rank correlation test: $\rho = 0.62$, p-value $= 7.114 \times 10^{-4}$. (**D**). Correlation between estimated genome size and the percentage of TEs. Spearman rank correlation test: $\rho = 0.47$, p-value $=1.421 \times 10^{-2}$.

3.5. Frequency of satDNA Families Occurrence

In Crustacea, and particularly in Decapoda, we annotated a large number of different satDNA families (Table 2) and evaluated the occurrence of each family in each genome (Figure 4). In each genome, the majority of satDNA families were detected one to nine times. Depending on the genomes, between one and thirty-four families appeared between 10 and 99 times. With nine out of the ninety-seven satDNA families repeated more than 1000 times, *P. clarkii* was the species with the highest number of highly repeated satDNA families. In contrast, five genomes do not have highly repeated satDNA families (more than 99 occurrences). Thus, although Decapoda has extremely large numbers of satDNA families (Table 2), only a few are predominant in each genome (Figure 4), as seen in several other studies [18–20]. The Decapoda and non-Decapoda species studied here are no exception. The Decapoda infraorders Astacidea and Anomura had the largest genome size estimation and assembly size (Table 1) and presented the largest numbers of families that were highly repeated in their genomes (Figure 4). They also tend to have the highest total number of families (Table 2). This suggests that satDNA is a key factor in explaining the huge variations in genome size observed in Decapoda.

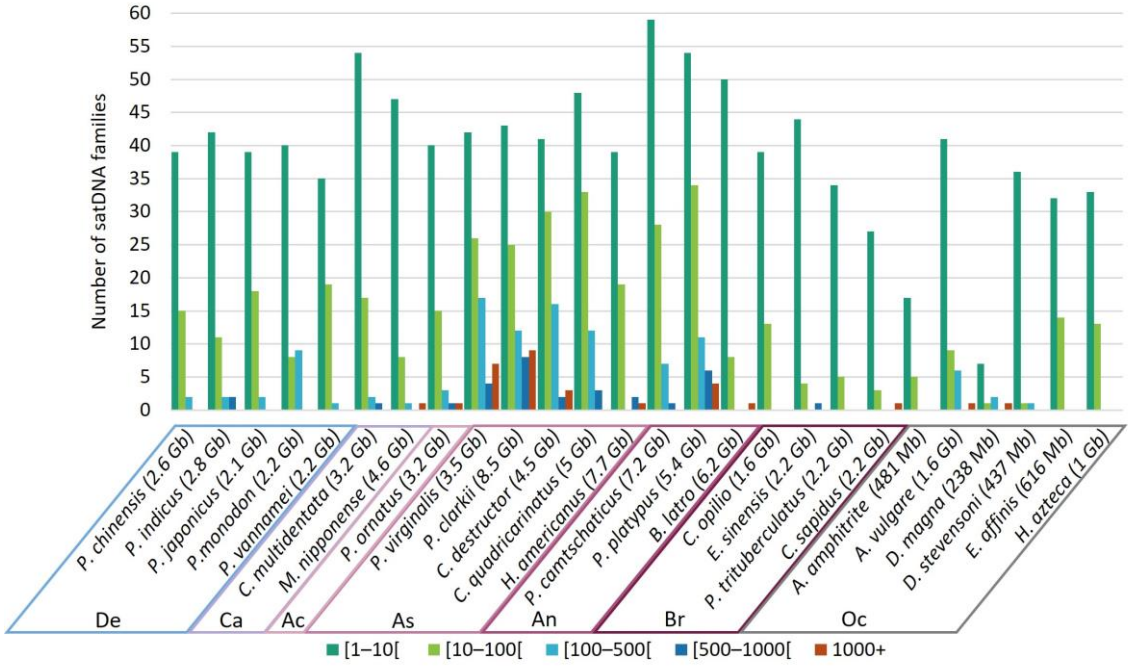

Figure 4. Distribution of satDNA families according to the number of occurrences in each genome. Low-frequency families (less than 10 occurrences) are indicated in dark green, while highly abundant families with more than 1000 occurrences are indicated in red. Number indicated for each species is the estimated genome size. De, Dendrobranchiata; Ca, Caridea; Ac, Achelata; As, Astacidea; An, Anomura; Br, Brachyura; Oc, other Crustacea.

3.6. Diversity of Repetitive Elements

To investigate the diversity of REs, we determined the number of copies (the load) of each superfamily of REs identified for each genome (Figure 5). With 67 superfamilies of TEs present in at least one species, the majority of the known superfamilies of REs were found in the investigated genomes, as seen in insects [86], and appear highly conserved across all the genomes (Figure 5). Among the studied Decapoda genomes, there was a clear pattern of high and low presence of repeat superfamilies, with only a few distinct variations between species by repeat suborder.

The load of REs of each superfamily was then used as a profile for each genome to construct the dendrogram by clustering of the RE profiles (Figure 5). This dendrogram mainly followed the currently known species phylogeny [4] except for *A. vulgare*, whose RE proportions and composition were more similar to Decapoda (Figure 2) and two Anomura species that were grouped with the Caridea. The genome of *A. vulgare* (1.6 Gb) was larger than the other Crustacea analysed in this study (238 Mb–1 Gb), with the highest percentage of repeats among the studied non-Decapoda crustacean species (Figure 2). This may explain why *A. vulgare* is clustered with Decapoda species and not with other crustaceans (Figure 5). Nevertheless, we could see a clear differentiation between Decapoda species and the other Crustacea that have a lower number and a distinct composition of REs, except for *A. vulgare*. Similarly, we could clearly distinguish Dendrobranchiata from Pleocyemata infraorders, with the presence of LINE *ingi* and SINE *MIR*. Within Pleocyemata, Caridea was also separated from the other Reptantia species, in agreement with the established phylogeny [4]. Many studies, including Petersen et al. (2019) [86], Sproul et al. (2022) [73], and Wu and Lu (2019) [87], based their RE analysis on already published phylogenetic trees. In our study, we clustered the repetitive profile of each genome and obtained a phylogenetic signal that respects the major classification (Figure 5) [1]. In fact, REs have been used recently as evidence for phylogenetic tree construction in plants, with RE abundance resolving species relationships in a similar manner to DNA sequences from plastid and nuclear ribosomal regions [90,91]. This can be explained by the capacity of

some REs to have a high conservation and synteny within species [92–94]. This approach could therefore be used in the future to determine the phylogeny of non-model species using low-coverage, low-cost sequencing.

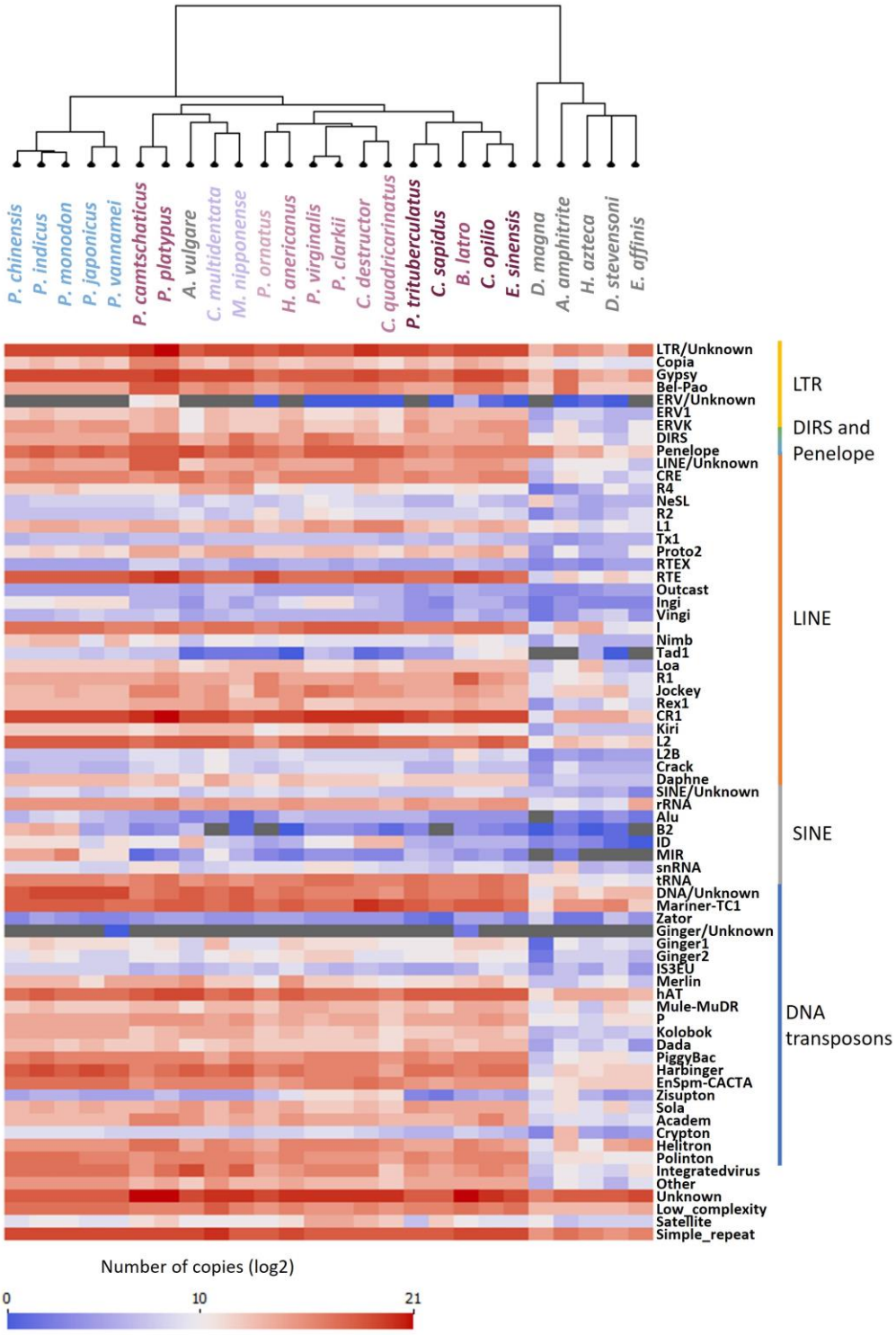

Figure 5. Diversity of repetitive elements. Log2 of the load of each family of repetitive elements identified for each genome was graduated between 0 (blue) and 21 (red). Gray colour indicates raw values of 0, before log2 transformation. The dendrogram was produced according to repeat profile by clustering.

3.7. Sequence Divergence Distribution of Transposable Elements

The genetic distance between each annotated TE copy and the consensus sequence of the respective TE family was calculated using the Kimura 2P distance in order to analyse the sequence divergence distribution and approximate the age and intensity of duplication events (Figure 6). The distribution shows the genomic coverage of TE copies according to the percentage of divergence from their family consensus estimated using the Kimura 2P distance. A peak indicates that a large group of TE copies shares the same divergence to the consensus sequence and suggests a major expansion event of these elements. This event is more recent if the peak is located at a low Kimura 2P distance from the consensus, i.e., at a low percentage of divergence. At a high Kimura 2P distance, a wide peak can indicate that TE copies have undergone genetic drift or other processes, leading to high sequence divergence and suggesting an ancient expansion event.

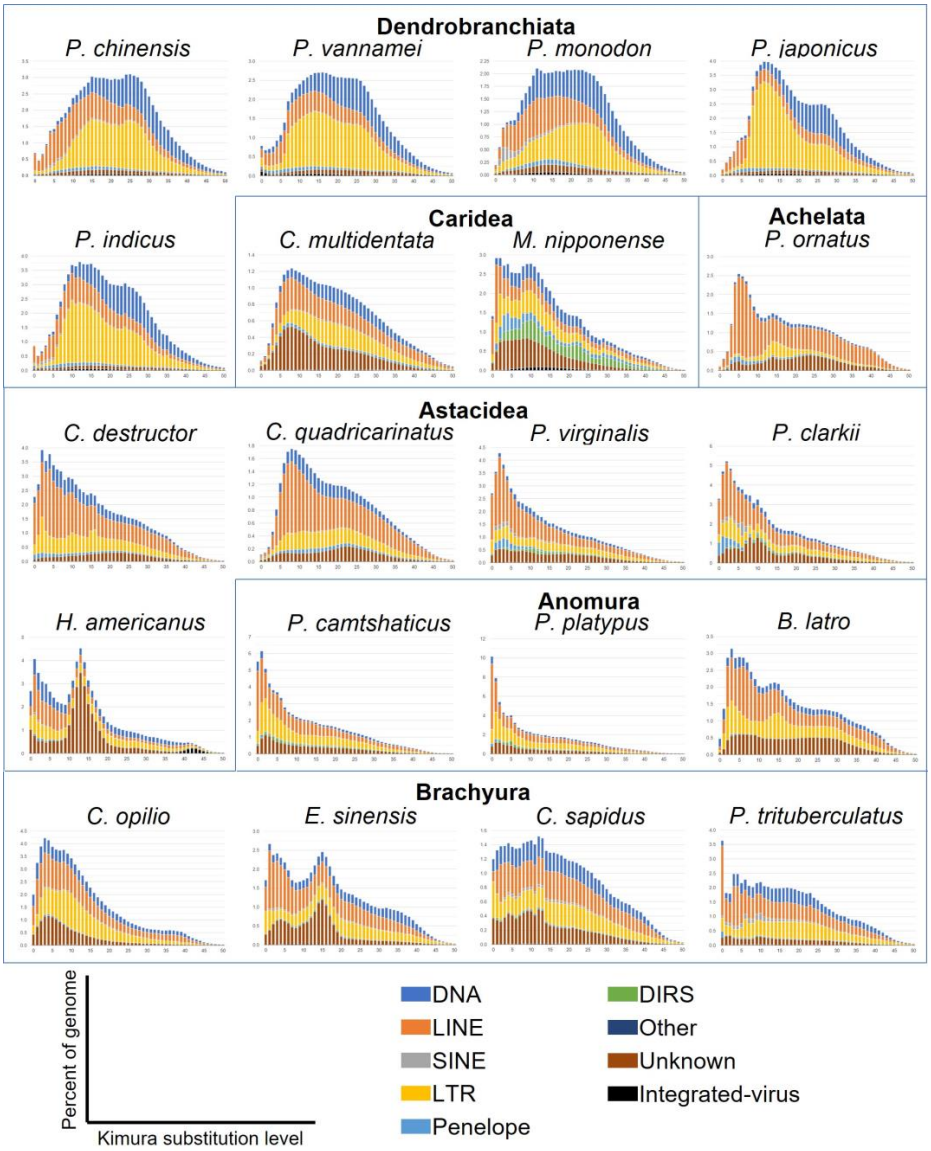

Figure 6. Sequence divergence distribution of TEs representing TE accumulation history based on Kimura 2P distance. Percentage of sequence divergence, or Kimura substitution level, is indicated on the *x*-axis. On the *y*-axis is the percentage of the genome occupied by each TE type; the scale is different for each genome depending on the percentage occupied. The TE type is indicated by the color chart.

In Dendrobranchiata, sequence divergence landscapes were similar for the five species (Figure 6). We observed two very similar peaks. The first one presented a larger number of LTRs and a smaller increase in LINE elements between 10% to 15% of divergence. The peak of LTRs was particularly high in *P. japonicus* and *P. indicus*. At the same time point, we observed an increasing amount of DNA transposons with the same distance to the consensus in *P. monodon*. A longer time ago, an augmentation of DNA transposons and LTR elements around 25% of divergence was shared by all species. This suggests that all the Dendrobranchiata shared the same old evolutionary events. The *P. monodon* genome was one of the few analysed Decapoda genomes showing a recent peak of SINE elements with the two *Procambarus* species. We would therefore expect to see a higher proportion of SINEs in *P. monodon* compared to other genomes. However, SINE elements were only slightly more abundant in this genome due to a higher presence of SINE MIR elements (Figures 2 and 5). Interestingly, the content of repeats showed that DNA transposons are the most widespread among the suborder (Figure 2). However, the expansion of DNA transposons was older and more spread out over time (Figure 6). In contrast, the landscape and diversity of repeats showed a higher peak of LTR elements over time in the suborder compared to the other species, with Gypsy being the most abundant (Figures 5 and 6). There were almost no sequences with low divergence. This quasi-absence of recent peaks in Dendrobranchiata suggests low activity of the TEs in recent times in these genomes (Figure 6).

The two Caridea species presented a different sequence divergence landscape (Figure 6). In *C. multidentata*, there was a recent peak of unknown elements between 5% to 10% of divergence. This peak could be caused by the expansion of one or several families of unknown TEs. We also observed that from high divergence, the fraction of the genome increased as the Kimura 2P distances decreased. This trend could be seen until the event at 5% to 10% of divergence. After this event, and more recently, the number of TEs with very low divergence decreased, with almost no TEs at 0% of divergence. This suggests that despite the peak of recently active unknown elements, TEs are not active anymore for this species. For *M. nipponense*, we observed two recent peaks at 1–4% and 10% of Kimura divergence corresponding to LINE, Penelope, and LTR elements for the first one and DIRS for the second one. We observed integrated virus expansion between 5% and 25% of divergence. This was in accordance with the diversification of repeats (Figure 5), where the *M. niponnense* genome was the Decapoda with the highest amount of integrated virus. The presence of sequences with little divergence from the consensus sequences suggests that TEs are active in this genome (Figure 6).

Within Astacidea, *H. americanus* has a different TE landscape compared to the other four species belonging to the infraorder (Figure 6). Indeed, the genome has a high peak at a divergence of 15% of unknown elements. Interestingly, we observed an ancient event concerning integrated viruses at 40% to 45% of Kimura 2P distance. The *H. americanus* genome was the only Decapoda genome studied here presenting this characteristic. Integrated virus could not be seen in the proportion of repeats because of their low presence in genomes and was included in the category "other REs" (Figure 2). Integrated virus in *H. americanus* sequences corresponds to the white spot syndrome virus (WSSV) [95], suggesting that *H. americanus* faced this virus a long time ago and these sequences were then propagated (Figure 6). Since WSSV is a worldwide threat to shrimps and potentially to many crustacean species, this interesting finding in a resistant species (i.e., *H. americanus*) could be important for future inferences into susceptibility/resistance to WSSV [96,97]. In the *H. americanus* genome, there was a clear increase in LINE, LTR, and DNA transposon coverage with a low percentage of divergence, which leads us to conclude that TEs are still active in this genome. TEs are also active in the *Procambarus* species, which has a similar landscape, with several elements at a low divergence and especially LINEs. We also observed an augmentation of Penelope and SINE elements at low divergence for both species. In *P. clarkii*, there was also a small peak at 10% of divergence of unknown elements. In contrast to the TEs in *C. quadricarinatus*, TEs seem to be active in *C. destructor*, with an

increase in LINEs at low divergence. The expansion of LINEs in *C. quadricarinatus* was, instead, more ancient, at 6% to 10% of divergence.

In Brachyura, all genomes seemed to have active TEs, but the TE landscapes across the genomes of this infraorder differ from each other (Figure 6). In *P. trituberculatus*, the LINEs with no divergence from consensus sequences were three times more abundant than LINEs at 1% of divergence. These LINEs were in a very active phase in this genome. Penelope elements were also more abundant at 0% of divergence. The *C. sapidus* genome showed an almost constant increased coverage of TEs with lower divergence for all elements. However, we observed an increasing number of LTRs with no divergence and a decreasing number of LINEs and DNA transposons. The genome of *E. sinensis* was the only Brachyura genome presenting two peaks. The oldest one was at 15% of Kimura 2P distance and was caused by unknown elements. The latest event involved LINE, LTR, and unknown elements at divergences between 0% and 7%. Of the Brachyura, *C. opilio* had the least active TEs. We observed a large peak between 0% to 20% of divergence, where LINEs and LTRs increased. The proportion of DNA transposons also increased during this time, but at a lower coverage.

Concerning the last two infraorders, in Achelata, the *P. ornatus* genome has a middle age peak at 15% of divergence, corresponding to LTRs (Figure 6). There was also a recent and high peak, around 4–8% of divergence, caused by the expansion of LINE elements, with 2% of the genome being represented by LINEs that are 6% divergent. This suggests that LINEs were, until recently, highly transcriptionally active in the genome but are now inactive. The high presence of LINE elements was also visible when considering the proportion of repeats in the genome (Figure 2). In Anomura, the intragroup with the highest percentage of LTRs within Decapoda (Figure 2), *B. latro* and the *Paralithodes* species had very different landscapes. The *B. latro* genome seemed to have inactive TEs, with two peaks of LTRs and LINEs at 3% and 15% of Kimura 2P distance (Figure 6). On the other hand, *Paralithodes* species had highly active LINEs and LTRs, with 6.8% and 3.6% of LINE elements without divergence to consensus sequences in *P. platypus* and *P. camtschaticus*, respectively. Finally, for other crustaceans, the amount of unknown elements in their genomes was predominant, making the analysis of the divergence distribution of TEs in their genomes difficult to interpret (Figure S2).

A clear differentiation in sequence divergence distribution between Dendrobranchiata and Pleocyemata species was observed, as seen with the proportion and diversity of repeats (Figure 6). Indeed, Dendrobranchiata have more non-transcriptionally active TEs compared to the majority of Pleocyemata. Among all Pleocyemata species studied here, almost all have at least one or more types of active TEs. The expansion of a particular subfamily of RE increases genome plasticity and can indicate periods of rapid evolutionary changes [14,33]. This suggests that Pleocyemata genomes had a rapid evolution on a recent timescale. Genomes with recent accumulations of repeats present highly similar repeats or types of repeats that can be long (mostly LTRs and LINEs). These long repetitive regions are more difficult to assemble, and so repeat resolution during assembly is even more problematic [98]. Indeed, we could argue that a large number of the genomes studied presented recent accumulation of long REs. These long REs, being difficult to assemble, can be a possible explanation of assembly fragmentation. Moreover, species with larger genome sizes tend to have more transcriptionally active TEs, but also more REs.

4. Conclusions

In this study, we annotated repetitive elements in twenty Decapoda and six other Crustacea genome assemblies publicly available, using a new pipeline for the annotation of repetitive elements. We showed that repetitive elements constitute a large fraction of Decapoda genomes, with a highly variable content of REs both between and within infraorders of Decapoda. Additionally, our analysis indicates that in Decapoda, both the load of repetitive elements and the number of RE families are correlated with the assembly size of the genome. Moreover, larger genomes tend to have more active TEs (high proportion

of sequences at 0% of divergence from their consensus), confirming the impact of REs in genome size expansion. We also demonstrated that, although the age distribution of TE superfamilies shows intra- and inter-lineage variation, the clustered RE profile reflects the phylogeny of the major groups analysed in this study. Compared to non-Decapoda Crustacea, Decapoda have a higher proportion and number of REs in their genome. Moreover, the pattern of RE families present in Decapoda is well-conserved across species. With our protocol, we showed that the combination of repeat libraries of all species provides an excellent tool to analyse content and diversification of repetitive elements with on average 8% more categorized elements. The new consensus sequences can improve the annotation of TEs in other Crustacea or Arthropoda species by increasing the number of consensuses for homology searches. We suggest using this two-step pipeline for all repeatome studies on non-model organisms that are often underrepresented in public databases. Our pipeline provides a baseline for future genomic analysis, producing standardized and reproducible analyses that will allow for much more rigorous and complete comparative analysis of repeats in non-model organisms.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes14081627/s1, File S1: crustaceans_RE_library.fa; Table S1: Assembly metrics.; Figure S1: Correlation between genome size and REs.; Figure S2: Sequence divergence distribution of TEs.

Author Contributions: Conceptualization, C.R., L.B., C.F., L.L.B., K.T. and O.L.; methodology, C.R., D.M., L.L.B., L.B., C.F. and O.L.; software, C.R. and A.K.; visualization, C.R.; writing—original draft preparation, C.R., K.T. and O.L.; writing—review and editing, C.R., C.F., L.B., L.L.B., K.T. and O.L.; supervision, K.T. and O.L.; project administration, K.T. and O.L.; funding acquisition, K.T. and O.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was produced within a framework of the GEODE project from the international collaborative research project co-funded by the Agence Nationale de la Recherche and the Deutsche Forschungsgemeinschaft (ANR-21-CE02-0028; DFG TH 1807/7-1). This work was supported by the French ministry of higher education and research and the doctoral school of Life Science of the University of Strasbourg.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: In this study, we generated a library of repetitive elements in crustacean species. Elements fully categorized were submitted to Repbase. The library of new repetitive elements found during this study is also provided in Supplementary Materials.

Acknowledgments: We thank the platform of Bioinformatics and Genomics BiGEst-ICube for bioinformatics supports. We are also very grateful to Julie Thompson for her critical reading of the manuscript and her valuable suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. De Grave, S.; Pentcheff, N.D.; Ahyong, S.T.; Chan, T.-Y.; Crandall, K.A.; Dworschak, P.C.; Felder, D.L.; Feldmann, R.M.; Fransen, C.H.; Goulding, L.Y.; et al. A Classification of Living and Fossil Genera of Decapod Crustaceans. *Raffles Bull. Zool.* **2009**, *21*, 1–109.
2. Reynolds, J.; Souty-Grosset, C.; Richardson, A. Ecological Roles of Crayfish in Freshwater and Terrestrial Habitats. *Freshw. Crayfish* **2013**, *19*, 197–218.
3. Souty-Grosset, C.; Holdich, D.D.M.; Noël, P.Y.; Reynolds, J.; Haffner, P. *Atlas of Crayfish in Europe*; Muséum national d'Histoire naturelle: Paris, France, 2006; Volume 187.
4. Wolfe, J.M.; Breinholt, J.W.; Crandall, K.A.; Lemmon, A.R.; Lemmon, E.M.; Timm, L.E.; Siddall, M.E.; Bracken-Grissom, H.D. A Phylogenomic Framework, Evolutionary Timeline and Genomic Resources for Comparative Studies of Decapod Crustaceans. *Proc. R. Soc. B Biol. Sci.* **2019**, *286*, 20190079. [CrossRef] [PubMed]
5. Boštjančić, L.L.; Bonassin, L.; Anušić, L.; Lovrenčić, L.; Besendorfer, V.; Maguire, I.; Grandjean, F.; Austin, C.M.; Greve, C.; Hamadou, A.B.; et al. The *Pontastacus Leptodactylus* (Astacidae) Repeatome Provides Insight into Genome Evolution and Reveals Remarkable Diversity of Satellite DNA. *Front. Genet.* **2021**, *11*, 611745. [CrossRef]
6. Lécher, P.; Defaye, D.; Noel, P. Chromosomes and Nuclear DNA of Crustacea. *Invertebr. Reprod. Dev.t* **1995**, *27*, 85–114. [CrossRef]

7. González-Tizón, A.M.; Rojo, V.; Menini, E.; Torrecilla, Z.; Martínez-Lage, A. Karyological Analysis of the Shrimp *Palaemon Serratus* (Decapoda: Palaemonidae). *J. Crustac. Biol.* **2013**, *33*, 843–848. [CrossRef]
8. Niiyama, H. On the Unprecedentedly Large Number of Chromosomes of the Crayfish, *Astacus Trowbridgii* Stimpson. *Annot. Zool. Japon.* **1962**, *35*, 229–233.
9. Crandall, K.A.; De Grave, S. An Updated Classification of the Freshwater Crayfishes (Decapoda: Astacidea) of the World, with a Complete Species List. *J. Crustac. Biol.* **2017**, *37*, 615–653. [CrossRef]
10. Gregory, T.R. Chapter 1—Genome Size Evolution in Animals. In *The Evolution of the Genome*; Gregory, T.R., Ed.; Academic Press: Burlington, NJ, USA, 2005; pp. 3–87.
11. Tørresen, O.K.; Star, B.; Mier, P.; Andrade-Navarro, M.A.; Bateman, A.; Jarnot, P.; Gruca, A.; Grynberg, M.; Kajava, A.V.; Promponas, V.J.; et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Res.* **2019**, *47*, 10994–11006. [CrossRef] [PubMed]
12. Treangen, T.J.; Salzberg, S.L. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nat. Rev. Genet.* **2012**, *13*, 36–46. [CrossRef] [PubMed]
13. Pop, M. Genome assembly reborn: Recent computational challenges. *Brief. Bioinform.* **2009**, *10*, 354–366. [CrossRef]
14. Shapiro, J.A.; von Sternberg, R. Why repetitive DNA is essential to genome function. *Biol. Rev.* **2005**, *80*, 227–250. [CrossRef]
15. Jurka, J.; Kapitonov, V.V.; Kohany, O.; Jurka, M.V. Repetitive Sequences in Complex Genomes: Structure and Evolution. *Annu. Rev. Genom. Hum. Genet.* **2007**, *8*, 241–259. [CrossRef]
16. Garrido-Ramos, M.A. Satellite DNA: An Evolving Topic. *Genes* **2017**, *8*, 230. [CrossRef] [PubMed]
17. Macas, J.; Neumann, P.; Navrátilová, A. Repetitive DNA in the pea (*Pisum sativum* L.) genome: Comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genom.* **2007**, *8*, 427. [CrossRef] [PubMed]
18. Ruiz-Ruano, F.J.; López-León, M.D.; Cabrero, J.; Camacho, J.P.M. High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Sci. Rep.* **2016**, *6*, 28333. [CrossRef]
19. Mravinac, B.; Plohl, M.; Ugarković, Đ. Preservation and High Sequence Conservation of Satellite DNAs Suggest Functional Constraints. *J. Mol. Evol.* **2005**, *61*, 542–550. [CrossRef]
20. Miga, K.H. Completing the human genome: The progress and challenge of satellite DNA assembly. *Chromosome Res.* **2015**, *23*, 421–426. [CrossRef]
21. Plohl, M.; Luchetti, A.; Meštrović, N.; Mantovani, B. Satellite DNAs between selfishness and functionality: Structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. *Gene* **2008**, *409*, 72–82. [CrossRef]
22. Plohl, M.; MešTrović, N.; Mravinac, B. Satellite DNA Evolution. *Repetitive DNA* **2012**, *7*, 126–152.
23. Pezer, Ž.; Brajković, J.; Feliciello, I.; Ugarković, Đ. Satellite DNA-Mediated Effects on Genome Regulation. *Genome Dyn.* **2012**, *7*, 153–169.
24. Biscotti, M.A.; Canapa, A.; Forconi, M.; Olmo, E.; Barucca, M. Transcription of tandemly repetitive DNA: Functional roles. *Chromosome Res.* **2015**, *23*, 463–477. [CrossRef] [PubMed]
25. Wang, S.Y.; Biesiot, P.M.; Skinner, D.M. Toward an Understanding of Satellite DNA Function in Crustacea. *Integr. Comp. Biol.* **1999**, *39*, 471–486. [CrossRef]
26. Bourque, G.; Burns, K.H.; Gehring, M.; Gorbunova, V.; Seluanov, A.; Hammell, M.; Imbeault, M.; Izsvák, Z.; Levin, H.L.; Macfarlan, T.S.; et al. Ten things you should know about transposable elements. *Genome Biol.* **2018**, *19*, 199. [CrossRef]
27. Bennetzen, J.L.; Wang, H. The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. *Annu. Rev. Plant Biol.* **2014**, *65*, 505–530. [CrossRef]
28. Deininger, P.L.; Moran, J.V.; Batzer, M.A.; Kazazian, H.H. Mobile elements and mammalian genome evolution. *Curr. Opin. Genet. Dev.* **2003**, *13*, 651–658. [CrossRef]
29. Craig, N.L.; Lambowitz, A.; Gragie, R.; Gellert, M. *Mobile DNA II*; ASM Press: Washington, DC, USA, 2002.
30. Kim, Y.-J.; Lee, J.; Han, K. Transposable Elements: No More "Junk DNA". *Genom. Inform.* **2012**, *10*, 226–233. [CrossRef]
31. Barrón, M.G.; Fiston-Lavier, A.-S.; Petrov, D.A.; González, J. Population Genomics of Transposable Elements in Drosophila. *Annu. Rev. Genet.* **2014**, *48*, 561–581. [CrossRef] [PubMed]
32. Burns, K.H.; Boeke, J.D. Human Transposon Tectonics. *Cell* **2012**, *149*, 740–752. [CrossRef] [PubMed]
33. Lanciano, S.; Mirouze, M. Transposable elements: All mobile, all different, some stress responsive, some adaptive? *Curr. Opin. Genet. Dev.* **2018**, *49*, 106–114. [CrossRef]
34. Slotkin, R.K.; Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* **2007**, *8*, 272–285. [CrossRef]
35. Wicker, T.; Sabot, F.; Hua-Van, A.; Bennetzen, J.L.; Capy, P.; Chalhoub, B.; Flavell, A.; Leroy, P.; Morgante, M.; Panaud, O.; et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **2007**, *8*, 973–982. [CrossRef] [PubMed]
36. Di Stefano, L. All Quiet on the TE Front? The Role of Chromatin in Transposable Element Silencing. *Cells* **2022**, *11*, 2501. [CrossRef] [PubMed]
37. Kojima, K.K. Structural and sequence diversity of eukaryotic transposable elements. *Genes Genet. Syst.* **2019**, *94*, 233–252. [CrossRef]
38. Flynn, J.M.; Hubley, R.; Goubert, C.; Rosen, J.; Clark, A.G.; Feschotte, C.; Smit, A.F. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 9451–9457. [CrossRef] [PubMed]

39. Smit, A.F.A.; Hubley, R.; Green, P. RepeatMasker Open-4.0 2013–2015. Available online: http://www.repeatmasker.org (accessed on 12 May 2021).
40. Bao, Z.; Eddy, S.R. Automated *De Novo* Identification of Repeat Sequence Families in Sequenced Genomes. *Genome Res.* **2002**, *12*, 1269–1276. [CrossRef]
41. Price, A.L.; Jones, N.C.; Pevzner, P.A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **2005**, *21* (Suppl. S1), i351–i358. [CrossRef]
42. Ou, S.; Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol.* **2018**, *176*, 1410–1422. [CrossRef]
43. Flutre, T.; Duprat, E.; Feuillet, C.; Quesneville, H. Considering Transposable Element Diversification in *De Novo* Annotation Approaches. *PLoS ONE.* **2011**, *6*, e16526. [CrossRef]
44. Novák, P.; Neumann, P.; Pech, J.; Steinhaisl, J.; Macas, J. RepeatExplorer: A Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **2013**, *29*, 792–793. [CrossRef]
45. Holt, C.; Campbell, M.; Keays, D.A.; Edelman, N.; Kapusta, A.; Maclary, E.; Domyan, E.T.; Suh, A.; Warren, W.C.; Yandell, M.; et al. Improved Genome Assembly and Annotation for the Rock Pigeon (*Columba livia*). *G3 Genes Genomes Genet.* **2018**, *8*, 1391–1398. [CrossRef]
46. Meng, X.; Fu, Q.; Luan, S.; Luo, K.; Sui, J.; Kong, J. Genome Survey and High-Resolution Genetic Map Provide Valuable Genetic Resources for Fenneropenaeus Chinensis. *Sci. Rep.* **2021**, *11*, 7533. [CrossRef] [PubMed]
47. Swathi, A.; Shekhar, M.S.; Katneni, V.K.; Vijayan, K.K. Genome Size Estimation of Brackishwater Fishes and Penaeid Shrimps by Flow Cytometry. *Mol. Biol. Rep.* **2018**, *45*, 951–960. [CrossRef]
48. Kawato, S.; Nishitsuji, K.; Arimoto, A.; Hisata, K.; Kawamitsu, M.; Nozaki, R.; Kondo, H.; Shinzato, C.; Ohira, T.; Satoh, N.; et al. Genome and Transcriptome Assemblies of the Kuruma Shrimp, *Marsupenaeus Japonicus*. *G3 Genes Genomes Genet.* **2021**, *11*, jkab268. [CrossRef] [PubMed]
49. Jin, S.; Bian, C.; Jiang, S.; Han, K.; Xiong, Y.; Zhang, W.; Shi, C.; Qiao, H.; Gao, Z.; Li, R.; et al. A Chromosome-Level Genome Assembly of the Oriental River Prawn, *Macrobrachium Nipponense*. *GigaScience* **2021**, *10*, giaa160. [CrossRef] [PubMed]
50. Veldsman, W.P.; Ma, K.Y.; Hui, J.H.L.; Chan, T.F.; Baeza, J.A.; Qin, J.; Chu, K.H. Comparative Genomics of the Coconut Crab and Other Decapod Crustaceans: Exploring the Molecular Basis of Terrestrial Adaptation. *BMC Genom.* **2021**, *22*, 313. [CrossRef]
51. Gutekunst, J.; Andriantsoa, R.; Falckenhayn, C.; Hanna, K.; Stein, W.; Rasamy, J.; Lyko, F. Clonal Genome Evolution and Rapid Invasive Spread of the Marbled Crayfish. *Nat. Ecol. Evol.* **2018**, *2*, 567–573. [CrossRef]
52. Shi, L.; Yi, S.; Li, Y. Genome Survey Sequencing of Red Swamp Crayfish *Procambarus Clarkii*. *Mol. Biol. Rep.* **2018**, *45*, 799–806. [CrossRef]
53. Austin, C.M.; Croft, L.J.; Grandjean, F.; Gan, H.M. The NGS Magic Pudding: A Nanopore-Led Long-Read Genome Assembly for the Commercial Australian Freshwater Crayfish, *Cherax Destructor*. *Front. Genet.* **2022**, *12*, 695763. [CrossRef]
54. Tan, M.H.; Gan, H.M.; Lee, Y.P.; Grandjean, F.; Croft, L.J.; Austin, C.M. A Giant Genome for a Giant Crayfish (*Cherax Quadricarinatus*) With Insights Into Cox1 Pseudogenes in Decapod Genomes. *Front. Genet.* **2020**, *11*, 201. [CrossRef]
55. Polinski, J.M.; Zimin, A.V.; Clark, K.F.; Kohn, A.B.; Sadowski, N.; Timp, W.; Ptitsyn, A.; Khanna, P.; Romanova, D.Y.; Williams, P.; et al. The American Lobster Genome Reveals Insights on Longevity, Neural, and Immune Adaptations. *Sci. Adv.* **2021**, *7*, eabe8290. [CrossRef]
56. Tang, B.; Wang, Z.; Liu, Q.; Wang, Z.; Ren, Y.; Guo, H.; Qi, T.; Li, Y.; Zhang, H.; Jiang, S.; et al. Chromosome-level Genome Assembly of *Paralithodes platypus* Provides Insights into Evolution and Adaptation of King Crabs. *Mol. Ecol. Resour.* **2021**, *21*, 511–525. [CrossRef]
57. Liu, L.; Cui, Z.; Song, C.; Liu, Y.; Hui, M.; Wang, C. Flow Cytometric Analysis of DNA Content for Four Commercially Important Crabs in China. *Acta Oceanol. Sin.* **2016**, *35*, 7–11. [CrossRef]
58. Jimenez, A.G.; Kinsey, S.T.; Dillaman, R.M.; Kapraun, D.F. Nuclear DNA Content Variation Associated with Muscle Fiber Hypertrophic Growth in Decapod Crustaceans. *Genome* **2010**, *53*, 161–171. [CrossRef] [PubMed]
59. Kim, J.-H.; Kim, H.; Kim, H.; Chan, B.; Kang, S.; Kim, W. Draft Genome Assembly of a Fouling Barnacle, *Amphibalanus Amphitrite* (Darwin, 1854): The First Reference Genome for Thecostraca. *Front. Ecol. Evol.* **2019**, *7*, 465. [CrossRef]
60. Chebbi, M.A.; Becking, T.; Moumen, B.; Giraud, I.; Gilbert, C.; Peccoud, J.; Cordaux, R. The Genome of *Armadillidium vulgare* (Crustacea, Isopoda) Provides Insights into Sex Chromosome Evolution in the Context of Cytoplasmic Sex Determination. *Mol. Biol. Evol.* **2019**, *36*, 727–741. [CrossRef] [PubMed]
61. Routtu, J.; Hall, M.D.; Albere, B.; Beisel, C.; Bergeron, R.D.; Chaturvedi, A.; Choi, J.-H.; Colbourne, J.; De Meester, L.; Stephens, M.T.; et al. An SNP-Based Second-Generation Genetic Map of *Daphnia magna* and Its Application to QTL Analysis of Phenotypic Traits. *BMC Genom.* **2014**, *15*, 1033. [CrossRef]
62. Tran Van, P.; Anselmetti, Y.; Bast, J.; Dumas, Z.; Galtier, N.; Jaron, K.S.; Martens, K.; Parker, D.J.; Robinson-Rechavi, M.; Schwander, T.; et al. First Annotated Draft Genomes of Nonmarine Ostracods (Ostracoda, Crustacea) with Different Reproductive Modes. *G3 Genes Genomes Genet.* **2021**, *11*, jkab043. [CrossRef] [PubMed]
63. Rasch, E.; Lee, C.; Wyngaard, G. DNA-Feulgen Cytophotometric Determination of Genome Size for the Freshwater-Invading Copepod Eurytemora Affinis. *Genome/Natl. Res. Counc. Can.* **2004**, *47*, 559–564. [CrossRef]

64. Poynton, H.; Hasenbein, S.; Benoit, J.; Sepulveda, M.; Poelchau, M.; Hughes, D.; Murali, S.; Chen, S.; Glastad, K.; Goodisman, M.; et al. The Toxicogenome of *Hyalella azteca*: A Model for Sediment Ecotoxicology and Evolutionary Toxicology. *Environ. Sci. Technol.* **2018**, *52*, 6009–6022. [CrossRef]
65. Chikhi, R.; Medvedev, P. Informed and Automated K-Mer Size Selection for Genome Assembly. *Bioinformatics* **2014**, *30*, 31–37. [CrossRef]
66. Novák, P.; Ávila Robledillo, L.; Koblížková, A.; Vrbová, I.; Neumann, P.; Macas, J. TAREAN: A Computational Tool for Identification and Characterization of Satellite DNA from Unassembled Short Reads. *Nucleic Acids Res.* **2017**, *45*, e111. [CrossRef]
67. Bao, W.; Kojima, K.K.; Kohany, O. Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes. *Mob. DNA* **2015**, *6*, 11. [CrossRef] [PubMed]
68. Demšar, J.; Curk, T.; Erjavec, A.; Gorup, Č.; Hočevar, T.; Milutinovič, M.; Možina, M.; Polajnar, M.; Toplak, M.; Starič, A.; et al. Orange: Data mining toolbox in python. *J. Mach. Learn. Res.* **2013**, *14*, 2349–2353.
69. Silva, B.S.M.L.; Picorelli, A.C.R.; Kuhn, G.C.S. In Silico Identification and Characterization of Satellite DNAs in 23 Drosophila Species from the Montium Group. *Genes* **2023**, *14*, 300. [CrossRef]
70. Pita, S.; Panzera, F.; Mora, P.; Vela, J.; Cuadrado, Á.; Sánchez, A.; Palomeque, T.; Lorite, P. Comparative Repeatome Analysis on Triatoma Infestans Andean and Non-Andean Lineages, Main Vector of Chagas Disease. *PLoS ONE* **2017**, *12*, e0181635. [CrossRef]
71. Palacios-Gimenez, O.M.; Koelman, J.; Palmada-Flores, M.; Bradford, T.M.; Jones, K.K.; Cooper, S.J.B.; Kawakami, T.; Suh, A. Comparative Analysis of Morabine Grasshopper Genomes Reveals Highly Abundant Transposable Elements and Rapidly Proliferating Satellite DNA Repeats. *BMC Biol.* **2020**, *18*, 199. [CrossRef]
72. Utsunomia, R.; Silva, D.M.Z.d.A.; Ruiz-Ruano, F.J.; Goes, C.A.G.; Melo, S.; Ramos, L.P.; Oliveira, C.; Porto-Foresti, F.; Foresti, F.; Hashimoto, D.T. Satellitome Landscape Analysis of *Megaleporinus Macrocephalus* (Teleostei, Anostomidae) Reveals Intense Accumulation of Satellite Sequences on the Heteromorphic Sex Chromosome. *Sci. Rep.* **2019**, *9*, 5856. [PubMed]
73. Sproul, J.S.; Hotaling, S.; Heckenhauer, J.; Powell, A.; Larracuente, A.M.; Kelley, J.L.; Pauls, S.U.; Frandsen, P.B. Repetitive Elements in the Era of Biodiversity Genomics: Insights from 600+ Insect Genomes. *bioRxiv* **2022**.
74. Logsdon, G.A.; Vollger, M.R.; Eichler, E.E. Long-Read Human Genome Sequencing and Its Applications. *Nat. Rev. Genet.* **2020**, *21*, 597–614. [CrossRef]
75. Paajanen, P.; Kettleborough, G.; López-Girona, E.; Giolai, M.; Heavens, D.; Baker, D.; Lister, A.; Cugliandolo, F.; Wilde, G.; Hein, I.; et al. A Critical Comparison of Technologies for a Plant Genome Sequencing Project. *GigaScience* **2019**, *8*, giy163. [CrossRef]
76. Xu, Z.; Gao, T.; Xu, Y.; Li, X.; Li, J.; Lin, H.; Yan, W.; Pan, J.; Tang, J. A chromosome-level reference genome of red swamp crayfish Procambarus clarkii provides insights into the gene families regarding growth or development in crustaceans. *Genomics* **2021**, *113*, 3274–3284. [CrossRef]
77. Wang, Q.; Ren, X.; Liu, P.; Li, J.; Lv, J.; Wang, J.; Zhang, H.; Wei, W.; Zhou, Y.; He, Y.; et al. Improved genome assembly of Chinese shrimp (*Fenneropenaeus chinensis*) suggests adaptation to the environment during evolution and domestication. *Mol. Ecol. Res.* **2022**, 334–344. [CrossRef] [PubMed]
78. Katneni, V.K.; Shekhar, M.S.; Jangam, A.K.; Krishnan, K.; Prabhudas, S.K.; Kaikkolante, N.; Baghel, D.S.; Koyadan, V.K.; Jena, J.; Mohapatra, T. A Superior Contiguous Whole Genome Assembly for Shrimp (*Penaeus indicus*). *Front. Mar. Sci.* **2022**, *8*, 808354. [CrossRef]
79. Uengwetwanit, T.; Pootakham, W.; Nookaew, I.; Sonthirod, C.; Angthong, P.; Sittikankaew, K.; Rungrassamee, W.; Arayamethakorn, S.; Wongsurawat, T.; Jenjaroenpun, P.; et al. A chromosome-level assembly of the black tiger shrimp (Penaeus monodon) genome facilitates the identification of growth-associated genes. *Mol. Ecol. Resour.* **2021**, *21*, 1620–1640. [CrossRef]
80. Yuan, J.; Zhang, X.; Li, F.; Xiang, J. Genome Sequencing and Assembly Strategies and a Comparative Analysis of the Genomic Characteristics in Penaeid Shrimp Species. *Front. Genet.* **2021**, *12*, 658619. [CrossRef]
81. Zhang, X.; Yuan, J.; Sun, Y.; Li, S.; Gao, Y.; Yu, Y.; Liu, C.; Wang, Q.; Lv, X.; Zhang, X.; et al. Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. *Nat. Commun.* **2019**, *10*, 356. [CrossRef]
82. Liu, M.; Ge, S.; Bhandari, S.; Fan, C.; Jiao, Y.; Gai, C.; Wang, Y.; Liu, H. Genome characterization and comparative analysis among three swimming crab species. *Front. Mar. Sci.* **2022**, *9*, 895119. [CrossRef]
83. Tang, B.; Zhang, D.; Li, H.; Jiang, S.; Zhang, H.; Xuan, F.; Ge, B.; Wang, Z.; Liu, Y.; Sha, Z.; et al. Chromosome-level genome assembly reveals the unique genome evolution of the swimming crab (*Portunus trituberculatus*). *GigaScience* **2020**, *9*, giz161. [CrossRef] [PubMed]
84. Bachvaroff, T.R.; McDonald, R.C.; Plough, L.V.; Chung, J.S. Chromosome-level genome assembly of the blue crab, Callinectes sapidus. *G3 Genes Genomes Genet.* **2021**, *11*, jkab212. [CrossRef]
85. Tang, B.; Wang, Z.; Liu, Q.; Zhang, H.; Jiang, S.; Li, X.; Wang, Z.; Sun, Y.; Sha, Z.; Jiang, H.; et al. High-Quality Genome Assembly of Eriocheir japonica sinensis Reveals Its Unique Genome Evolution. *Front. Genet.* **2020**, *10*, 1340. [CrossRef] [PubMed]
86. Petersen, M.; Armisén, D.; Gibbs, R.A.; Hering, L.; Khila, A.; Mayer, G.; Richards, S.; Niehuis, O.; Misof, B. Diversity and Evolution of the Transposable Element Repertoire in Arthropods with Particular Reference to Insects. *BMC Ecol. Evol.* **2019**, *19*, 11. [CrossRef]
87. Wu, C.; Lu, J. Diversification of Transposable Elements in Arthropods and Its Impact on Genome Evolution. *Genes* **2019**, *10*, 338. [CrossRef]

88. Shao, C.; Sun, S.; Liu, K.; Wang, J.; Li, S.; Liu, Q.; Deagle, B.E.; Seim, I.; Biscontin, A.; Wang, Q.; et al. The Enormous Repetitive Antarctic Krill Genome Reveals Environmental Adaptations and Population Insights. *Cell* **2023**, *186*, 1279–1294.e19. [CrossRef] [PubMed]
89. Lyon, M.F. Do LINEs Have a Role in X-Chromosome Inactivation? *J. Biomed. Biotechnol.* **2006**, *2006*, 59746. [CrossRef]
90. Dodsworth, S.; Chase, M.W.; Kelly, L.J.; Leitch, I.J.; Macas, J.; Novák, P.; Piednoël, M.; Weiss-Schneeweiss, H.; Leitch, A.R. Genomic Repeat Abundances Contain Phylogenetic Signal. *Syst. Biol.* **2015**, *64*, 112–126. [CrossRef]
91. Dodsworth, S.; Jang, T.-S.; Struebig, M.; Chase, M.W.; Weiss-Schneeweiss, H.; Leitch, A.R. Genome-Wide Repeat Dynamics Reflect Phylogenetic Distance in Closely Related Allotetraploid Nicotiana (Solanaceae). *Plant Syst. Evol.* **2017**, *303*, 1013–1020. [CrossRef]
92. Zhu, L.; Swergold, G.D.; Seldin, M.F. Examination of Sequence Homology between Human Chromosome 20 and the Mouse Genome: Intense Conservation of Many Genomic Elements. *Hum. Genet.* **2003**, *113*, 60–70. [CrossRef] [PubMed]
93. Silva, J.C.; Shabalina, S.A.; Harris, D.G.; Spouge, J.L.; Kondrashovi, A.S. Conserved Fragments of Transposable Elements in Intergenic Regions: Evidence for Widespread Recruitment of MIR- and L2-Derived Sequences within the Mouse and Human Genomes. *Genet Res* **2003**, *82*, 1–18. [CrossRef]
94. Vitales, D.; Garcia, S.; Dodsworth, S. Reconstructing Phylogenetic Relationships Based on Repeat Sequence Similarities. *Mol. Phylogenet. Evol.* **2020**, *147*, 106766. [CrossRef]
95. Bao, W.; Tang, K.F.J.; Alcivar-Warren, A. The Complete Genome of an Endogenous Nimavirus (Nimav-1_LVa) From the Pacific Whiteleg Shrimp *Penaeus* (*Litopenaeus*) *vannamei*. *Genes* **2020**, *11*, 94. [CrossRef] [PubMed]
96. Cawthorn, R.J. Diseases of American Lobsters (*Homarus Americanus*): A Review. *J. Invertebr. Pathol.* **2011**, *106*, 71–78. [CrossRef] [PubMed]
97. Clark, K.F.; Greenwood, S.J.; Acorn, A.R.; Byrne, P.J. Molecular Immune Response of the American Lobster (*Homarus Americanus*) to the White Spot Syndrome Virus. *J. Invertebr. Pathol.* **2013**, *114*, 298–308. [CrossRef] [PubMed]
98. Sotero-Caio, C.G.; Platt, R.N., II; Suh, A.; Ray, D.A. Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biol. Evol.* **2017**, *9*, 161–177. [CrossRef]

4.3 Discussion

4.3.1 Repetitive elements in Decapoda genomes in comparative light

Using my new approach combining RE databases allowed a substantial increase in the number of annotated repeats, with around 10% more REs identified compared to traditional approaches. The number of unclassified REs largely decreased, with between 1% to 20% more repeats identified. In terms of REs diversity, studies of repeatome composition commonly relied on established phylogenies to enhance visualisation and analysis. However, in my study, the calculation of the phylogenetic tree was based on the pattern of presence and absence of REs in decapods. This phylogenetic reconstruction remarkably mainly aligns with established phylogenies, suggesting a phylogenetic signal within repetitive elements. The potential for REs to harbour such a signal has been previously explored (Dodsworth et al., 2017, 2015; Vitales et al., 2020), supporting the rationale behind the strategy of creating a merged database of *ab initio* annotations, enabling the transfer of annotations between phylogenetically related sequences.

In addition, our satDNA annotation protocol disclosed a very diverse array of repeat families such as observed in *P. platypus* with 109 satDNA families. However, the occurrence of each satDNA family shows that only a few families are widely distributed in genomes, as observed in other studies (Miga, 2015; Mravinac et al., 2005; Ruiz-Ruano et al., 2016). The prevalence of satDNA families in genomes signifies their substantial contribution to genome architecture and function. Furthermore, the dynamic nature of satDNAs, characterised by rapid turnover rates and differential expansion or contraction among species, highlights their importance as evolutionary drivers (Garrido-Ramos, 2017).

Our analysis revealed a generally larger RE proportion of the decapod genomes, ranging from 58% to 79%, compared to other non-decapod crustaceans analysed in this study, which ranged from 26% to 52% (except *Armadilidium vulgare* with a RE content of 76%). The proportion of TEs in decapods also seems higher than in insects, which is typically less than 50% TEs, although exceptions exist with levels reaching up to 80% (Petersen et al., 2019; Sproul et al., 2023). The proportion of REs in decapods also appears to be higher than the levels detected in vertebrates (Chalopin et al., 2015), including mammals typically hovering around 50% (Platt

et al., 2018). On the contrary, plants present a proportion of REs closer to that of decapods, generally exceeding 50% (Luo et al., 2022).

It is generally stated that larger the genome, the more are REs present. In the case of my study, correlation analyses between TEs and genome size demonstrated a positive and significant relationship across decapods, affirming the general trend observed at different phylogenetic scales in eukaryotes (vertebrates, animals, fungi, plants, arthropods) (Chalopin et al., 2015; Elliott and Gregory, 2015; Li et al., 2017; Petersen et al., 2019; Wu and Lu, 2019). However, this trend is not observed in all taxa. In the MetaInvert project (Supplementary 1; Collins et al., 2023), the annotation of REs in 232 soil invertebrate genomes revealed that the correlation holds when considering all species, but that it is no longer significant within specific taxa like Myriapoda or Nematoda. This underscores the complexity of genome size evolution and the need to consider additional factors beyond TE content. Future investigations could explore this aspect, to elucidate potential relationships between repeat content and other genomic features such as GC content or ecological traits.

4.3.2 Limits of the chosen approach

The study was limited by the low number of decapod genomes available, which resulted in the underrepresentation of some infraorders or even the inclusion of only a single representative. Additionally, the comparison between decapods and non-decapod species could be enhanced by expanding the number of non-decapod genomes. Given that *A. vulgare* exhibited a pattern similar to that of decapods, it's possible that other groups of crustaceans share a profile closer to decapods than the non-decapod crustaceans represented in the study. For instance, the recently published genome of the Antarctic krill exhibits an exceptionally high level of REs with 92% repeats (Shao et al., 2023).

Another limitation of the study lies in the high degree of fragmentation of the assemblies used, which may have led to a significant underestimation of REs, particularly satDNAs which is notoriously difficult to assemble. SatDNAs can represent more than 30% of plant genomes and more than 50% of genomes from the animal kingdom (reviewed in Garrido-Ramos, 2017). In our study, the content of satDNAs detected in Decapoda varied from 4 to 5% only. However, an analysis of the crayfish *P. leptodactylus* repeatome estimated the proportion of satDNAs at

28% of the genome (Boštjančić et al., 2021). This study was based on analysing read sequences rather than assemblies, which may explain the observed discrepancy.

The annotation process was time-consuming, even when employing two powerful servers (64 cores and 1.47 Tb RAM each), due to factors such as genome fragmentation and the size of the assembly. For species-specific annotation with RepBase alone, which is widely used in other studies, annotation typically took around one week. In contrast, using the custom database extended the annotation process from one to two weeks, primarily due to the larger size of the database. Considering the time of annotation, despite RepeatMasker being widely employed, some optimisation strategies can enhance its efficiency for analysing large genomes and reduce annotation time. For example, utilisation of hash tables while storing data or probabilistic data structures such as Bloom filters could enable an increase in speed in processing the large amounts of data.

Despite the substantial increase in RE annotation, a considerable number of unidentified repeats remained, indicating potential for improvement of classification tools such as RepeatClassifier (Flynn et al., 2020), PASTEC (Hoede et al., 2014), or TEsorter (Zhang et al., 2022), which proved to be less time consuming than RepeatClassifier with improved annotation (Zhang et al., 2022). However, time constraints and modifications made to the classification of REs (such as reclassifying Penelope as an LTR element) made these tools inappropriate for this study. Obviously, a manual curation of REs could have significantly improved RE classification, but is notably labour-intensive, requiring approximately one week for an expert to curate a single genome. Despite its potential benefits, the resource-intensive nature of manual curation limited its feasibility within the context of this study.

4.3.3 REs impact on genomes and assemblies

The substantial presence of REs can exert diverse impacts on genome assembly processes (see Chapter 3). Notably, REs that lack full read support can pose challenges for assemblers, which must navigate to either include or exclude reads containing repeats with potentially multiple possible positions in the assembly. This can occur for example when a RE is longer than the read fragment, which means it may not be entirely recognized as an RE by the software. On the other hand, satDNA is often truncated by assemblers to their minimal size if insufficiently

supported by reads, owing to the difficulty in accurately resolving their length and leading to underestimation (Miga, 2015). Such conditions may precipitate misassemblies, as sequences may erroneously connect to disparate regions.

The prevalence of REs, typically associated with larger genome sizes (Chalopin et al., 2015; Elliott and Gregory, 2015; Li et al., 2017; Petersen et al., 2019; Wu and Lu, 2019), necessitates extensive read datasets including long reads. However, managing these voluminous datasets can strain computational resources and necessitate powerful computing infrastructure. The presence of RE-rich segments complicating assembly processes also elevates RAM usage for assemblers. Consequently, it becomes imperative to examine the ramifications of REs in genomes before assembly.

This highlights the critical need to comprehend the dynamics of repeats for accurate genome characterisation and evolutionary analysis, especially for large genomes such as the noble crayfish. By acknowledging and addressing the challenges posed by repetitive elements, assembly methodologies can be refined. This, in turn, can enhance our understanding of genomic architecture and evolutionary processes in genomes.

Chapter 5 - Comparison of Decapoda proteomes

5.1 Introduction

In the previous chapter, we highlighted the diversification of repetitive elements and their importance in the evolution of Decapoda genomes. Here, we focus on the evolution of protein coding gene repertoires in Decapoda. Comparing protein sequences across species offers a powerful method to study evolutionary conservation, allowing identification of the core biological functions and adaptations across taxa. This process relies heavily on the concept of homology, which refers to the similarity in protein sequences or structures that arise from common ancestry. Identifying homologous proteins allows to infer functional similarities and evolutionary relationships between different species. Within homologs, it is crucial to distinguish between paralogs that derived from a common ancestor after a duplication event and orthologs that emerged after a speciation event (Fitch, 1970). It is widely accepted that orthologs generally retain the same function, while paralogs frequently evolve towards different functions. This explains why the search for orthologs is at the heart of comparative genomics.

Orthology prediction

To analyse protein conservation across species, various bioinformatics tools have been developed. BLAST+ (Basic Local Alignment Search Tool) (Camacho et al., 2009) is widely used for identifying homologous sequences due to its sensitivity and customisation options. More specialised tools are needed to distinguish between orthologs and paralogs. There are three main types of approaches (reviewed in Nevers et al., 2020): (i) methods based on the analysis of graph of homology relationships between proteins after an all-against-all similarity searches between proteins from two genomes; (ii) tree-based methods that compare the gene family tree to the species tree to infer orthologs and paralogs and (iii) hybrid methods combining graph and gene trees. Tree-based approaches are generally more demanding in terms of computation and therefore not well suited to rapid comparison of large proteomes. Among graph-based approaches, OrthoMCL (Li et al., 2003) uses Markov clustering to group orthologous sequences based on similarity, making it effective for studies involving closely related species. The OrthoFinder hybrid approach (Emms and Kelly, 2019) offers a more advanced solution by combining similarity data with phylogenetic information to improve

accuracy in ortholog and paralog identification while also generating gene trees to visualize evolutionary relationships. While fast and accurate for closely related species, OrthoFinder automated approach allows limited customisation and can struggle with more distantly related species due to high divergence. OrthoInspector (Nevers et al., 2019) is a graph-based tool with an excellent balance between accuracy and sensitivity as demonstrated by a recent benchmarking (Nevers et al., 2022). OrthoInspector is optimised for whole genome orthology detection across extensive datasets and provides a pre-calculated database of orthology relations, making it highly suitable for studying broad evolutionary patterns.

Proteome comparison in Decapoda

Several massive proteome comparisons have been carried out in Arthropoda, revealing essential pathways, as well as lineage-specific adaptations that may relate to diverse habitats and ecological roles. For example, a large-scale study on 76 arthropods revealed some highly conserved proteins with core metabolic pathways among arthropods and the emergence of numerous gene families in insects such as 1038 gene families in the Lepidoptera ancestor with an enrichment in odorant binding function (Thomas et al., 2020). It also have been revealed that carbohydrate-active enzyme content is correlated with herbivorous adaptations in 815 arthropods (Ojeda-Martinez et al., 2024). These valuable studies are however mainly focused on insects with a limited number of crustacean species and very few, if any, Decapoda species. At the decapod level, proteome comparison between the blue king crab *Paralithodes platypus* and 12 other Arthropoda species highlighted an expansion of different genes families in the blue king crab (Tang et al., 2021). These expanded families included genes involved in inflammatory regulation that could explain the strong environmental adaptation ability of king crabs. However, this comparative analysis only includes four Decapoda species. Decapods counts 17 genomes with annotated proteins (see Chapter 2.3.2). The overall protein-coding gene count in decapod species is typically around 20,000 to 25,000 genes (see Chapter 2.1.3). The exception is for two crab species that stand out with significantly larger gene counts, estimated at 40,000 genes each. At the time of writing, no large-scale study focusing on protein conservation in decapod species has been conducted. Nor are there any massive comparative studies between the proteomes of decapods and more distantly related eukaryotic species, that would allow a better understanding of both shared and unique features of these organisms and provide valuable insights into their evolution and biology.

In this study, I present a comprehensive comparative analysis of proteins among decapod species and across a broad range of eukaryotic species from the main kingdoms. This approach allows us to explore not only the core, evolutionarily conserved proteins within Decapoda but also to determine the functional features of genes conserved across specific groups of taxa.

5.2 Material and methods

5.2.1 Phylogenetic profiling of Decapoda proteomes

We selected the 15 annotated proteomes of Decapoda species that had a complete Busco score of more than 60% for both genome assembly and protein annotation. These 15 proteomes were compared to each other to determine the pairwise orthology relationships within Decapoda. They were also compared with the public version of the OrthoInspector eukaryote database containing 1472 species (Table 5-1) to determine orthologous proteins across a wide range of species. Using this custom version of the Eukaryotic database of OI, we generated the phylogenetic profile of each protein, i.e. the presence or absence of orthologs of the considered protein in the set of eukaryotic proteomes. The profiles where then split by species, generating a binary matrix where each row represents a gene, each column represents a eukaryotic species, and the “1” or “0” in each row refers to the presence or absence of the ortholog in each species, respectively.

Table 5-1: Composition of the OrthoInspector eukaryote database

Taxons	Number of species
Discoba	17
Metamonada	5
Metazoa	438
Fungi	734
Other Opisthokonta	4
Amoebozoa	9
Apusozoa	1
Haptista	3
SAR	83
Cryptophyta	1
Rhodophyta	5
Viridiplantae	172
Total	**1472**

5.2.2 Analysis of orthology relationships among Decapoda proteomes

We used phylogenetic profiles and taxonomic information contained in our customized OI database to determine the core proteome of Decapoda. The Decapoda core proteome was defined as the set of protein families conserved in at least 13 decapod species out of the 15 to take account of the fragmentation of some genomes. The remaining proteins were hierarchically analysed to determine whether they were conserved at lower taxonomic ranks. For Pleocyemata, Dendrobranchiata, Astacidea, Brachyura, the threshold was respectively 8 out of 10, 4 out of 5, 2 out of 3, and 3 out of 4 species. For each group of conserved proteins, we defined the subset of specific proteins. Proteins were considered conserved and specific if they were present in fewer than 15 species outside their group of species.

Each relevant group of conserved proteins in *P. japonicus* and *P. clarkii* was then used to make an enrichment analysis in STRING database (Szklarczyk et al., 2023) using the uploaded corresponding protein set. For technical reasons, if the number of proteins exceeded 2000, a random set of 2000 proteins was taken to make the analysis. The proteins corresponding to the main enrichment were then used to create a heatmap to obtain a detailed view of the distribution of orthologs across species.

5.2.3 Heatmap generation, gene clustering and functional analysis

To visualize and analyse at larger scale the proteome of Decapoda, we selected the crayfish *P. clarkia*. We used seaborn (Waskom, 2021) to generate the heatmap representing the phylogenetic profile of each protein of *P. clarkii* in all species present in our customized OI database. The Ward hierarchical clustering method based on the Euclidean distances between phylogenetic profiles was used to cluster proteins. For clarity, the species-specific proteins, showing only absence of orthologues were removed from the dataframe. To order the dataframe according to taxonomy, a dendrogram of OI species was computed using taxonomic information contained in OI.

The heatmap was then decomposed into 20 different clusters of proteins. Each protein cluster was then analysed with STRING (Szklarczyk et al., 2023) to make a functional enrichment

analysis. As reference set for the enrichment analysis, we used the *P. clarkii* proteome by uploading the proteome to the STRING database.

5.3 Results and discussion

5.3.1 Evolution of proteomes in Decapoda

Among the 17 Decapoda species presenting an annotated genome (with gene prediction and protein annotation), I selected 15 species with a BUSCO score (arthropoda_odb10) higher than 70% (

Table 5-2). This threshold was applied in order to have proteomes of reasonable quality while retaining most of the species to make the comparison. For all species, we used the protein set from GenBank (Clark et al., 2015) or RefSeq (O'Leary et al., 2016), except for *Chionoecetes opilio.* The proteome of the latter was retrieved from the Uniprot database (Boutet et al., 2007) and was classified as Standard by the Complete Proteome Detector (CPD) with the most similar statistics to those observed in GenBank or RefSeq. These proteomes were added to the eukaryotic database of OrthoInspector (OI) to construct a custom database and generate the profile of presence or absence of orthologs among eukaryotic species.

Table 5-2: Studied Decapoda species.

order/infraorder		Name	Busco genome	Busco protein	Number of proteins
Dendrobranchiata		*Penaeus japonicus*	93.6	98.0	22301
		Penaeus chinensis	90.7	95.6	20076
		Penaeus vannamei	84.8	93.5	24987
		Penaeus monodon	83.9	91.2	24011
		Penaeus indicus	88.6	92.7	21824
Pleocyemata	Caridea	*Halocaridina rubra*	87.9	81.8	25341
	Astacidea	*Procambarus clarkii*	94.3	98.0	26417
		Homarus americanus	93.0	97.4	22368
		Cherax quadricarinatus	69.9	88.2	18152
	Anomura	*Petrolisthes manimaculis*	92.8	94.8	40296
		Petrolisthes cinctipes	92.0	94.3	44511
	Brachyura	*Portunus trituberculatus*	93.5	97.0	17292
		Eriocheir sinensis	92.6	97.9	19615
		*Chionoecetes opilio**	91.0	62.1	22650
		Scylla paramamosain	96.0	98.7	19839

*proteome from UniProt

I searched for both conserved proteins (CNS), and conserved and specific proteins (CS) in some specific taxonomic groups (Figure 5-1). For each species, the proteins were first categorised and counted as CNS in crustaceans if present in at least 22 species, and CS if they weren't present in more than 15 other species. These proteins were then removed for the dataset to search CNS and CS proteins to Decapoda (≥ 13), Dendrobranchiata (≥ 4) for Dendrobranchiata species, and in Pleocyemata (≥ 8) for Pleocyemata species. Once again, proteins categorised in Dendrobranchiata or Pleocyemata were removed from the dataset. Then for Pleocyemata we categorised the proteins according to their respective suborder as Astacidea (≥ 2) or Brachyura (≥ 3). Once these proteins were removed from the dataset, the proteins without orthologs were categorised as specific to the species.

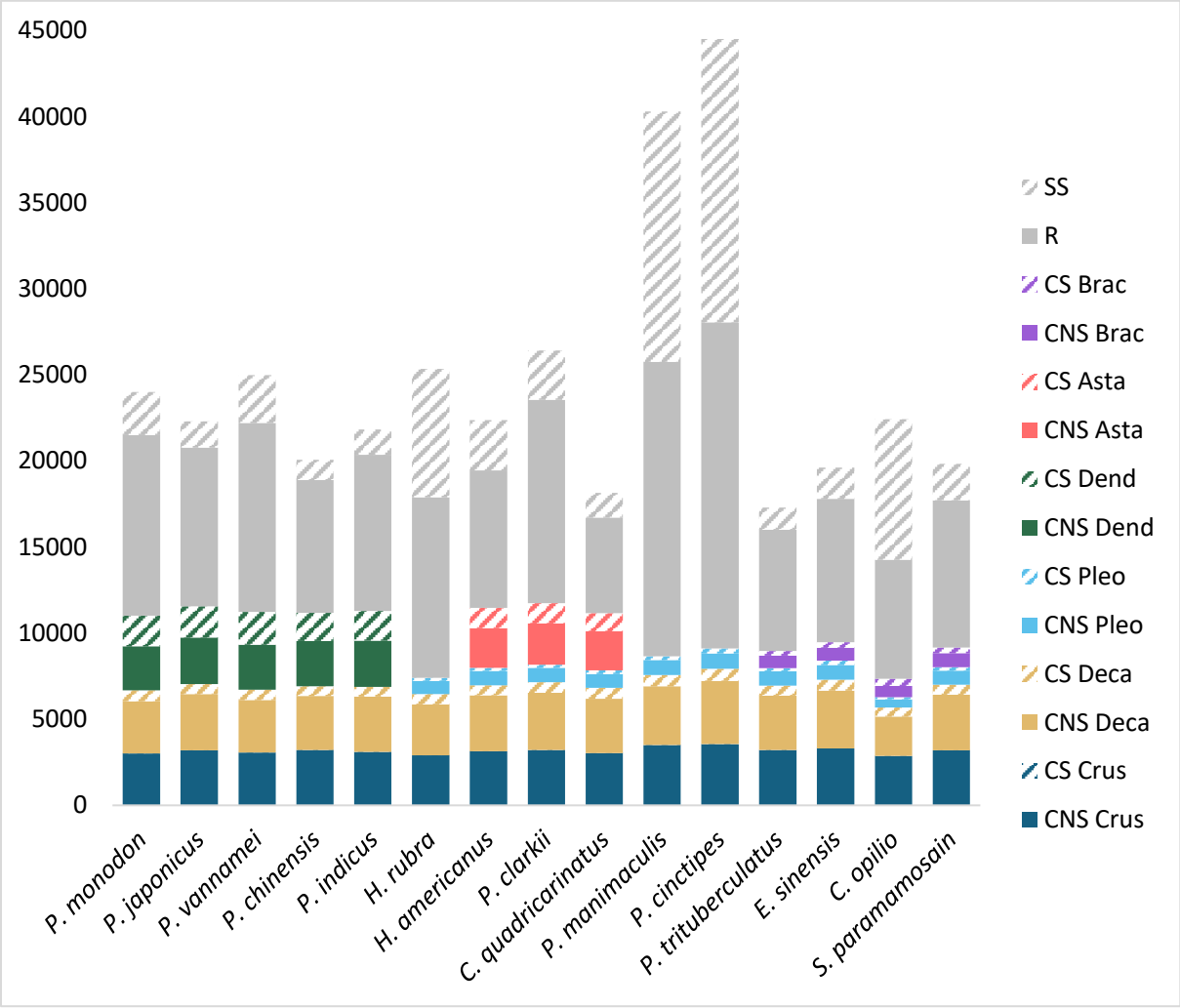

Figure 5-1: Decapoda conserved proteins. CNS: Conserved non-specific, CS: conserved and specific. For each species, the number of CNS and CS are indicated at different taxonomic levels: Crustaceans (Crus), Decapoda (Deca), Pleocyemata (Pleo), Dendrobranchiata (Dend), Astacidea (Asta), Brachyura (Brac).

Number of CNS proteins in crustacean species varies from 2 871 for *C. opilio* to 3 553 for *P. cinctipes* with a mean of 3 168 (Figure 5-1, Supplementary 2). A detailed view of the phylogenetic distribution of this set of proteins in *P. japonicus* and *P. clarkii* (Figure 5-2) shows that most of these proteins are well conserved in metazoan, including proteins conserved in all eukaryotes. Enrichment analyses in *P. japonicus* highlight essential general functions such as gene expression with the presence of 359 proteins in the cluster on the 1 251 proteins related to this biological process and a signal of 2.25. Results are similar for *P. clarkii* with the main enrichments related to organelle organisation (230/861 proteins, signal 2.33) and gene expression (341/1 396 proteins, signal 2.26).

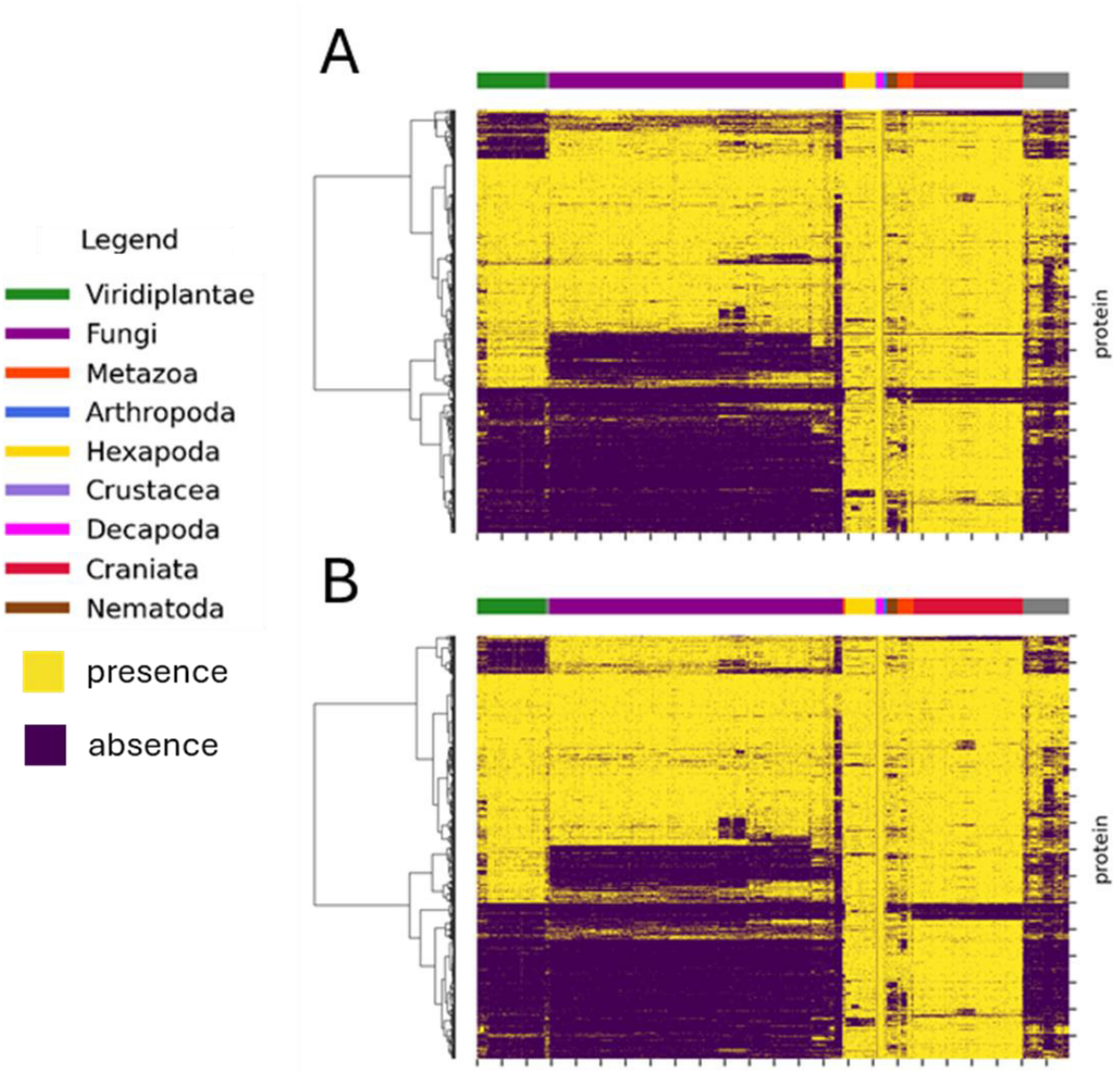

Figure 5-2: Heatmap of conserved proteins in crustaceans. A. *Penaeus japonicus*. B. *Procambarus clarkii*. Each row corresponds to a gene and each column correspond to an eukaryotic species. Presence of an ortholog is indicated by a yellow dot while absence is

indicated by a purple dot. Species are ordered and coloured according to taxonomy. The dendrogram show the clustering of the proteins profile.

The numbers of CS are really low in crustaceans with 1 to 8 proteins depending on the species. Even if most of these crustacean-specific proteins are uncharacterized, some of them are annotated as neurotrophin 1 or neurotrophin, and spaetzle 3 (Table 5-3). Neurotrophins and spaetzle belong to the neurotrophin superfamily that includes numerous paralogous families. Neurotrophins are known to regulate neuronal survival, targeting, synaptic plasticity, memory and cognition and maintains neuronal survival in flies (Zhu et al., 2008). They have been showed to be highly conserved in insects. Our results suggest the emergence of several specific neurotrophin families in the common ancestor of Crustacea.

Table 5-3: Annotated proteins conserved and specific to Crustacea. Only proteins being annotated are showed.

	Species	CS Crustacea	Protein with annotation
Dendrobranchiata	*P. monodon*	4	neurotrophin 1-like, neurotrophin 1-like, neurotrophin 1-like
	P. japonicus	4	neurotrophin 1-like, neurotrophin 1-like, neurotrophin 1-like
	P. vannamei	3	ecdysone-inducible protein E75-like, PHD finger protein rhinoceros-like
	P.chinensis	2	adhesion G protein-coupled receptor L2-like
	P. indicus	1	
Caridea	*H. rubra*	4	Spaetzle, Spaetzle
Astacidea	*H. americanus*	5	neurotrophin 1-like, neurotrophin 1-like, neural-cadherin-like
	P. clarkii	2	ribosome-binding protein 1-like (predicted)
	C. quadricarinatus	3	
Anomura	*P. manimaculis*	8	
	P. cinctipes	1	
Brachyura	*P. trituberculatus*	2	neurotrophin 1-like
	E. sinensis	6	protein spaetzle 3-like, protein spaetzle 3-like, neurotrophin 1-like
	C. opilio	1	
	S. paramamosain	3	neurotrophin 1-like, neurotrophin 1-like

CS = conserved and specific.

Number of CNS proteins in decapod species (after removal of proteins conserved in Crustacea) varies from 2 288 for *C. opilio* to 3 664 for *P. cinctipes* with a mean of 3 165 proteins. These conserved proteins (CNS + CS) in *P. japonicus* and *P. clarkii* show an overall good conservation within metazoan species (except from non-decapoda crustacean species) but some of them are absent from Nematoda and Hexapoda (Figure 5-3). The enrichment analyses only reveal general metabolic processes such as cellular metabolic process (695/4 767 proteins, signal 1.1 for *P. japonicus*, 708/4 504 proteins, signal 1.53 for *P. clarkii*), and for *P. clarkii* small molecule metabolic process (174/892, signal 1.54), cellular lipid metabolic process (105/469 proteins,

signal 1.54). The mean number of CS proteins is 600. As in the case of CNS, extreme values for CS proteins to decapod species are observed in *C. opilio* and *P. cinctipes,* both Pleocyemata species, with respectively 516 and 711 proteins. The higher number of CNS and CS observed in *P. cinctipes* can be related to the higher number of identified proteins in the genome. It suggests recent expansions of some gene families in this genome. The lowest number of CNS and CS observed in *C. opilio* can be linked to the lower BUSCO score observed at the protein level in this species despite a high BUSCO score at the assembly level. It could reflect recent gene losses or more probably a low-quality annotation with unpredicted genes. Among decapods, with a mean of 2 647 CS proteins, Dendrobranchiata present a generally higher number of CS proteins in decapod species than Pleocyemata that have a mean of 795 CS proteins.

By combining the sets of conserved protein families (CNS and CS) in crustaceans and decapods, we can delineate a Decapoda core proteome of around 7 000 proteins, corresponding to half the Crustacean core proteome.

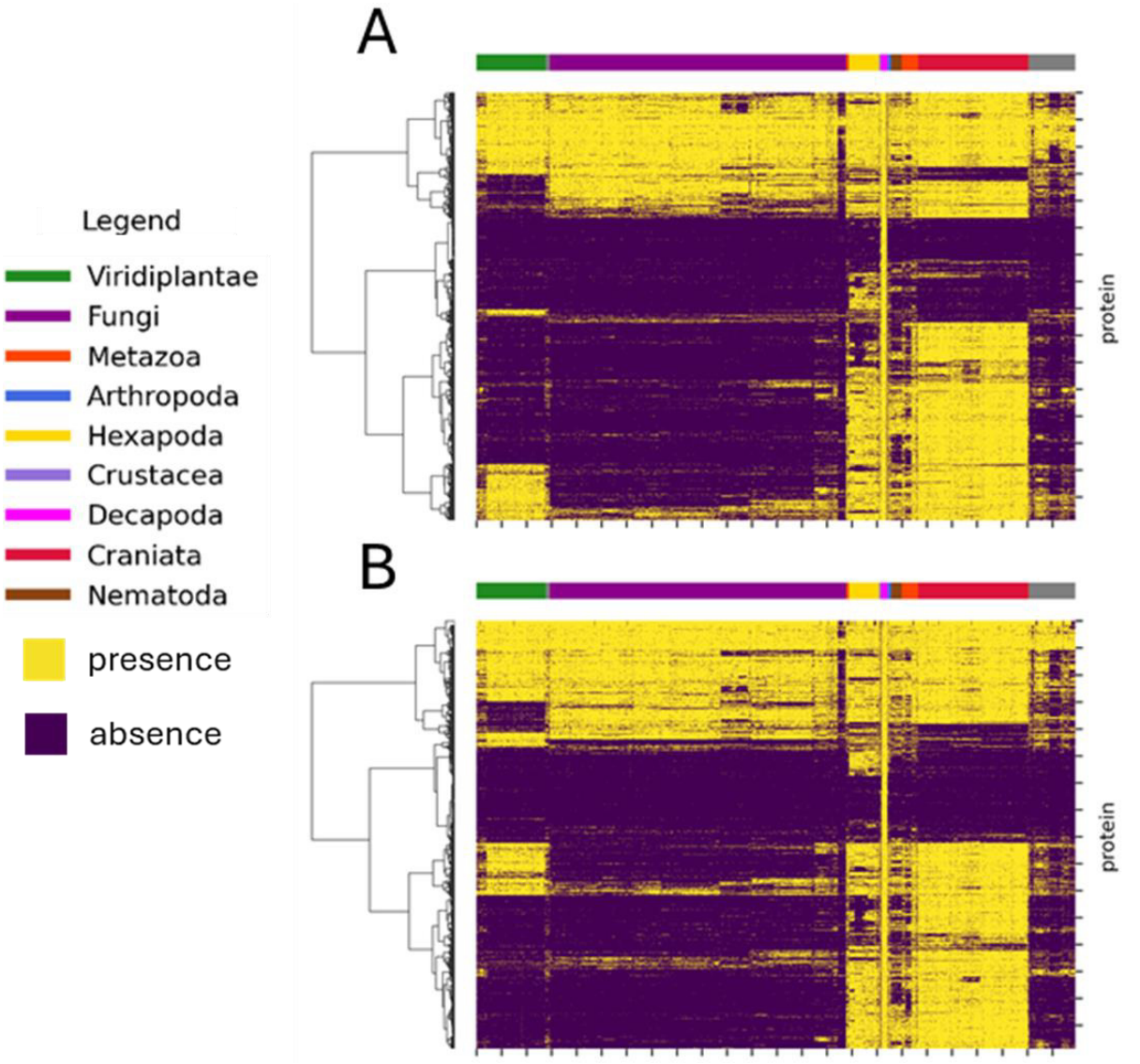

Figure 5-3: Heatmap of conserved proteins in Decapoda. A. *Penaeus japonicus*. B. *Procambarus clarkii*. Each row corresponds to a gene and each column correspond to an eukaryotic species. Presence of an ortholog is indicated by a yellow dot while absence is indicated by a purple dot. Species are ordered and coloured according to taxonomy. The dendrogram show the clustering of the proteins profile.

Dendrobranchiata species show homogeneous numbers of CNS and CS proteins with a mean of 2 647 and 1763 respectively. Conserved proteins of Dendrobranchiata (CNS + CS) still show some groups of proteins conserved among metazoan or across most eukaryotics species (Figure 5-4). Some proteins are also conserved in plants and metazoan but absent in fungi. However, an important set of proteins is restricted to arthropods. No functional enrichment was detected in this set. Among Pleocyemata, number of CNS is homogeneous too with an average number of 841 proteins. The 770 conserved proteins of *P. clarkii* (CNS + CS) show

similar phylogenetical profile than *P. japonicus* in Dendrobranchiata (Figure 5-4). Similarly to *P. japonicus*, no specific functional enrichment was detected. The number of CS in Pleocyemata is variable depending on the infraorder. Highest values can be explained by some recent gene duplications generating the presence of inparalogs. On the opposite, lowest values could be explained by recent gene lost. The higher amount CNS and CS in Dendrobranchiata than in Pleocyemata species can also be explained by the fact that the four considered species belong to the same genus with a recent common ancestor compared to the more divergent Pleocyemata species that are separated into different infraorders.

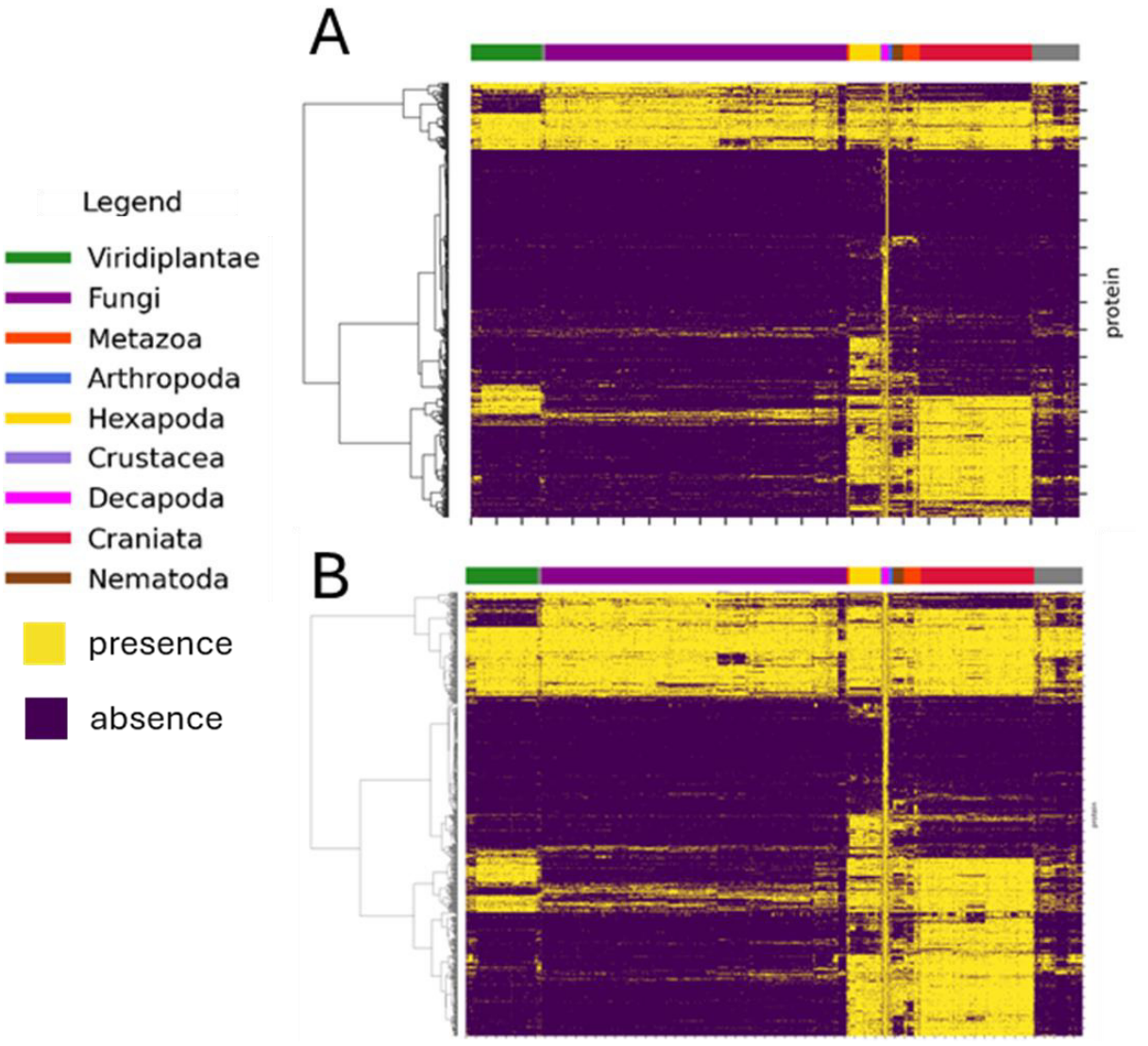

Figure 5-4: Heatmap of conserved proteins in Dendrobranchiata and Pleocyemata. A. *Penaeus japonicus*. B. *Procambarus clarkii*. Each row corresponds to a gene and each column correspond to an eukaryotic species. Presence of an ortholog is indicated by a yellow dot while absence is indicated by a purple dot. Species are ordered and coloured according to taxonomy. The dendrogram show the clustering of the proteins profile.

Among Pleocyemata infraorders, Brachyura species show the lowest amount of CNS and CS proteins with a mean of 745 and 325 respectively. This is much lower than in the three Astacidae species which have an average of 2,316 CNS and 1,145 CS. Although these Astacidae representatives belong to three different superfamilies, our study shows a significant set of common proteins, similar to that observed in the Dendrobrachiata with the four *Penaeus* species.

Interestingly, the two *Petrolisthes* species, that have a significantly higher number of proteins (> 40 000 proteins), also have a higher number of species-specific genes (> 14 500 proteins). This high number of species-specific proteins can be explained by an overestimation of the number of genes due to an incomplete assembly or annotation errors with a lot of false positives (Ko et al., 2022; Scalzitti et al., 2020). *C. opilio* that were mainly presenting less conserved proteins among the different studied taxa, is also showing a large number of species-specific proteins with 8 156 of them. With a similar amount (7 453 proteins) the Caridea *H. rubra* is also presenting a large number of species-specific proteins, however, as the only representant of its infraorder, the number of CNS and CS couldn't be estimated for Caridea. All other species show less than 3 000 specie specific proteins with a mean of 1 992 proteins per species.

5.3.2 Analysis of *Procambarus clarkii* phylogenetic profiles

Using the phylogenetic profiles of *P. clarkii*, we generated a heatmap to visualise the presence and absence of orthologues across various eukaryotic species, allowing us to explore patterns in protein conservation (Figure 5-5). After clustering these proteins based on profile similarity, we conducted enrichment analyses to identify biological processes associated with specific groups of proteins (Table 5-4).

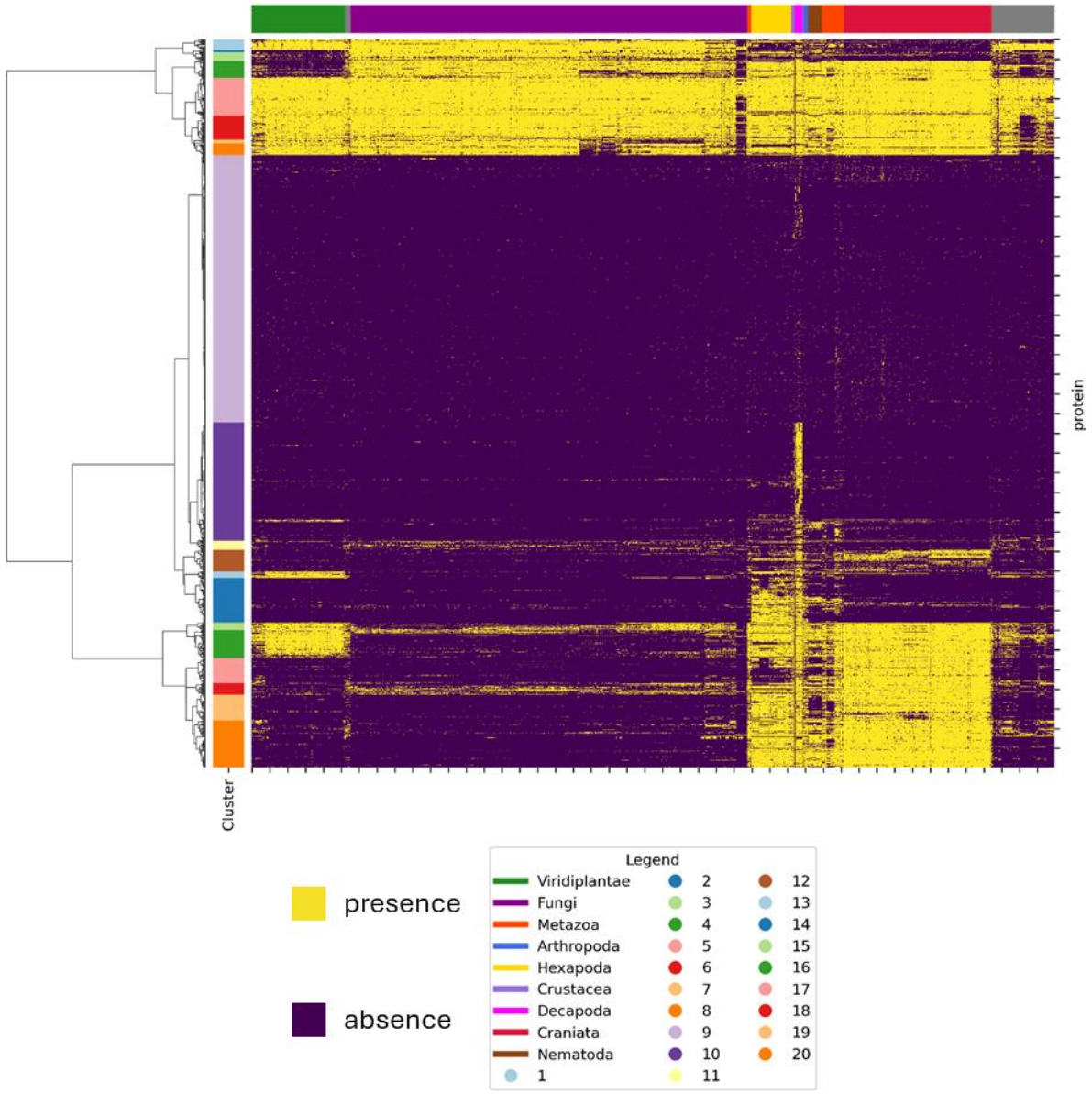

Figure 5-5: Heatmap of the phylogenetic profiles of *Procambarus clarkii*. Each row corresponds to a gene and each column correspond to an eukaryotic species. Presence of an ortholog is indicated by a yellow dot while absence is indicated by a purple dot. Species are ordered and coloured according to taxonomy. The dendrogram show the clustering of the proteins profile and clusters are annotated and differentiated by colours.

Table 5-4: Enrichment analysis by cluster. Number of proteins in enrichment represents the number of proteins from the cluster that present this annotation against the total number of proteins of the genome fitting in this category.

	number of proteins in cluster	Enrichment annotation	number of proteins in enrichment	signal
Cluster 5	1217	Cytoplasmic translation	67/91	6
Cluster 6	784	Cellular amino acid biosynthetic process	22/53	2,72
Cluster 8	379	Regulation of G2/M transition of mitotic cell cycle	41/79	7,65
Cluster 7	111	Glycerophospholipid metabolic process	13/126	2,47
Cluster 15	245	Spliceosomal complex disassembly	9/11	2,85
Cluster 16	907	Nucleic acid metabolic process	174/1485	1,97
Cluster 4	558	Carboxylic acid metabolic process	58/426	2,56
Cluster 1	330	Nucleosome assembly	86/241	9,96
Cluster 2	97	NA		
Cluster 18	372	Peptide catabolic process	13/25	2,89
Cluster 3	267	Cellular hormone metabolic process	14/37	3,04
Cluster 19	838	Protein insertion into ER membrane	30/37	4,85
Cluster 20	1514	Transpcrition initiation from RNA polymerase II promoter	68/105	4,65
Cluster 17	813	Regulation of intrinsic apoptotic signalling pathway	15/31	2,04
Cluster 12	695	Regulation of DNA recombination	24/78	2,4

Cluster 13		213	Glycosphingolipid biosynthetic process	12/22	3,49
Cluster 14		1435	Chitin-based cuticle development	57/176	2,45
Cluster 11		301	Mitotic cytokinesis	20/81	3,14
Cluster 10	2	689	Sensory perception of chemical stimulus	44/136	3,83
	1_1	617	Response to reactive oxygen species	17/59	1,75
	1_2	413	No biological process enrichment		
	1_3	1198	No biological process enrichment		
Cluster 9		8647	-		

As expected, some proteins (clusters 5, 6, 7, 8) are highly conserved across all represented eukaryotic kingdoms, indicating that the corresponding genes are essential and were present in the last eukaryotic common ancestor. Cluster 5 and 6 present an enrichment in proteins associated with cytoplasmic translation and cellular amino acid biosynthesis respectively. These functions are related to fundamental roles in cellular machinery and involved in processes in maintaining cellular health and metabolic balance. The **cluster 5** shows mainly proteins annotated as being ribosomal proteins. For **cluster 6**, the two main annotations are NADH dehydrogenase complex proteins (playing a crucial role in cellular respiration and energy production) and ATP synthase and associated factors (also essential for cellular energy production in the mitochondria) (Rich, 2003; Tielens and Van Hellemond, 1998). **Cluster 7 and 8** are also largely conserved across all species, however, these clusters are characterized by gene losses in some lineages, in particular in some groups of fungi and in some hexapods in the case of the cluster 8. These clusters are also related to fundamental roles in cellular machinery with enrichment in glycerophospholipid metabolic process, and the regulation of G2/M transition of mitotic cell cycle.

Several clusters are well conserved in metazoan but exhibit diverse conservation patterns in other kingdoms. **Cluster 4** is conserved in fungi and metazoan species and enriched in proteins related to carboxylic acid metabolic process, illustrating the close proximity between Metazoa

and Fungi that both belong to Opisthokonta. **Cluster 15**, conserved in plants, metazoans and some fungi, is enriched is spliceosomal complex disassembly proteins. Cluster 16 is well conserved in metazoans and plants, and almost absent in fungi with an enrichment in nucleic acid metabolic process. This suggests a different evolution of fundamental generic process in fungi compared to plants and metazoans. **Cluster 18**, conserved in metazoan species and less conserved in fungi, is enriched in peptide catabolic process, highlighting its role in protein degradation pathways. **Cluster 19** is almost exclusively conserved in metazoans, excluding nematodes and some other metazoans with proteins enriched in protein insertion into the endoplasmic reticulum membrane, reflecting cellular complexity in metazoan organisms. **Cluster 20**, conserved in metazoans too, is enriched in transcription initiation from RNA polymerase II promoter, a fundamental process in eukaryotic transcriptional regulation. **Cluster 17** is conserved mainly in craniate species and decapods, also in some nematodes and hexapods, proteins are enriched in regulation of intrinsic apoptotic signalling pathway.

Other clusters present a patchy distribution within metazoan species, highlighting evolutionary diversification of some specific pathways or processes. **Cluster 12** is well conserved in crustaceans but poorly conserved in other protostomes while more present in deuterostome species. This cluster is enriched in regulation of DNA recombination. **Cluster 14**, present in arthropods and some nematodes present an enrichment in chitin-based cuticle development proteins. **Cluster 11**, enriched in mitotic cytokinesis proteins, is well conserved in Decapoda but exhibits a patchy distribution in other species.

Four clusters (1, 2, 3, 13) exhibit particularly atypical distribution. **Cluster 1** is well conserved in plants, most fungi and some protists but strikingly, it is mainly absent in Craniata species and sparsely distributed in protostomes. This cluster is enriched in proteins related to nucleosome assembly. This disparity could indicate specificities in chromatin dynamics and gene regulation in decapods and some protostomes compared to craniate. We might think of the peculiarities noted in spermatogenesis in decapods, with spermatozoa characterised by a nucleus with a uncondensed chromatin nucleus (Chen et al., 2020), or the chromatin reduction reported in some protostome species including crustaceans (Grishanin, 2024). The small **cluster 2**, conserved in fungi and nematode species, doesn't present any enrichment. **Cluster 3** mainly conserved in fungi and decapods presents an enrichment in cellular hormone

metabolic process. **Cluster 13** exhibits an atypical distribution, being mainly present in plants and some arthropods. The cluster is enriched in glycosphingolipid biosynthetic process. Corresponding genes could be involved in the synthesis of arthropod-specific glycosphingolipids (Kimura et al., 2014) but the presence of orthologs of these genes in plants remains enigmatic.

Finally, the large clusters 9 (8647 proteins) and 10 (2917 proteins) exhibit a very restricted distribution. Proteins from **cluster 9** are present only in a few disparate species. **Cluster 10** is of particular interest since it is conserved and specific to decapods. As it contains too many proteins for an enrichment analysis in STRING, it was further clustered into three subclusters. it Cluster 10 **subcluster 1_1** presents some orthologues in other species and the enrichment analysis showed an enrichment in proteins related to response to reactive oxygen species. Subclusters 1_2 and 1_3, strictly conserved in decapods, don't present any biological process enrichment. This absence of enrichment can be due to poor annotation of these proteins. The **subcluster 1_2** presents 175 annotated proteins out of 414 proteins. Among all the proteins, several proteins are annotated as Phospholipase D Beta and integrin Alpha-4. Phospholipase D is involved in lipid metabolism and membrane signalling, aiding cellular responses to stress and damage (Jenkins and Frohman, 2005). Integrin Alpha-4 plays a role in cell adhesion, signalling, and immune responses (Hynes, 2002). The **subcluster 1_3** is composed of 509 annotated proteins out of 1199 with several proteins annotated as glutamate receptor ionotropic and E3 Ubiquitin-Protein Ligase RNF168. Glutamate Receptor Ionotropic mediates excitatory neurotransmission, vital for processes such as learning, memory, and sensory perception such as chemical cues in aquatic environments (Lüscher and Malenka, 2012). E3 Ubiquitin-Protein Ligase RNF168 plays a role in DNA damage response, signalling, and repair by tagging damaged proteins for degradation (Brinkmann et al., 2015). This cluster now presents an enrichment in cellular detoxification of nitrogen compound with 7 out of 12 proteins and a signal of 0.96. Finaly, the **subcluster 2** from cluster 10 show an enrichment in sensory perception of chemical stimulus.

5.4 Conclusions and perspectives

In this study, I conducted a comparative analysis of the Decapoda proteomes, examining the evolution of protein-coding genes both within Decapoda and across a broader set of eukaryotic taxa. Our study has provided a first overview of the Decapod core proteome. Based on the genomes available at the time of the study, this represents around 7000 proteins, i.e. approximately a quarter of a Decapod proteome. This number will certainly evolve with the arrival of new genomes, and in particular genomes from representatives of other infra-orders and superfamilies. The comparison of Pleocyemata proteomes reveals a wide diversity of gene repertoires, since the proteins conserved in the latter represent a restricted pool of around 795 proteins. Regarding the Dendrobranchiata, it is difficult to draw any general conclusions about gene conservation, as the only Dendrobranchiata genomes available at the time of the study were limited to the *Penaeus* genus. However, the number of conserved proteins appears very limited between species of the same genus. These results highlight the evolutionary and functional diversity among decapod species and suggest a range of lineage-specific adaptations.

Considering broader conservation patterns, phylogenetic profiling of the *P. clarkii* proteome has revealed the diversity of evolutionary histories of Decapoda genes and a large pool of Decapod-specific genes. Specific clusters within Decapoda, particularly those enriched in pathways like cellular detoxification and chemical stimulus perception, point to adaptations that may contribute to the ecological success and resilience of this group. Phylogenetic profiling also revealed gene clusters with atypical phylogenetic distributions such as genes involved in nucleosome assembly that constitute a promising avenue for understanding chromatin dynamics in these species.

This study represents the first large-scale proteome comparison specifically focused on Decapoda and their evolutionary relationships within Arthropoda and the main eukaryotic kingdoms. Our initial profiling proved successful in identifying gene modules that share a similar evolutionary history. However, some clusters are more heterogeneous. In the future, it would be possible to test other methods of evaluating the distances between clusters and/or to carry out a second round of clustering to refine clusters and detect the associated functions more precisely. In addition, we have analysed the presence/absence of orthologs without

specifically studying recent duplications, i.e. inparalogs that reveal gene family expansions. Even though it can be improved, our phylogenetic profiling provides a valuable basis for more detailed functional studies of relevant clusters. In addition, phylogenetic profiles can be used to predict functional link (Pellegrini et al., 1999) between genes. Given the large number of proteins with unknown functions in Decapoda, as in many non-model species, such approach could be precious in identifying proteins involved in interesting pathways or complexes.

Chapter 6 – Genome assembly of the noble crayfish, *Astacus astacus*

6.1 Choice of the species

As seen in Chapter 1 – Crayfish, keystone species in aquatic ecosystems, crayfish are keystone species and environmental engineers of freshwater ecosystems. The noble crayfish is the most widespread crayfish species in Europe. It stands as an emblem of Europe but faces critical endangerment, listed on the IUCN Red List Index of Threatened Species (Edsman et al., 2010; Füreder et al., 2010). Over the past 150 years, its populations have decreased by over 95%, nearing extinction in France and experiencing a 56% decline in range in northwest Germany over 22 years. Despite their ecological importance, crayfish are still poorly studied at the genomic level. Currently, only four genomes have been assembled (*Procambarus virginalis*: 3.7 Gb (Maciaszek et al., 2022), *Procambarus clarkii*: 4.0 Gb (Liao et al., 2024), *Cherax destructor*: 3.3 Gb (Austin et al., 2022), and *Cherax quadricarinatus*: 3.9 Gb (Liu et al., 2024)), none of which represent European species that are experiencing population declines.

To safeguard the biodiversity and ecological integrity of aquatic ecosystems in Europe, the implementation of effective management and prevention strategies of invasive crayfish have become an urgent imperative. To achieve this goal, it is essential to gain comprehensive insights into the behavioural and molecular distinctions existing among both European and non-European crayfish species, as well as within European species themselves. This will allow genomic comparison between species and populations to discriminate possible genes responsible to resistance to the crayfish plague. In this regard, acquiring the genome of a representative European species is of paramount importance, as it enables a conservative genomic approach that is vital for informed conservation efforts.

Considering the literature, the genome size of the noble crayfish was initially estimated between 2 Gb to 3 Gb (Gregory, 2023; Gutekunst et al., 2018; Tan et al., 2020). This is why an assembly of 50x coverage of short reads with a scaffolding with 10x of long reads was initially planned when the sequencing project started in 2020. Some assemblies trial was done, and then a flow cytometry analysis was conducted, estimating a genome size of 17 Gb (Theissinger et al., unpublished results). As seen in Chapter 2, sequencing, assembly and annotation methods have rapidly evolved, allowing better exploitation of genomic content. However,

Chapter 3 highlights the challenges of sequencing and assembling large genomes such as the noble crayfish. Indeed, considering the size of the noble crayfish genome, numerous REs are expected. Moreover, our study on REs in decapod genomes (Chapter 4) highlights the generally high presence of REs in crayfish genomes, being 40% to 70% (Chapter 4; Rutz et al., 2023). Large genomes presenting numerous REs are complicated for both the sequencing, with a need of a combination of sequencing from different platforms, and the assembly, regarding the amount of data and unresolved REs. With advances of technologies and the corrected genome size estimation, both sequencing and assembly strategies had been adapted. We present here the evolution of the assembly of the noble crayfish genome.

6.2 Sequencing

6.2.1 Illumina

The noble crayfish specimen used for the genome sequencing was a male from Finland population of the lake Rytky. DNA isolation was done on muscle tail tissue based on a phenol-chloroform extraction protocol. We aimed at sequencing 50x coverage with paired-end Illumina reads of 150 bp with an insert size of 350 bp. To accomplish this, we employed different sequencing platforms. Initially, we used the NovaSeq platform for the first three sequencing runs. However, many of the reads contained polyX sequences, where a single nucleotide was repeated multiple times. In some cases, the reads consisted of just one repeated nucleotide, the guanine. This issue could be attributed to the NovaSeq platform's use of only two filters for detecting nucleotide fluorophores, as discussed in Chapter 2. Due to the polyX presence, we switched to the HiSeq platform for a new run. Although the occurrence of polyX sequences was significantly reduced, a high number of duplicated reads occurred, which resulted in a similar number of reads being discarded with the HiSeq platform than to the NovaSeq platform. Considering this, we decided to complete the remaining sequencing run using NovaSeq platform. In total, we obtained 1 294 Gbp (Table 6-1).

Table 6-1: Illumina sequencing results. Sequencing results before and after preprocessing.

		Run 1	Run 2	Run 3	Run 4*	Run 5	Total
Raw reads	**Number of reads in millions**	760	13	671	914	6 266	8 624
	Number of Gb	114	2	101	137	940	1 294
Pre-processed reads	**Number of reads in millions**	613	10	585	767	4 869	6 844
	Number of Gb	87	1	84	109	684	965
Deduplicated reads	**Number of reads in millions**	859			509		5 251
	Number of Gb	123			73		745

*Sequenced using HiSeq platform.

For each run, the same preprocessing protocol was used on raw reads. I first used Trimmomatic (Bolger et al., 2014) version 0.27 with options CROP:125 HEADCROP:3 ILLUMINACLIP:adapters LEADING:3 TRAILING:3. This step was followed by the use of AfterQC (S. Chen et al., 2017) version 0.9.6 to removed polyX, especially for polyG removal, longer than 30 nucleotides using options -f 0 -t 0 -u 0 -n20 -s 80 –barcode=False –no_correction –no_overlap -p 30 -a 2 -u 0 –qc_sample=1000.

After merging all remaining reads, I used dedup.ssh from BBMap (https://sourceforge.net/projects/bbmap/) version 38.87 suit to eliminate duplicated reads, using default parameters. To check for potential contamination, I used FastQScreen (Wingett and Andrews, 2018) version 0.14.0. In addition to the default databases (Adapters, *Arabidopsis*, *Drosophila*, *E_coli*, Human, Lambda, Mitochondria, Mouse, PhiX, Rat, Vectors, Worm, Yeast, rRNA), I also added custom databases for proteobacteria (alpha, beta, delta, epsilon, gamma, zeta), *Aphanomyces astaci*, and Crustaceans (Supplementary 3). While a

significant number of sequences fell into the “no hit” category, no major contamination was detected. After preprocessing I obtained a total of 5.2 billion of reads corresponding to 745 Gbp (Table 6-1). The final coverage of short read sequencing is 43x.

6.2.2 Nanopore

Nanopore was a more accessible and cheaper technology than PacBio at the start of the project. All sequencing runs were performed with muscle tissue from two male noble crayfish individuals (Supplementary 4), the first individual being the same as for Illumina sequencing. Unfortunately, a recurrent issue occurred using Nanopore technologies as previously observed in molluscs sequencing (Adema, 2021): the nanopores rapidly became obstructed after the sequencing started, resulting in 10 different bad quality runs (output < 2 Gb, method expectation is > 50 Gb) with different protocols used (Table 6-2). Different kits for extraction and purification methods were tested (Lena Bonassin, unpublished results), without positive conclusions. Runs 1 and 2 were performed with the original extracted DNA. The low sequencing yield, which turned out to be due to the blocking of the nanopores, led to several optimization tests: DNA fragmentation to 8 and 20 kb with Covaris G-tubes and DNA clean-up with CTAB. Run 3 was performed after fragmentation of the HMW DNA to 8 kb. The nanopore activity improved, although with a lower N50 (as expected), and the yield was still way below the 50 Gb expected. Run 4 was performed after DNA cleanup using CTAB buffer, without fragmentation, but produced similar results as the original HMW DNA (Run 2). Runs 5 and 6 were performed without further DNA treatments (like runs 1 and 2) and gave similarly poor results. Runs 7 to 10 were performed after fragmentation of the HMW DNA to 20 kb. That gave similar outcome as with the 8 kb fragmentation for run 3, including the N50. The main difference was the broader distribution of read lengths with the 20 kb fragmentation. These results forced us to conclude that ONT was not a suitable platform for sequencing the noble crayfish.

Table 6-2: Nanopore sequencing results.

	Run 1	Run 2	Run 3*	Run 4**	Run 5	Run 6	Run 7*	Run 8*	Run 9*	Run 10*
Number of reads (thousand)	142.7	99.4	363.7	96	81.6	82.2	289.9	255.1	299.2	236.3
Number of bases (Mb)	1 090	737	1 690	677	633	655	1 800	1 460	1 700	1 470

*Fragmented DNA **CTAB-clean-up

6.2.3 Pacific Bioscience

With advancement of technologies, PacBio has emerged as a more promising option than Nanopore for sequencing large genomes like the noble crayfish. All PacBio sequencing was done using one male crayfish individual, obtained from the breeder Flusskrebszucht Frömel (Kavelstorf, Germany) (Supplementary 4). Firstly, we used PacBio sequelII to sequence CLR reads to obtain long reads (> 10 000 bp) in order to capture long stretches of repetitive elements (Table 6-3, Run 1). CLR provided 1.7 Gb on the 25 Gb expected. We then tested HiFi sequencing, which was gaining in popularity and could improve the quality of sequencing. HiFi sequencing results remained disappointing with a yield of less than 2 Gb on the 25 Gb expected (Run 2 and Run 3). In the vein of the challenges faced with nanopore sequencing, the native DNA of the noble crayfish appeared to easily fragment, impeding long read sequencing. The secondary structures of the DNA and/or chemical contaminants such as polysaccharides seemed to negatively affect long read sequencing. Given the difficulties with the native DNA, we shifted our approach and opted for the PacBio ultra-low input protocol, typically used for small organisms where only minimal amounts of DNA can be extracted. In this method, native DNA undergoes a PCR amplification step to generate sufficient material for sequencing (Run 4 through Run 27). This protocol allowed us to sequence non-native DNA, producing around 25 Gb to 30 Gb of data each, as expected by the method. In total, we generated 641 Gb of long read sequencing data, equivalent to approximately 38x genome coverage. However, because the PCR step duplicates the DNA, we needed to remove these duplicates. This was accomplished using the pbmarkdup version 1.0.2 tool (https://github.com/PacificBiosciences/pbbioconda) with default parameters, resulting in a final dataset of 550 Gb, corresponding to 32x genome coverage.

Table 6-3: PacBio sequencing results.

	Number of reads in millions	Number of Gb		Number of reads in millions	Number of Gb
Run1*+	0.2	1.7	**Run15**	2.4	18.6
Run2*	0.02	0.1	**Run16**	2.9	28.2
Run3*	0.2	1.7	**Run17**	2.9	27.6
Run4	2.2	19.9	**Run18**	2.9	24.1
Run5	3.2	30.8	**Run19**	3.1	25.1
Run6	2.0	22.0	**Run20**	3.2	29.9
Run7	2.4	24.8	**Run21**	3.0	28.2
Run8	3.0	28.6	**Run22**	3.2	26.0
Run9	2.8	27.6	**Run23**	3.4	27.3
Run10	2.6	26.3	**Run24**	3.2	26.4
Run11	2.9	29.7	**Run25**	3.0	27.2
Run12	2.7	25.7	**Run26**	2.9	26.2
Run13	3.0	30.3	**Run27**	3.0	27.8
Run14	2.9	28.8			

*Low input protocol, +CLR.

6.3 Short read assembly

6.3.1 Standard assembly

Preliminary test using Illumina runs 1 to 3

Based on literature research, the genome size was initially estimated to be approximately 2 Gb to 3 Gb (Gregory, 2023; Gutekunst et al., 2018; Tan et al., 2020). After the first three runs of Illumina sequencing, we achieved a coverage of 41x, assuming a genome size of around 3 Gb. Using KmerGenie version 1.7051 (Chikhi and Medvedev, 2014) for read-based genome size estimation, the genome was further assessed to 2.3 Gb.

To produce an assembly, we tested different tools. The tool SOAPdenovo2 (Luo et al., 2012) decomposed reads into kmers, and the graph is constructed by connecting overlapping kmers. This graph is then simplified by removing errors and resolving complex structures such as repeats. One advantage of SOAPdenovo2 is its ability to automatically select the optimal kmer

size for assembly, in contrast to ABySS (Jackman et al., 2017; Simpson et al., 2009), which requires manual specification of kmer size. Using the k63 pipeline, limiting the kmer size to 63, SOAPdenovo2 version 2.04-r241 and options -F -m 63 -N 2.3g, we generate statistics for all predicted assemblies for all tested kmers to evaluate the impact of kmer size on the assembly with the three first runs (Figure 6-1). We can see a bigger differentiation between kmer 49 and 50 than between all other kmer sizes, with the number of contigs and the length of the assembly increasing while the N50 decreased. Increasing kmer size increases the overlap between reads to be assembled and then eliminate more reads presenting repetitive elements and then produce a more accurate assembly (Cha and Bird, 2016).

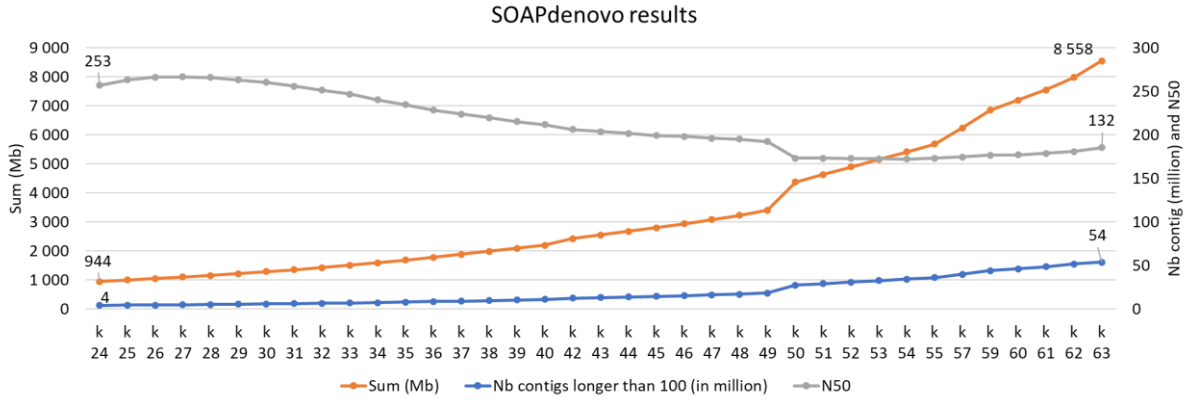

Figure 6-1: Statistics of the assemblies tested by SOAPdenovo2. Nb = number. The impact of the kmer size chosen for the assembly of the three first runs is showed for the total size of the assembly (Sum in Mb), the number of contigs, plotted in million on the right axis, that longer than 100 bp, and the N50 size on the same axis as the number of contigs.

SOAPdenovo determined the best assembly based on the statistics to be the one computed with a kmer of 63 and produced the assembly using this kmer. Contigs of less than 500 pb were removed from the assembly. The resulting assembly is composed of 1 315 456 contigs (minimum size: 500 bp), with a total assembly size of 1.3 Gb. The assembly N50 was only 956 bp, and the longest contig was 18 387 bp.

ABySS uses de Bruijn graph, similarly to SOAPdenovo2. While ABySS is generally slower compared to other tools, it is known for producing high-quality assemblies. However, unlike SOAPdenovo, ABySS requires manual specification of kmer length. To determine the optimal kmer size for our assembly, I tested various kmer lengths ranging from 20 to 75 using version 4.1 (Table 6-4). The best assembly, based on contig length and assembly size, was achieved

with a kmer size of 75 bp, yielding a result comparable to the SOAPdenovo assembly, with a total size of 1.3 Gb and an N50 of 856, remaining lower than the estimated size.

Table 6-4: ABySS assembly statistics for different kmer sizes. Only contigs over 500 bp were used. Assemblies were made using the three first runs.

	Number of contigs	**N50**	**Longest contig**	**Assembly size (Mb)**
K=20	165	557	867	0.09
K=25	203 014	722	5 811	151
K=30	282 068	752	5 929	216
K=35	386 741	759	5 941	298
K=37	529 978	883	14 794	464
K=39	569 958	889	14 794	502
K=41	611 429	896	13 930	541
K=43	666 061	904	13 351	593
K=45	712 157	907	13 351	636
K=50	841 769	915	13 770	756
K=55	978 980	915	14 111	877
K=65	1 267 309	897	13 874	1 110
K=75	1 605 558	856	9 176	1 365

For all short read assemblies, the N50 remains under 1 000 bp, indicating that the assemblies were highly fragmented. The N50 improved very slightly between kmer size of 20 and 55, and then decreased as the kmer size increased. The longest contigs were produced using a kmer size of 37 and 39. However, longest assemblies were produced using a kmer of 65 and 75. Using a kmer size of 75, the total assembly size of 1.3 Gb, although lower, was relatively close to our initial genome size estimates. To further assess the quality of the assemblies, I used the BUSCO (Manni et al., 2021; Seppey et al., 2019) score with the arthropoda_odb10 database (Figure 6-2), which includes a set of 1 013 genes expected to be conserved as single copy across all Arthropoda species. BUSCO evaluates the proportion of these conserved genes present in the assemblies. The SOAPdenovo assembly had the highest BUSCO score, retrieving 11.5% of complete genes and 15.1% of fragmented genes. To further investigate this low

BUSCO score, I mapped the raw reads back to the ABySS assembly (kmer size 75) using Bowtie2. The alignment rate was 78%, with an estimated coverage of 8x based on the data from the initial three sequencing runs, explaining also the high fragmentation of the genome assemblies. Considering the high alignment rate and the low coverage, we started doubting the initial expectations of a genome size of 2-3 Gb. To refine the genome size estimate, we conducted flow cytometry, which suggested a genome size of 17 Gb. (Theissinger et al., unpublished results). This confirmed that the first three runs provided an 8x coverage. Recognizing the need for additional sequencing, we performed two more sequencing runs: the fourth on the HiSeq platform and the fifth on the NovaSeq platform (Table 6-1).

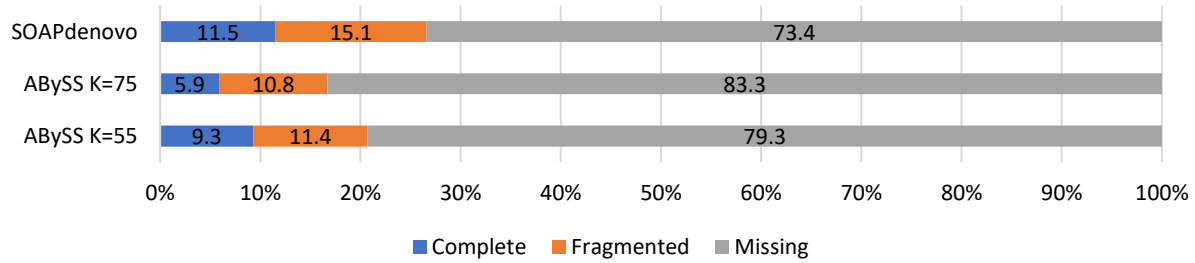

Figure 6-2: Busco scores for assembly trials for contigs over 500 bp. The BUSCO scores are provided for each assembly compared to the arthropoda_odb10 database.

Assemblies using all short read runs (1 to 5)

Once all the sequencing reads were available, and based on our previous assembly trials, I used SOAPdenovo to produce an assembly. However, SOAPdenovo2 didn't support the data amount and failed without producing any assembly. I then used ABySS tool (version 4.3), increasing the kmer length even further as it seemed to produce better assemblies. I tested kmer lengths of 97 and 127 (Table 6-5). The best assembly was produced by ABySS using a kmer length of 97 with an assembly size of 3 Gb and a N50 of 1 892 bp. To assess the quality, I used compleasm (Huang and Li, 2023), a BUSCO-like tool that uses minimap and is much faster than BUSCO, also using arthropoda_odb10. The compleasm score for complete genes was 16.58. I then tried the rresolver module from ABySS tool with default parameters on the ABySS assembly (K=97) to attempt to resolve some repeat-rich regions that increased the assembly size to 3.3 Gb, for a small reduction of the N50 and compleasm score.

At the same time, I also tested the Megahit assembler (Li et al., 2015), which is more commonly used for metagenomic approaches, because of its ability to manage large amount of data for a small amount of memory usage. It uses advanced graph simplification techniques to efficiently handle the complexity of metagenomic data. Megahit uses an iterative kmer strategy, in contrast to SOAPdenovo which tests different kmers to produce the best assembly. Megahit starts an assembly at a small kmer size and then increases its size to improve the previous assembly, which helps to resolve complex repeats and ambiguities. Megahit version 1.2.9 with options –k-min 41 –k-steps 10 produced a better assembly than ABySS with a length of 5 Gb, and an N50 of 2 kb (Figure 6-3). As Megahit does not have a scaffolding step using paired-end information, it recommended that we use SOAPdenovo-fusion -D -K 41 to make this step. This scaffolding step increased the assembly length to 6.8 Gb with a similar N50 of 2 kb. The Megahit assembly with scaffolding using SOAPdenovo fusion showed a completeness of 18.16% with 53.7 % of fragmented genes.

Table 6-5: Assembly statistics for contigs over 500 bp using all Illumina runs.

	Number of contigs (millions)	**N50**	**Longest contig**	**Assembly size (Mb)**
Megahit	3.4	2 076	326 357	4 999
Megahit SOAPdenovo fusion	4.6	2 034	529 289	6 862
ABySS K=97	2.2	1 892	194 926	3 005
ABySS K=97 rresolver	2.5	1 837	274 140	3 360
ABySS K= 127	2.6	807	28 578	2 254

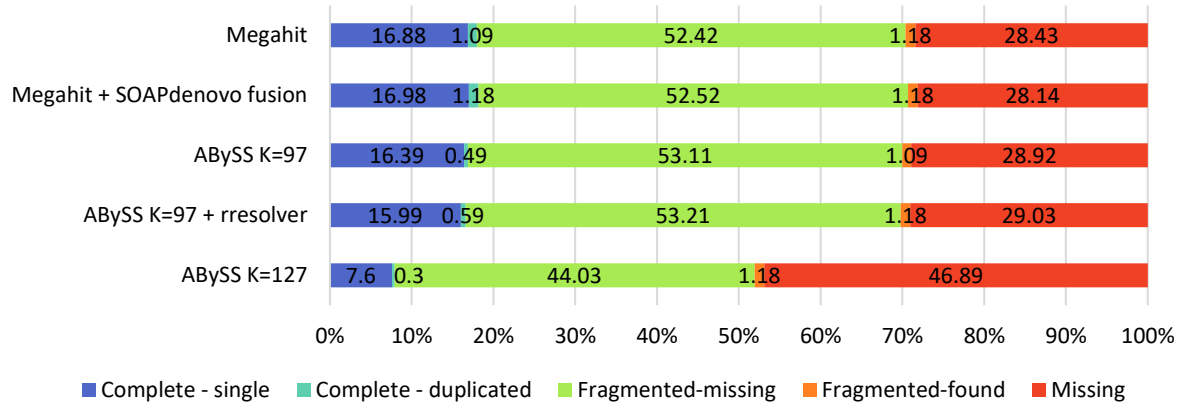

Figure 6-3: Compleasm scores for assemblies with contigs over 500 bp using all runs. The compleasm scores are provided for each assembly compared to the arthropoda_odb10 database. "Complete – single" corresponds to genes found complete in the genome in a single copy. "Complete – duplicated" corresponds to genes found complete in the genome in more than one copy. "Fragmented – missing" corresponds to genes found incomplete in the genome. "Fragmented – found" corresponds to genes found incomplete but span into different contigs. "Missing" corresponds to missing genes from the assembly.

6.3.2 Stepwise assembly

Even if all assemblies were conducted on a 192 cores server with 2.95 Tb of memory, I faced memory issues with standard assembly. SOAPdenovo has the great advantage of automatically choosing the best kmer size and I wanted to use this ability to produce an assembly. Considering the amount of data and the memory issues assembling the genome using SOAPdenovo, I tested a stepwise assembly. The stepwise assembly strategy was adapted from the one used for the Siberian larch (12 Gb) (Kuzmin et al., 2019). The strategy consists of splitting all the reads (forward and reverse) into five different groups of equal size, representing a coverage of 8X each (Figure 6-4). A sixth group of the same size as the first five was created using reads randomly sampled from all the available reads. Those six groups were assembled using SOAPdenovo with 127mer module using options -F -m 127 -N 17g (Table 6-6). The whole set of reads were used to create a seventh group of data, however, on this group all the information about pairs was removed to lower the use of computational memory. Reads in the seventh group were then considered as simple reads in this group. However, even by removing information about pairs, the SOAPdenovo program (with option to assemble simple end reads) used to assemble every other group couldn't manage the whole set of reads. We then used the program Megahit to assemble this group of reads, as Megahit considers all

reads as simple reads anyway. Results showed that each group assembled using SOAPdenovo had similar statistics with an assembly size of 3.3 Gb and a N50 of 1k (Table 5). The complete pool assembled using Megahit without the information about pairs using options –k-min 25 –k-max 127 –k-steps 2. It produced a shorter assembly of 1.7 Gb and a N50 slightly longer at 1.2 kb. The longest contig produced by Megahit was also much longer, with 200 kb, compared to the other pool of around 30 kb.

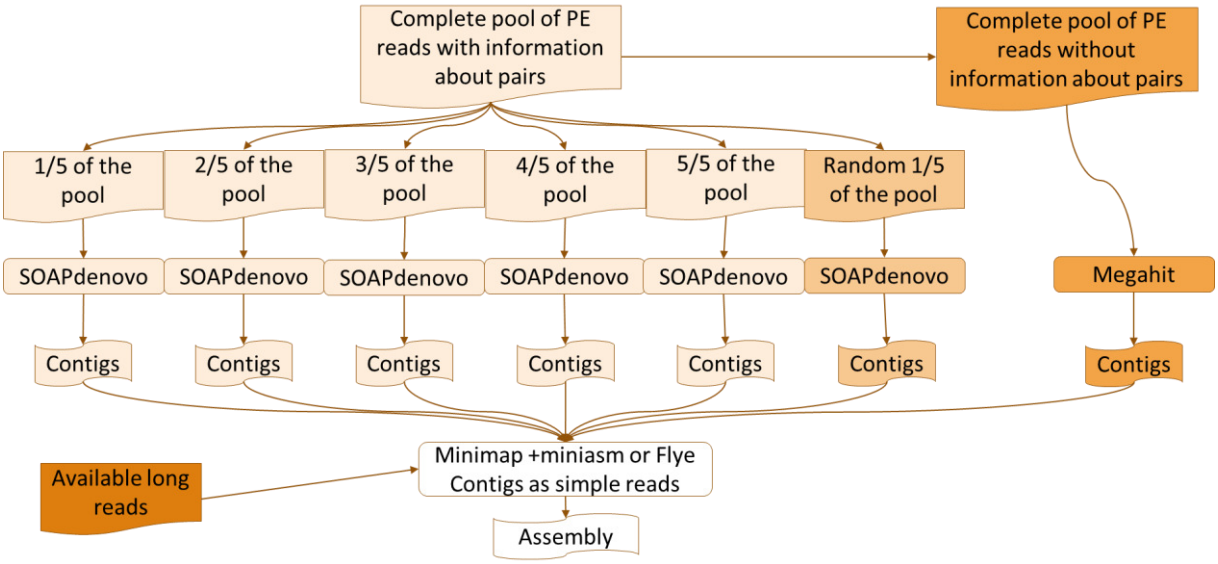

Figure 6-4: Stepwise assembly strategy. The complete pool of read is separated in five pools to be assembled separately. A sixth group is formed by an equal number of reads as the five first but randomly took in the complete pool. A seventh group consists of all reads without information about pairs. All contig generated and available long reads are then assembled to produce the assembly.

Table 6-6: Stepwise assembly statistics for each pool. Statistics are showed for all contigs generated using short reads.

	Part 1	Part 2	Part 3	Part 4	Part 5	Random	Single end
Sum (Gb)	3.375	3.374	3.373	3.375	3.375	3.373	1.788
N ctg	3 461 228	3 459 798	3 459 122	3 461 260	3 461 246	3 458 658	1 609 165
Max length	27 114	34 267	25 494	37 013	25 372	30 740	200 632
N50	1 025	1 025	1 025	1 025	1 025	1 025	1 229

To assemble the contigs generated in all of these seven groups we tried the tools Quickmerge, Flye, the subassembly method of Flye (Kolmogorov et al., 2019), and minimap and miniasm

(Li, 2016). All these assembly methods were attempted without any real success, with no statistical improvements, or assembly failed. The Flye version 2.8.1-b1676 pacbio-hifi module, with option -g 17g was used to produce an assembly. In addition to all contigs coming from the different pools I added the seven first PacBio runs available (101 Gb, 6X). This assembly provided a draft assembly, before the consensus step using Minimap, of 12.5 Gb with an N50 of 52 kb before failing (Table 6-7). The compleasm score was 22.9%. Compared to more traditional assemblies, compleasm complete score of the stepwise was better, probably because of the use of the long reads, that are not used in the produced traditional assemblies, providing more material for the assembly. However, the percentage of missing genes is higher in the Stepwise assembly, 40.77% against 28.14% in the Megahit + SOAPdenovo2 fusion approach. This is probably due to the low coverage used in the stepwise assembly and the fact that Flye require a minimum of 1 000 bp aligned to merge reads/contigs.

Once all the long reads available, I tried to assemble the short read contigs again with the use of all long reads. The Flye subassembly module failed again, so I tried wtdbg2 (Ruan and Li, 2020). Unlike Flye which uses an OLC approach, wtdbg2 uses a de Bruijn graph approach based on w-kmers. W-kmers are subsequences from the reads that do not require an exact match. Wtdbg is a fast tool that however have some troubles handling repeats due to the use of de Bruijn graph. Wtdbg version 0.0 was used with options -x preset4 -l 500 -g 17g -m 1000 and generated an assembly of 8.6 Gb with a N50 of 29k (Table 6-7). The compleasm complete score is 32.28%. The wtdbg2 assembly produced a better assembly than the Flye subassembly in terms of completeness, probably due to the use of the complete set of long reads. The contiguity statistics are however much lower than for the Flye subassembly.

Table 6-7: Stepwise assembly statistics after merging assembly of each pool and long reads.Statistics are shown for all contigs generated using short reads contigs and the seven first long reads runs for the Flye subassembly. These results are before the consensus step using Minimap before Flye failed. For wtdbg2 the complete set of long reads were used.

	Flye – pacbio_hifi	**wtdbg2**
Sum (Gb)	12.5	8.6
N ctg	263 159	425 839
Max length	1 146 521	642 783
N50	52 076	29 190

Standard genome assemblies have the advantage of minimizing bias, as the only major biases introduced stem from the assemblers themselves. In contrast, stepwise assembly approaches can potentially introduce misassemblies, particularly when coverage is low (an average of 8x for each pool in this case). This reduced coverage can lead to errors in connecting contigs, resulting in a less accurate assembly and single end connected reads are subject to errors (Jünemann et al., 2014). Assemblies based on short reads are limited by the high number of REs present in the genome. As a result, these assemblies typically have low compleasm scores and contiguity metrics, reflecting an incomplete and fragmented assembly (P. Wang et al., 2021). The stepwise assembly was conducted in order to reduce computational resources. However, it took longer than the traditional assembly, considering that seven different assemblies were produced. Moreover, while connecting the contigs produced by the assemblies, I faced computational issues again and various tools returned an error due to memory consumption.

6.3.3 Scaffolding

Several attempts to scaffold the different short read assemblies were made. The first attempt was based on transcriptomic data, while subsequent attempts were based on long-reads.

Scaffolding standard short read assembly with transcriptomic data

I tried to exploit transcriptomes to scaffold the Megahit and SOAPdenovo fusion assembly. Transcriptome scaffolding takes advantage of the fact that expressed genes are generally localized in specific regions of the genome, and the transcriptome provides information about

the order and orientation of these gene regions. To do that, we used the two available *A. astacus* transcriptomes, GeneBank ID: GJEB00000000.1 (Boštjančić et al., 2022) and GEDF00000000.1 (Theissinger et al., 2016). The tools L_RNA_scaffolder (Xue et al., 2013) and SCUBAT (https://github.com/elswob/SCUBAT/) were used. For L_RNA_scaffolder, transcripts are first aligned to the assembled genome contigs using BLAT. L_RNA_scaffolder has the advantage of having the ability to use contigs instead of reads. If a transcript spans multiple contigs, this suggests that these contigs should be scaffolded together. It uses the splicing information from the RNA-Seq reads to determine the relative position and orientation of contigs. Scubat works the same way, without the use of splicing information, however the alignment of contigs is not included in the program. However, using both tools, no scaffolds were produced.

Scaffolding standard short read assembly with long reads

The scaffolding of short read assemblies using long reads really improves the contiguity of genome assembly (Coombe et al., 2021). Using the PacBio reads, I made a scaffolding on the Megahit and SOAPdenovo fusion assembly using LongStitch (Coombe et al., 2021) version 1.0.3 with the tigmint-ntLink-arks module (options: -G 17g -k_ntLink 63 -k_arks 63 -z 500 and gap_fill True). The overall assembly size improved from 6.9 Gb to 13.8 Gb, however the N50 decreased from 2 kb to 723 bp and the longest contig from 529 289 bp to 326 357 bp. The decrease of the N50 could be due to some initial misassemblies that are now corrected by long reads alignments. The compleasm complete score increased from 16.98% to 25.67% with a decrease of missing genes going from 28.14% to 25.56% (Figure 6-5).

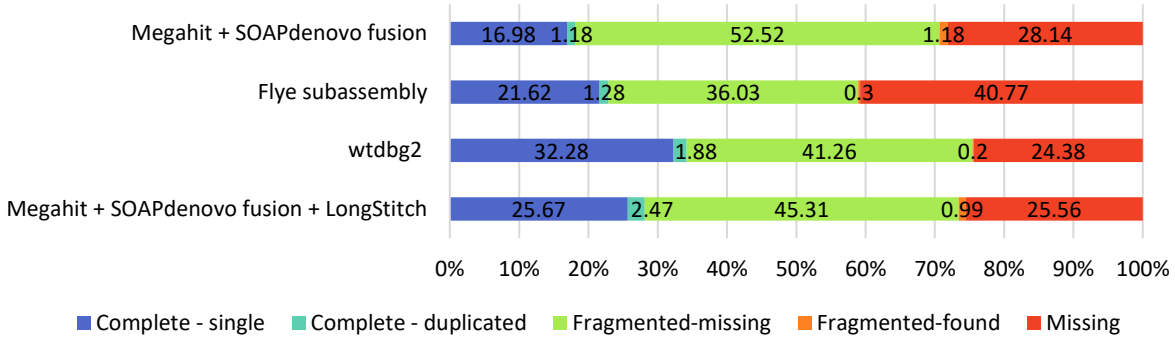

Figure 6-5: Compleasm scores for short reads assemblies before and after scaffolding.The compleasm scores are provided for the initial Megahit and SOAPdenovo-fusion assembly (first row) and for the same assembly scaffolded by LongStitch using long reads (last row). The Flye subassembly was obtained after merging sub-assemblies of 7 short-read pools and 6X long

reads. The stepwise assembly with wtdbg was produced using short read sub-assemblies and all long reads generated. The compleasm scores are provided compared to the arthropoda_odb10 database. “Complete – single” corresponds to genes found complete in the genome in a single copy. “Complete – duplicated” corresponds to genes found complete in the genome in more than one copy. “Fragmented – missing” corresponds to genes found incomplete in the genome. “Fragmented – found” corresponds to genes found incomplete but span into different contigs. “Missing” corresponds to missing genes from the assembly.

6.4 Long read assembly

Different tools were tested at low coverage as sequencing data were coming in. Finally, the software hifiasm (Cheng et al., 2021) was chosen for the long read assembly. Hifiasm is an assembler designed specifically for high-quality, long read sequencing data, particularly from PacBio HiFi reads. Hifiasm takes advantage of the HiFi reads quality to directly assemble reads without requiring heavy error correction. This saves computational time and avoids the need for a consensus-building step, fastening the assembly. Moreover, hifiasm seems to be one of the most efficient tools to assemble HiFi PacBio sequences and use a k-mer aproach (Cosma et al., 2023). Using the 32x long reads, we finally produced an assembly with hifiasm version 0.19.5-r592 with options -k 63 –hg-size 17g and -l0 to not purge haplotigs. The assembly is composed of 300 478 contigs for a total length of 21.8 Gb and a N50 contig of 127 926 bp, the longest contig being 6 176 326 bp. The compleasm complete score is of 42.54% and 36.72% of fragmented (Figure 6-6). This assembly shows a tiny amount of short-sized contigs, with only 7 contigs shorter than 5 kb and a mean read size of 10 kb (Table 6-8). The hifiasm assembly considering only contigs over 50 kb is 16.7 Gb and is so really close to the estimate genome size. However, the compleasm score for the contigs over 50 kb is much lower than without filtering (Figure 6-6).

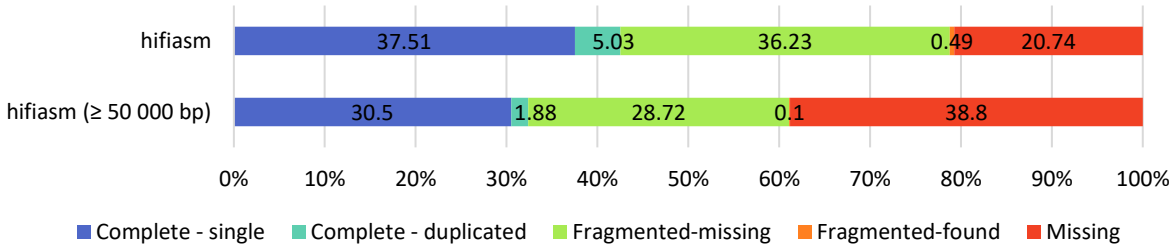

Figure 6-6: Assembly compleasm score for long reads assembly. The compleasm scores are provided compared to the arthropoda_odb10 database for the hifiasm assembly and for the hifiasm assembly after filtering contigs below 50 kb. “Complete – single” corresponds to genes

found complete in the genome in a single copy. “Complete – duplicated” corresponds to genes found complete in the genome in more than one copy. “Fragmented – missing” corresponds to genes found incomplete in the genome. “Fragmented – found” corresponds to genes found incomplete but span into different contigs. “Missing” corresponds to missing genes from the assembly.

Table 6-8 : Contigs statistics for the hifiasm assembly for different contig size.

≥ 0 bp	Number of contigs	300 478	Total length (Gb)	21.8
≥ 1 000 bp		300 478		21.8
≥ 5 000 bp		300 471		21.8
≥ 10 000 bp		297 578		21.8
≥ 25 000 bp		204 875		20.1
≥ 50 000 bp		107 641		16.7

In contrast to short reads, long read assemblies, as expected, produce a more contiguous genome due to the superior ability of long read technologies to span repetitive regions (Cechova, 2020). This leads to better handling of REs, however, a larger than expected genome size was observed. The inflated genome size of 21 Gb instead of 17 Gb could be attributed to unresolved REs being duplicated within the assembly resulting in non-connected contigs Additionally, in the Hifiasm assembly, the haplotigs were not purged considering the low completeness score, contributing further to the overall genome size (Roach et al., 2018). Haplotigs purging was tried but consequently reduced completeness of the assembly. Compared to the Megahit assembly, which was scaffolded using long reads, and stepwise assembly produced with wtdbg2, the long read assembly shows both a larger assembly size and a higher number of complete genes, making it the better strategy among all those tested, including standard assembly with ABySS, Megahit without scaffolding and stepwise assembly with Flye subassembly.

6.5 Comparison of the preliminary assembly to available Arthropoda genomes

A study comparing the completeness and contiguity of Arthropoda genomes available on NCBI provided a valuable basis for assessing the quality of our assembly in comparison to related genomes (Feron and Waterhouse, 2022). This study highlights a huge variability in the quality

of arthropod assemblies both in terms of contiguity (assessed by the N50) and in terms of completeness (assessed by the BUSCO score). This is particularly true for decapod (21 genomes analysed), with N50 values ranging from 1 kb up to 100 Mb, with the median at 100 kb (Figure 6-7). BUSCO scores ranged from 25% to nearly 100%, with a fourth of the assemblies being between 35% to 75% completeness. While a completeness score of 42.54% is not ideal, our genome is comparable in quality to other decapod species. Concerning contiguity, with a N50 of 128 kb, we are at the median of decapods assemblies. This result is remarkable considering the size of the other assemblies, which are all under 5 Gb, compared to the huge size of our preliminary assembly (21.8 Gb). In fact, our noble crayfish genome is the largest assembly obtained for all Arthropods (according to the NCBI Genome database, consulted in November 2024), surpassing the recently published genome of *Meganyctiphanes norvegica* (Northern krill) that is 19.7 Gb (Unneberg et al., 2024). Published in another database, the 48.1 Gb genome of *Euphausia superba* (Antarctic krill) is the biggest metazoan genome published to date (Shao et al., 2023). Apart from *A. asta*cus and krills, all Arthropod assemblies are smaller than 10 Gb.

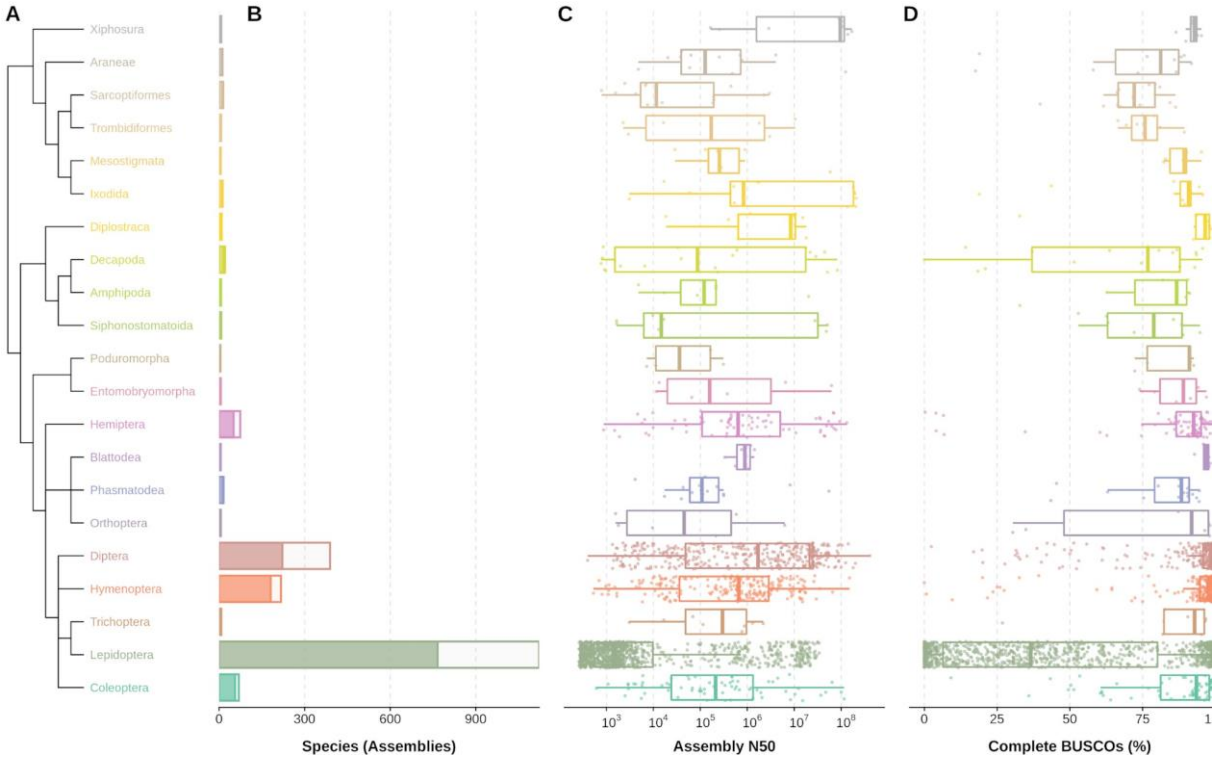

Figure 6-7: Assembly statistics for arthropod species. A: Phylogenetic tree. B: Number of assemblies for each group. C: Assembly N50 for each group. D: BUSCO completeness for each group. Figure from Feron and Waterhouse, 2022.

6.6 Conclusion and perspectives

This chapter outlined the difficulties in sequencing and assembling such a large genome. Despite numerous attempts, we were unable to obtain a satisfactory assembly from short reads. Sequencing with long reads was also challenging. Nanopore sequencing was hindered by pore blockages. PacificBioscience long read sequencing proved more promising, though challenges persisted with DNA fragmentation and low sequencing yield, which were solved using the ultra-low input protocol, which uses DNA amplified by PCR. The ability to sequence non-native DNA pinpoints the presence of some chemicals or tertiary structure of the DNA preventing its sequencing at a native state. Despite these obstacles, a draft assembly based on 32x long read coverage using PacBio HiFi reads yielded a final genome assembly of 21.8 Gb with 42.54% completeness. This assembly provides a solid foundation for future improvements, especially considering that it is still in its early stages, as it would require both assembly polishing and obtaining new data such as HiC to improve the different assemblies, which couldn't be achieved within the time constraints of this thesis. Considering the impact of kmer size on the assembly quality (Figure 6-1), we can regret that the maximum kmer size is limited to 63 in hifiasm. Increasing kmer size could improve the contiguity and the quality of huge genome assembly. One of the key next steps for the long read assembly would be to map the long reads to the genome assembly to fill gaps between contigs and to try to extend some contigs and correct potential misassemblies. Following this, the short reads would be mapped for the same purpose, as short reads might have more deeply sequenced some parts of the genome than long reads. Additionally, nanopore reads, though fewer in number, could be leveraged to connect reads that remain unresolved, potentially closing persistent gaps in the assembly. These steps could improve both contiguity, completeness and accuracy of the genome. Finally, integrating Hi-C data will allow to establish the higher-order structure, improving the accuracy of scaffolding by correctly positioning contigs and resolving misassemblies. Hi-C sequencing captures the three-dimensional organization of the genome by identifying interactions between different chromosomal regions. This method would provide a more comprehensive view of the genome architecture, ultimately increasing the accuracy of the final assembly.

Chapter 7 - Conclusion and perspectives

My PhD thesis represents the first comprehensive comparative analysis of decapod genomes, an area where genomic research remains relatively limited despite the economic importance and the evolutionary and ecological significance of this group. While substantial genomic work has focused on other arthropods, particularly insects (Misof et al., 2014; Petersen et al., 2019; Thomas et al., 2020), decapods have been underexplored, primarily due to the complexity and size of their genomes (Iannucci et al., 2022). Despite the limited number of species studied here, this work provides insights into decapod genomic features, particularly regarding REs and protein-coding genes, both of which play critical roles in genome structure and function (Biscotti et al., 2015).

The RE annotation pipeline that I have developed in my PhD thesis (Chapter 4) successfully increased the RE detection by 10% over traditional protocols, reducing the number of unclassified REs. This improved annotation of REs using a standardised approach has implications for understanding the functional and evolutionary significance of REs, as REs are known to drive genome expansion and variation (Nie et al., 2024; Treangen and Salzberg, 2012; Yuan et al., 2024). This is particularly true for decapods, which show extreme variations in genome size. Decapod genomes studied here showed higher RE content levels than those of many other taxa (58% and 79%; Rutz et al., 2023). The genome size and RE content correlation observed here follow a general trend as seen in other complex genomes (Tenaillon et al., 2011), although exceptions exist within certain taxa. Despite the fragmentation of some analysed assemblies, I was able to detect a clear phylogenetic signal from the RE content. The RE diversity observed in decapods also supports evidence from other taxa that lineage-specific expansions of certain RE families can contribute to evolutionary divergence and adaptation (Chalopin et al., 2015; Sotero-Caio et al., 2017). The resource-intensive annotation process highlighted the need for potential optimisations, while the high RE content underscored the impact of REs on genome assembly, pointing to the need for refined bioinformatical methods for accurate genomic analysis.

The preliminary comparative analysis of Decapoda proteomes identified both highly conserved protein-coding genes and lineage-specific adaptations (Chapter 5), echoing findings

in other arthropod comparative genomics (Misof et al., 2014; Thomas et al., 2020). My analysis allowed to delineate Decapoda and infra-orders core proteomes, which are subject to shared evolutionary pressures that maintain key functions (Thomas et al., 2020). Within the Decapoda core proteome, Decapoda-specific proteins are prime candidates for future studies aimed at understanding the characteristics of decapods. Large-scale phylogenetic profiling of the *P. clarkii* proteome revealed particularly promising clusters with atypical distribution and emphasised the diverse selective forces that have shaped decapod evolution.

My comparative analysis of REs and proteomes have prepared for my analysis of the genome assembly of the noble crayfish. Obtaining this giant genome was challenging, both from sequencing and bioinformatics aspects (Chapter 6). The genome assembly of the noble crayfish posed challenges typical of large, RE-rich genomes. Among all tested strategies, short-read approaches failed to produce a satisfactory assembly due to low statistics and completeness despite a decent sequencing depth. Long-read sequencing through Nanopore and PacBio technologies had different issues such as extensive pore blockage and DNA fragmentation. The PacBio ultra-low input protocol, which uses PCR-amplified DNA, eventually enabled sequencing, leading to a draft assembly of 21.8 Gb with 42.54% completeness. This first *A. astacus* genome assembly demonstrates substantial progress in quality compared to previous efforts, providing a foundational framework for future decapod genomics. Future improvements should include assembly polishing, adding Hi-C data for accurate contig positioning, and leveraging long and short reads to close gaps and correct errors (Lieberman-Aiden et al., 2009). Together, these steps could improve genome contiguity, completeness, and accuracy. While the noble crayfish genome remains incomplete, it represents a crucial step toward understanding the genetic landscape of decapods. The ultra-low input protocol employed for PacBio sequencing is promising for other large and hard to sequence genomes, facilitating the sequencing of large DNA fragments even with complex genomic structures. It is essential to note that currently sequenced decapod genomes come from species with relatively moderate genome sizes (< 5 Gb). The sequencing of larger and more complex genomes remains a significant challenge due to the abundance of REs and the substantial genome size, marking the assembly process highly complicated but also demanding sophisticated and diverse bioinformatic approaches (Treangen and Salzberg, 2012).

Further research on the noble crayfish genome should focus on comprehensive annotation, including gene prediction, regulatory element mapping, and non-coding sequence characterization. These steps are essential for detailed studies of gene pathways, evolutionary dynamics, and functional adaptations in decapods (Yandell and Ence, 2012). In addition, investigating structural aspects, such as intron size variation, could reveal genomic differences compared to related groups like other crustaceans and insects (Misof et al., 2014). Improved annotations will facilitate comparative studies and complement the preliminary comparative analysis done during this thesis, shedding light on the genetic elements that contribute to the unique biology and adaptability of decapods.

The RE study in decapods paves the way for the study of these elements in the giant genome of the noble crayfish, potentially uncovering how they shape genomic size and architecture and evolutionary adaptability. For instance, a preliminary analysis of the assembly using RepeatExplorer (Novák et al., 2017) identified 370 satDNA families in the genome of the noble crayfish and provides a valuable starting point for the RE analysis still ongoing. Given the significant presence of REs across decapod genomes, further research should focus on characterizing the diversity and functional impact of these elements. By exploring RE families in detail and comparing them to the REs of decapod species already studied across this thesis and other species, we may uncover insights into how these elements shape the genomic and functional characteristics of decapods.

My work sets the stage for sequencing additional crayfish genomes, including European species. Comparative analysis between European crayfish and related North American and Australian species presents a unique opportunity to investigate genetic factors influencing susceptibility or tolerance to the crayfish plague disease agent. Identifying genetic differences or conserved regions of tolerant hosts could offer insights into the mechanisms behind the disease, which could benefit conservation strategies and population management efforts against crayfish plague (J. Jussila et al., 2021; Jussila et al., 2017; Svoboda et al., 2017). Comparisons of immune-related gene families and regulatory elements could provide insights into decapod adaptation. Functionally, comparing gene families involved in biological processes like immunity and metabolism can yield insights into adaptive strategies in decapods. At a larger scale, examining gene families involved in immunity and metabolism

across crayfish and decapods could help identify adaptive strategies specific to decapods, which may have applications in aquaculture and conservation biology. For instance, high-throughput DNA chip technology could leverage decapod genomic data for precise species identification, population monitoring, and disease detection, allowing for more precise monitoring of population health and stability. Additionally, identifying genetic markers for disease resistance could support the selection of resilient individuals suited to aquaculture, promoting sustainable farming practices.

In summary, my thesis significantly advances the field of decapod genomics by providing the first comparative analysis of decapod genomes, with a particular focus on REs and protein-coding genes. Through developing a standardised RE annotation pipeline and conducting comprehensive proteome analyses, I have illuminated some of the genomic characteristics that shape decapod evolution and functional adaptation. The assembly of the noble crayfish genome, despite challenges, establishes a foundational resource for future genomic research in decapods. This work not only enriches our understanding of decapod genome architecture but also offers practical insights for conservation, aquaculture, and the study of disease resistance, paving the way for future investigations into adaptive traits and species resilience.

References

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

Adema, C.M., 2021. Sticky problems: extraction of nucleic acids from molluscs. Philos. Trans. R. Soc. B Biol. Sci. 376. https://doi.org/10.1098/rstb.2020.0162

Ahrenfeldt, J., Skaarup, C., Hasman, H., Pedersen, A.G., Aarestrup, F.M., Lund, O., 2017. Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods. BMC Genomics 18, 19. https://doi.org/10.1186/s12864-016-3407-6

Akhan, S., Bektas, Y., Berber, S., Kalayci, G., 2014. Population structure and genetic analysis of narrow-clawed crayfish (Astacus leptodactylus) populations in Turkey. Genetica 142, 381–395. https://doi.org/10.1007/s10709-014-9782-5

Alderman, D.J., 1996. Geographical spread of bacterial and fungal diseases of crustaceans. Rev. Sci. Tech. Int. Off. Epizoot. 15, 603–632. https://doi.org/10.20506/rst.15.2.943

Alvarez, M., Schrey, A.W., Richards, C.L., 2015. Ten years of transcriptomics in wild populations: what have we learned about their ecology and evolution? Mol. Ecol. 24, 710–725. https://doi.org/10.1111/mec.13055

Andriantsoa, R., Tönges, S., Panteleit, J., Theissinger, K., Carneiro, V.C., Rasamy, J., Lyko, F., 2019. Ecological plasticity and commercial impact of invasive marbled crayfish populations in Madagascar. BMC Ecol. 19, 8. https://doi.org/10.1186/s12898-019-0224-1

Anger, K., 2016. Adaptation to Life in Fresh Water by Decapod Crustaceans: Evolutionary Challenges in the Early Life-History Stages, in: Kawai, T., Cumberlidge, N. (Eds.), A Global Overview of the Conservation of Freshwater Decapod Crustaceans. Springer International Publishing, Cham, pp. 127–168. https://doi.org/10.1007/978-3-319-42527-6_5

Angthong, P., Uengwetwanit, T., Pootakham, W., Sittikankaew, K., Sonthirod, C., Sangsrakru, D., Yoocha, T., Nookaew, I., Wongsurawat, T., Jenjaroenpun, P., Rungrassamee, W., Karoonuthaisiri, N., 2020. Optimization of high molecular weight DNA extraction methods in shrimp for a long-read sequencing platform. PeerJ 8, e10340. https://doi.org/10.7717/peerj.10340

Ansorge, W.J., 2009. Next-generation DNA sequencing techniques. New Biotechnol. 25, 195–203. https://doi.org/10.1016/j.nbt.2008.12.009

Arakawa, K., 2023. Ultralow-Input Genome Library Preparation for Nanopore Sequencing with Droplet MDA, in: Arakawa, K. (Ed.), Nanopore Sequencing: Methods and Protocols, Methods in Molecular Biology. Springer US, New York, NY, pp. 91–100. https://doi.org/10.1007/978-1-0716-2996-3_7

Ardui, S., Ameur, A., Vermeesch, J.R., Hestand, M.S., 2018. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res. 46, 2159–2168. https://doi.org/10.1093/nar/gky066

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene

Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29. https://doi.org/10.1038/75556

Athanasio, C.G., Chipman, J.K., Viant, M.R., Mirbahai, L., 2016. Optimisation of DNA extraction from the crustacean Daphnia. PeerJ 4, e2004. https://doi.org/10.7717/peerj.2004

Audo, D., Kawai, T., Letenneur, C., Huang, D., 2023. Crayfishes from the Jehol biota. Geodiversitas 45. https://doi.org/10.5252/geodiversitas2023v45a24

Austin, C.M., Croft, L.J., Grandjean, F., Gan, H.M., 2022. The NGS Magic Pudding: A Nanopore-Led Long-Read Genome Assembly for the Commercial Australian Freshwater Crayfish, Cherax destructor. Front. Genet. 12. https://doi.org/10.3389/fgene.2021.695763

Averof, M., Patel, N.H., 1997. Crustacean appendage evolution associated with changes in Hox gene expression. Nature 388, 682–686. https://doi.org/10.1038/41786

Bachvaroff, T.R., McDonald, R.C., Plough, L.V., Chung, J.S., 2021. Chromosome-level genome assembly of the blue crab, Callinectes sapidus. G3 GenesGenomesGenetics 11, jkab212. https://doi.org/10.1093/g3journal/jkab212

Balasubramanian, S., 2015. Solexa Sequencing: Decoding Genomes on a Population Scale. Clin. Chem. 61, 21–24. https://doi.org/10.1373/clinchem.2014.221747

Bang-Jensen, J., Gutin, G., Yeo, A., 2004. When the greedy algorithm fails. Discrete Optim. 1, 121–127. https://doi.org/10.1016/j.disopt.2004.03.007

Bao, W., Kojima, K.K., Kohany, O., 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6, 11. https://doi.org/10.1186/s13100-015-0041-9

Barrón, M.G., Fiston-Lavier, A.-S., Petrov, D.A., González, J., 2014. Population Genomics of Transposable Elements in Drosophila. Annu. Rev. Genet. 48, 561–581. https://doi.org/10.1146/annurev-genet-120213-092359

Bennetzen, J.L., Wang, H., 2014. The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. Annu. Rev. Plant Biol. 65, 505–530. https://doi.org/10.1146/annurev-arplant-050213-035811

Bertocchi, S., Brusconi, S., Gherardi, F., Grandjean, F.S.-G., 2008. Genetic variability of the threatened crayfish Austropotamobius italicus in Tuscany (Italy): implications for its management. Fundam. Appl. Limnol. 153–164. https://doi.org/10.1127/1863-9135/2008/0173-0153

Biscotti, M.A., Canapa, A., Forconi, M., Olmo, E., Barucca, M., 2015. Transcription of tandemly repetitive DNA: functional roles. Chromosome Res. 23, 463–477. https://doi.org/10.1007/s10577-015-9494-4

Bláha, M., Weiperth, A., Patoka, J., Szajbert, B., Balogh, E.R., Staszny, Á., Ferincz, Á., Lente, V., Maciaszek, R., Kouba, A., 2022. The pet trade as a source of non-native decapods: the case of crayfish and shrimps in a thermal waterbody in Hungary. Environ. Monit. Assess. 194, 795. https://doi.org/10.1007/s10661-022-10361-9

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

Boštjančić, L.L., Bonassin, L., Anušić, L., Lovrenčić, L., Besendorfer, V., Maguire, I., Grandjean, F., Austin, C., Greve, C., Hamadou, A., Mlinarec, J., 2021. The Pontastacus leptodactylus (Astacidae) Repeatome Provides Insight Into Genome Evolution and Reveals Remarkable Diversity of Satellite DNA, Frontiers in Genetics. https://doi.org/10.3389/fgene.2020.611745

Boštjančić, L.L., Francesconi, C., Rutz, C., Hoffbeck, L., Poidevin, L., Kress, A., Jussila, J., Makkonen, J., Feldmeyer, B., Bálint, M., Schwenk, K., Lecompte, O., Theissinger, K.,

2022. Dataset of the de novo assembly and annotation of the marbled crayfish and the noble crayfish hepatopancreas transcriptomes. BMC Res. Notes 15, 281. https://doi.org/10.1186/s13104-022-06137-6

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., Mager, D.L., Feschotte, C., 2018. Ten things you should know about transposable elements. Genome Biol. 19, 199. https://doi.org/10.1186/s13059-018-1577-z

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bairoch, A., 2007. UniProtKB/Swiss-Prot. Methods Mol. Biol. Clifton NJ 406, 89–112. https://doi.org/10.1007/978-1-59745-535-0_4

Boyko, C., Ravoahangimalala, O., Randriamasimanana, D., Razafindrazaka, T., 2005. Astacoides hobbsi, a new crayfish (Crustacea: Decapoda: Parastacidae) from Madagascar. Zootaxa 1091, 41–51. https://doi.org/10.11646/zootaxa.1091.1.3

Braband, A., Kawai, T., Scholtz, G., 2006. The phylogenetic position of the East Asian freshwater crayfish Cambaroides within the Northern Hemisphere Astacoidea (Crustacea, Decapoda, Astacida) based on molecular data. J. Zool. Syst. Evol. Res. 44, 17–24. https://doi.org/10.1111/j.1439-0469.2005.00338.x

Bracken Grissom, H., Toon, A., Felder, D., Martin, J., Finley, M., Rasmussen, J., Palero, F., Crandall, K., 2009. The Decapod Tree of Life: Compiling the Data and Moving toward a Consensus of Decapod Evolution. Arthropod Syst Phylogeny 67. https://doi.org/10.3897/asp.67.e31691

Breinholt, J., Pérez-Losada, M., Crandall, K., 2009. The Timing of the Diversification of the Freshwater Crayfishes, in: Decapod Crustacean Phylogenetics. pp. 305–318. https://doi.org/10.1201/9781420092592-c17

Brent, M.R., 2005. Genome annotation past, present, and future: How to define an ORF at each locus. Genome Res. 15, 1777–1786. https://doi.org/10.1101/gr.3866105

Brinkmann, K., Schell, M., Hoppe, T., Kashkar, H., 2015. Regulation of the DNA damage response by ubiquitin conjugation. Front. Genet. 6, 98. https://doi.org/10.3389/fgene.2015.00098

Briones-Fourzán, P., Hendrickx, M.E., 2022. Ecology and Diversity of Marine Decapod Crustaceans. Diversity 14, 614. https://doi.org/10.3390/d14080614

Burns, K.H., Boeke, J.D., 2012. Human Transposon Tectonics. Cell 149, 740–752. https://doi.org/10.1016/j.cell.2012.04.019

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

Campbell, L.J., Hammond, S.A., Price, S.J., Sharma, M.D., Garner, T.W.J., Birol, I., Helbing, C.C., Wilfert, L., Griffiths, A.G.F., 2018. A novel approach to wildlife transcriptomics provides evidence of disease-mediated differential expression and changes to the microbiome of amphibian populations. Mol. Ecol. 27, 1413–1427. https://doi.org/10.1111/mec.14528

Carroll, J., Van Oostende, N., Ward, B.B., 2022. Evaluation of Genomic Sequence-Based Growth Rate Methods for Synchronized Synechococcus Cultures. Appl. Environ. Microbiol. 88, e0174321. https://doi.org/10.1128/AEM.01743-21

Cechova, M., 2020. Probably Correct: Rescuing Repeats with Short and Long Reads. Genes 12, 48. https://doi.org/10.3390/genes12010048

Cerenius, L., Bangyeekhun, E., Keyser, P., Söderhäll, I., Söderhäll, K., 2003. Host prophenoloxidase expression in freshwater crayfish is linked to increased resistance to the crayfish plague fungus, Aphanomyces astaci. Cell. Microbiol. 5, 353–357. https://doi.org/10.1046/j.1462-5822.2003.00282.x

Cerenius, L., Söderhäll, K., Persson, M., Ajaxon, R., 1988. The crayfish plague fungus Aphanomyces astaci-diagnosis, isolation and pathobiology. Freshw. Crayfish 7, l–144.

Cha, S., Bird, D.M., 2016. Optimizing k-mer size using a variant grid search to enhance de novo genome assembly. Bioinformation 12, 36–40. https://doi.org/10.6026/97320630012036

Chalopin, D., Naville, M., Plard, F., Galiana, D., Volff, J.-N., 2015. Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates. Genome Biol. Evol. 7, 567–580. https://doi.org/10.1093/gbe/evv005

Check Hayden, E., 2012. Nanopore genome sequencer makes its debut. Nature. https://doi.org/10.1038/nature.2012.10051

Chen, H., Zhang, R., Liu, Feng, Shao, C., Liu, Fangfang, Li, W., Ren, J., Niu, B., Liu, H., Lou, B., 2023. The chromosome-level genome of Cherax quadricarinatus. Sci. Data 10, 215. https://doi.org/10.1038/s41597-023-02124-z

Chen, Q., Lan, C., Zhao, L., Wang, J., Chen, B., Chen, Y.-P.P., 2017. Recent advances in sequence assembly: principles and applications. Brief. Funct. Genomics 16, 361–378. https://doi.org/10.1093/bfgp/elx006

Chen, S., Huang, T., Zhou, Y., Han, Y., Xu, M., Gu, J., 2017. AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. BMC Bioinformatics 18, 80. https://doi.org/10.1186/s12859-017-1469-3

Chen, T., Mu, S., Sun, Z., Zhang, H., Li, C., Guo, M., Li, Y., Kang, X., Wang, Z., 2020. Spermiogenic histone transitions and chromatin decondensation in Decapoda. Theriogenology 156, 242–252. https://doi.org/10.1016/j.theriogenology.2020.07.003

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li, H., 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods 18, 170–175. https://doi.org/10.1038/s41592-020-01056-5

Chikhi, R., Medvedev, P., 2014. Informed and automated k-mer size selection for genome assembly. Bioinformatics 30, 31–37. https://doi.org/10.1093/bioinformatics/btt310

Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., Turner, S.W., Korlach, J., 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods 10, 563–569. https://doi.org/10.1038/nmeth.2474

Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2015. GenBank. Nucleic Acids Res. 44, D67. https://doi.org/10.1093/nar/gkv1276

Collins, G., Schneider, C., Boštjančić, L.L., Burkhardt, U., Christian, A., Decker, P., Ebersberger, I., Hohberg, K., Lecompte, O., Merges, D., Muelbaier, H., Romahn, J., Römbke, J., Rutz, C., Schmelz, R., Schmidt, A., Theissinger, K., Veres, R., Lehmitz, R., Pfenninger, M., Bálint, M., 2023. The MetaInvert soil invertebrate genome resource provides insights into below-ground biodiversity and evolution. Commun. Biol. 6, 1–12. https://doi.org/10.1038/s42003-023-05621-4

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., Mortazavi, A., 2016. A survey of best practices for RNA-seq data analysis. Genome Biol. 17, 13. https://doi.org/10.1186/s13059-016-0881-8

Coombe, L., Li, J.X., Lo, T., Wong, J., Nikolic, V., Warren, R.L., Birol, I., 2021. LongStitch: high-quality genome assembly correction and scaffolding using long reads. BMC Bioinformatics 22, 534. https://doi.org/10.1186/s12859-021-04451-7

Cosma, B.-M., Shirali Hossein Zade, R., Jordan, E.N., van Lent, P., Peng, C., Pillay, S., Abeel, T., 2023. Evaluating long-read de novo assembly tools for eukaryotic genomes: insights and considerations. GigaScience 12, giad100. https://doi.org/10.1093/gigascience/giad100

Cottee-Jones, H.E.W., Whittaker, R.J., 2012. perspective: The keystone species concept: a critical appraisal. Front. Biogeogr. 4. https://doi.org/10.21425/F5FBG12533

Cowley, M., Oakey, R.J., 2013. Transposable Elements Re-Wire and Fine-Tune the Transcriptome. PLOS Genet. 9, e1003234. https://doi.org/10.1371/journal.pgen.1003234

Crandall, K.A., Buhay, J.E., 2008. Global diversity of crayfish (Astacidae, Cambaridae, and Parastacidae—Decapoda) in freshwater, in: Balian, E.V., Lévêque, C., Segers, H., Martens, K. (Eds.), Freshwater Animal Diversity Assessment, Developments in Hydrobiology. Springer Netherlands, Dordrecht, pp. 295–301. https://doi.org/10.1007/978-1-4020-8259-7_32

Crandall, K.A., De Grave, S., 2017. An updated classification of the freshwater crayfishes (Decapoda: Astacidea) of the world, with a complete species list. J. Crustac. Biol. 37, 615–653. https://doi.org/10.1093/jcbiol/rux070

Crandall, Keith A., Fetzner, J.W., Jr., Jara, C.G., Buckup, L., 2000. On the Phylogenetic Positioning of the South American Freshwater Crayfish Genera (Decapoda: Parastacidae). J. Crustac. Biol. 20, 530–540. https://doi.org/10.1163/20021975-99990069

Crandall, K A, Harris, D.J., Fetzner, J.W., 2000. The monophyletic origin of freshwater crayfish estimated from nuclear and mitochondrial DNA sequences. Proc. R. Soc. B Biol. Sci. 267, 1679–1686. https://doi.org/10.1098/rspb.2000.1195

Creed, R., 2009. Decapoda, in: Likens, G.E. (Ed.), Encyclopedia of Inland Waters. Academic Press, Oxford, pp. 271–279. https://doi.org/10.1016/B978-012370626-3.00169-1

Creed, R.P., Reed, J.M., 2004. Ecosystem engineering by crayfish in a headwater stream community. J. North Am. Benthol. Soc. 23, 224–236. http://dx.doi.org/10.1899/0887-3593(2004)023%3C0224:EEBCIA%3E2.0.CO;2

Cui, Z., Liu, Y., Yuan, J., Zhang, X., Ventura, T., Ma, K.Y., Sun, S., Song, C., Zhan, D., Yang, Y., Liu, H., Fan, G., Cai, Q., Du, J., Qin, J., Shi, C., Hao, S., Fitzgibbon, Q.P., Smith, G.G., Xiang, J., Chan, T.-Y., Hui, M., Bao, C., Li, F., Chu, K.H., 2021. The Chinese mitten crab genome provides insights into adaptive plasticity and developmental regulation. Nat. Commun. 12, 2395. https://doi.org/10.1038/s41467-021-22604-3

Cukerzis, J.M., 1987. On the Origin of Freshwater Crayfish (Astacura). Freshw. Crayfish 7, 343–349.

Davies, K., 2002. Cracking the Genome. Johns Hopkins University Press. https://doi.org/10.56021/9780801871405

De Grave, S., Decock, W., Dekeyzer, S., Davie, P.J.F., Fransen, C.H.J.M., Boyko, C.B., Poore, G.C.B., Macpherson, E., Ahyong, S.T., Crandall, K.A., De Mazancourt, V., Osawa, M., Chan, T.-Y., Ng, P.K.L., Lemaitre, R., Van Der Meij, S.E.T., Santos, S., 2023. Benchmarking global biodiversity of decapod crustaceans (Crustacea: Decapoda). J. Crustac. Biol. 43, ruad042. https://doi.org/10.1093/jcbiol/ruad042

De Grave, S., Pentcheff, N.D., Ahyong, S.T., Chan, T.-Y., Crandall, K.A., Dworschak, P.C., Felder, D.L., Feldmann, R.M., Fransen, C.H.J.M., Goulding, L.Y.D., Lemaitre, R., Low, M.E.Y., Martin, J.W., Ng, P.K.L., Schweitzer, C.E., Tan, S.H., Tshudy, D., Wetzer, R., 2009. A Classification of Living and Fossil Genera of Decapod Crustaceans.

Deiana, A.M., Cau, A., Coluccia, E., Cannas, R., Milia, A., Salvadori, S., Libertini, A., 1999. Genome Size and At-Dna Content in Thirteen Species of Decapoda, in: Crustaceans and the Biodiversity Crisis. Brill, pp. 981–985. https://doi.org/10.1163/9789004630543_076

Deininger, P.L., Moran, J.V., Batzer, M.A., Kazazian, H.H., 2003. Mobile elements and mammalian genome evolution. Curr. Opin. Genet. Dev. 13, 651–658. https://doi.org/10.1016/j.gde.2003.10.013

Diéguez-Uribeondo, J., Huang, T.-S., Cerenius, L., Söderhäll, K., 1995. Physiological adaptation of an Aphanomyces astaci strain isolated from the freshwater crayfish Procambarus clarkii. Mycol. Res. 99, 574–578. https://doi.org/10.1016/S0953-7562(09)80716-8

Diéguez-Uribeondo, J., Royo, F., Souty-Grosset, C., Ropiquet, A., Grandjean, F., 2008. Low genetic variability of the white-clawed crayfish in the Iberian Peninsula: its origin and management implications. Aquat. Conserv. Mar. Freshw. Ecosyst. 18, 19–31. https://doi.org/10.1002/aqc.811

Dodsworth, S., Chase, M.W., Kelly, L.J., Leitch, I.J., Macas, J., Novák, P., Piednoël, M., Weiss-Schneeweiss, H., Leitch, A.R., 2015. Genomic Repeat Abundances Contain Phylogenetic Signal. Syst. Biol. 64, 112–126. https://doi.org/10.1093/sysbio/syu080

Dodsworth, S., Jang, T.-S., Struebig, M., Chase, M.W., Weiss-Schneeweiss, H., Leitch, A.R., 2017. Genome-wide repeat dynamics reflect phylogenetic distance in closely related allotetraploid Nicotiana (Solanaceae). Plant Syst. Evol. 303, 1013–1020. https://doi.org/10.1007/s00606-016-1356-9

Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J.-F., Vlasova, A., Leskosek, B.L., Soler, L., Binzer-Panchal, M., Lantz, H., 2018. Ten steps to get started in Genome Assembly and Annotation. F1000Research 7, ELIXIR-148. https://doi.org/10.12688/f1000research.13598.1

Dorn, N.J., Wojdak, J.M., 2004. The role of omnivorous crayfish in littoral communities. Oecologia 140, 150–159. https://doi.org/10.1007/s00442-004-1548-9

Du, Z., Jin, Y., Ren, D., 2016. In-depth comparative transcriptome analysis of intestines of red swamp crayfish, Procambarus clarkii, infected with WSSV. Sci. Rep. 6, 26780. https://doi.org/10.1038/srep26780

Edsman, L., Füreder, L., Gherardi, F., Souty-Grosset, C., 2010. IUCN Red List of Threatened Species: Astacus astacus. IUCN Red List Threat. Species.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., Turner, S., 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. Science 323, 133–138. https://doi.org/10.1126/science.1162986

Ekblom, R., Wolf, J.B.W., 2014. A field guide to whole-genome sequencing, assembly and annotation. Evol. Appl. 7, 1026–1042. https://doi.org/10.1111/eva.12178

Elliott, T.A., Gregory, T.R., 2015. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos. Trans. R. Soc. B Biol. Sci. 370, 20140331. https://doi.org/10.1098/rstb.2014.0331

Emms, D.M., Kelly, S., 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20, 238. https://doi.org/10.1186/s13059-019-1832-y

Fernandez-Gutierrez, A., Gutierrez-Gonzalez, J.J., 2021. Bioinformatic-Based Approaches for Disease-Resistance Gene Discovery in Plants. Agronomy 11, 2259. https://doi.org/10.3390/agronomy11112259

Feron, R., Waterhouse, R.M., 2022. Assessing species coverage and assembly quality of rapidly accumulating sequenced genomes. GigaScience 11, giac006. https://doi.org/10.1093/gigascience/giac006

Finkers, R., van Kaauwen, M., Ament, K., Burger-Meijer, K., Egging, R., Huits, H., Kodde, L., Kroon, L., Shigyo, M., Sato, S., Vosman, B., van Workum, W., Scholten, O., 2021. Insights from the first genome assembly of Onion (Allium cepa). G3 GenesGenomesGenetics 11, jkab243. https://doi.org/10.1093/g3journal/jkab243

Fitch, W.M., 1970. Distinguishing Homologous from Analogous Proteins. Syst. Biol. 19, 99–113. https://doi.org/10.2307/2412448

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., Smit, A.F., 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. 117, 9451–9457. https://doi.org/10.1073/pnas.1921046117

Freedman, K.J., Otto, L.M., Ivanov, A.P., Barik, A., Oh, S.-H., Edel, J.B., 2016. Nanopore sensing at ultra-low concentrations using single-molecule dielectrophoretic trapping. Nat. Commun. 7, 10217. https://doi.org/10.1038/ncomms10217

Fu, A., Wang, Q., Mu, J., Ma, L., Wen, C., Zhao, X., Gao, L., Li, J., Shi, K., Wang, Y., Zhang, Xuechuan, Zhang, Xuewen, Wang, F., Grierson, D., Zuo, J., 2021. Combined genomic, transcriptomic, and metabolomic analyses provide insights into chayote (Sechium edule) evolution and fruit development. Hortic. Res. 8, 1–15. https://doi.org/10.1038/s41438-021-00487-1

Füreder, L., Gherardi, F., Holdich, D., Reynolds, J., Sibley, P., Souty-Grosset, C., 2010. IUCN Red List of Threatened Species: Austropotamobius pallipes. IUCN Red List Threat. Species.

Garrido-Ramos, M.A., 2017. Satellite DNA: An Evolving Topic. Genes 8, 230. https://doi.org/10.3390/genes8090230

Gatzmann, F., Falckenhayn, C., Gutekunst, J., Hanna, K., Raddatz, G., Carneiro, V.C., Lyko, F., 2018. The methylome of the marbled crayfish links gene body methylation to stable expression of poorly accessible genes. Epigenetics Chromatin 11, 57. https://doi.org/10.1186/s13072-018-0229-6

Gnerre, S., Lander, E.S., Lindblad-Toh, K., Jaffe, D.B., 2009. Assisted assembly: how to improve a de novo genome assembly by using related species. Genome Biol. 10, R88. https://doi.org/10.1186/gb-2009-10-8-r88

Gong, Y., Li, Y., Liu, X., Ma, Y., Jiang, L., 2023. A review of the pangenome: how it affects our understanding of genomic variation, selection and breeding in domestic animals? J. Anim. Sci. Biotechnol. 14, 73. https://doi.org/10.1186/s40104-023-00860-1

González-Tizón, A.M., Rojo, V., Menini, E., Torrecilla, Z., Martínez-Lage, A., 2013. Karyological Analysis of the Shrimp Palaemon Serratus (Decapoda: Palaemonidae). J. Crustac. Biol. 33, 843–848. https://doi.org/10.1163/1937240X-00002185

Goodwin, S., McPherson, J.D., McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333–351. https://doi.org/10.1038/nrg.2016.49

Gouin, N., Grandjean, F., Souty-Grosset, C., 2006. Population genetic structure of the endangered crayfish Austropotamobius pallipes in France based on microsatellite variation: biogeographical inferences and conservation implications. Freshw. Biol. 51, 1369–1387. https://doi.org/10.1111/j.1365-2427.2006.01570.x

Graveley, B.R., 2001. Alternative splicing: increasing diversity in the proteomic world. Trends Genet. 17, 100–107. https://doi.org/10.1016/S0168-9525(00)02176-4

Green, E., 2023. Genomics [WWW Document]. Genome.gov. URL https://www.genome.gov/genetics-glossary/genomics (accessed 9.21.23).

Green II, D.A., Kronforst, M.R., 2019. Monarch butterflies use an environmentally sensitive, internal timer to control overwintering dynamics. Mol. Ecol. 28, 3642–3655. https://doi.org/10.1111/mec.15178

Gregory, T.R., 2023. Animal Genome Size Database. http: //www.genomesize.com.

Grishanin, A., 2024. Chromatin diminution as a tool to study some biological problems. Comp. Cytogenet. 18, 27–49. https://doi.org/10.3897/compcytogen.17.112152

Gross, R., Palm, S., Kõiv, K., Prestegaard, T., Jussila, J., Paaver, T., Geist, J., Kokko, H., Karjalainen, A., Edsman, L., 2013. Microsatellite markers reveal clear geographic structuring among threatened noble crayfish (Astacus astacus) populations in Northern and Central Europe. Conserv. Genet. 14, 809–821. https://doi.org/10.1007/s10592-013-0476-9

Gutekunst, J., Andriantsoa, R., Falckenhayn, C., Hanna, K., Stein, W., Rasamy, J., Lyko, F., 2018. Clonal genome evolution and rapid invasive spread of the marbled crayfish. Nat. Ecol. Evol. 2, 567–573. https://doi.org/10.1038/s41559-018-0467-9

Hamr, P., 1992. A revision of the Tasmanian freshwater crayfish genus Astacopsis Huxley (Decapoda: Parastacidae). Pap. Proc. R. Soc. Tasman. 126, 91–94. https://doi.org/10.26749/rstpp.126.91

Henry, R.P., Lucu, Č., Onken, H., Weihrauch, D., 2012. Multiple functions of the crustacean gill: osmotic/ionic regulation, acid-base balance, ammonia excretion, and bioaccumulation of toxic metals. Front. Physiol. 3, 431. https://doi.org/10.3389/fphys.2012.00431

Hessen, D.O., Persson, J., 2009. Genome size as a determinant of growth and life-history traits in crustaceans. Biol. J. Linn. Soc. 98, 393–399. https://doi.org/10.1111/j.1095-8312.2009.01285.x

Heuertz, M., Carvalho, S.B., Galindo, J., Rinkevich, B., Robakowski, P., Aavik, T., Altinok, I., Barth, J.M.I., Cotrim, H., Goessen, R., González-Martínez, S.C., Grebenc, T., Hoban, S., Kopatz, A., McMahon, B.J., Porth, I., Raeymaekers, J.A.M., Träger, S., Valdecantos, A., Vella, A., Vernesi, C., Garnier-Géré, P., 2023. The application gap: Genomics for biodiversity and ecosystem service management. Biol. Conserv. 278, 109883. https://doi.org/10.1016/j.biocon.2022.109883

Hobbs, H.H., 1989. An Illustrated Checklist of the American Crayfishes (Decapoda, Astacidae, Cambaridae, Parastacidae), in: Smithsonian Contributions to Zoology. pp. 1–236. https://doi.org/10.5479/si.00810282.480

Hobbs, H.H., 1987. A Review of the Crayfish Genus Astacoides (Decapoda: Parastacidae), in: Smithsonian Contributions to Zoology. pp. 1–50. https://doi.org/10.5479/si.00810282.443

Hobbs, H.H., 1974. Synopsis of the Families and Genera of Crayfishes (Cruatacea: Decapoda).

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., Quesneville, H., 2014. PASTEC: an automatic transposable element classification tool. PloS One 9, e91929. https://doi.org/10.1371/journal.pone.0091929

Holdich, D.M., Reynolds, J.D., Souty-Grosset, C., Sibley, P.J., 2009. A review of the ever increasing threat to European crayfish from non-indigenous crayfish species. Knowl. Manag. Aquat. Ecosyst. 11. https://doi.org/10.1051/kmae/2009025

Holthuis, L.B., 1982. Freshwater Crustacea Decapoda of New Guinea, in: Gressitt, J.L. (Ed.), Biogeography and Ecology of New Guinea: Part One - Seven, Monographiae Biologicae. Springer Netherlands, Dordrecht, pp. 603–619. https://doi.org/10.1007/978-94-009-8632-9_28

Holthuis, L.B., 1952. The Crustacea Decapoda Macrura of Chile: Con Resumen en Español. C.W.K. Gleerup.

Hood, L.E., Hunkapiller, M.W., Smith, L.M., 1987. Automated DNA sequencing and analysis of the human genome. Genomics 1, 201–212. https://doi.org/10.1016/0888-7543(87)90046-2

Hotaling, S., Wilcox, E.R., Heckenhauer, J., Stewart, R.J., Frandsen, P.B., 2023. Highly accurate long reads are crucial for realizing the potential of biodiversity genomics. BMC Genomics 24, 117. https://doi.org/10.1186/s12864-023-09193-9

Huang, N., Li, H., 2023. compleasm: a faster and more accurate reimplementation of BUSCO. Bioinformatics 39, btad595. https://doi.org/10.1093/bioinformatics/btad595

Huang, T., Cerenius, L., Söderhäll, K., 1994. Analysis of genetic diversity in the crayfish plague fungus, Aphanomyces astaci, by random amplification of polymorphic DNA. Aquaculture 126, 1–9. https://doi.org/10.1016/0044-8486(94)90243-7

Huerlimann, R., Cowley, J.A., Wade, N.M., Wang, Y., Kasinadhuni, N., Chan, C.-K.K., Jabbari, J.S., Siemering, K., Gordon, L., Tinning, M., Montenegro, J.D., Maes, G.E., Sellars, M.J., Coman, G.J., McWilliam, S., Zenger, K.R., Khatkar, M.S., Raadsma, H.W., Donovan, D., Krishna, G., Jerry, D.R., 2022. Genome assembly of the Australian black tiger shrimp (Penaeus monodon) reveals a novel fragmented IHHNV EVE sequence. G3 GenesGenomesGenetics 12, jkac034. https://doi.org/10.1093/g3journal/jkac034

Hurt, C., Hildreth, P., Williams, C., 2022. A genomic perspective on the conservation status of the endangered Nashville crayfish (Faxonius shoupi). Conserv. Genet. 23, 589–604. https://doi.org/10.1007/s10592-022-01438-6

Hynes, R.O., 2002. Integrins: Bidirectional, Allosteric Signaling Machines. Cell 110, 673–687. https://doi.org/10.1016/S0092-8674(02)00971-6

Iannucci, A., Saha, A., Cannicci, S., Bellucci, A., Cheng, C.L.Y., Ng, K.H., Fratini, S., 2022. Ecological, physiological and life-history traits correlate with genome sizes in decapod crustaceans. Front. Ecol. Evol. 10. https://doi.org/10.3389/fevo.2022.930888

Jackman, S.D., Vandervalk, B.P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S.A., Jahesh, G., Khan, H., Coombe, L., Warren, R.L., Birol, I., 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res. 27, 768–777. https://doi.org/10.1101/gr.214346.116

Jain, M., Tyson, J.R., Loose, M., Ip, C.L.C., Eccles, D.A., O'Grady, J., Malla, S., Leggett, R.M., Wallerman, O., Jansen, H.J., Zalunin, V., Birney, E., Brown, B.L., Snutch, T.P., Olsen, H.E., 2017. MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. F1000Research 6, 760. https://doi.org/10.12688/f1000research.11354.1

Jeck, W.R., Reinhardt, J.A., Baltrus, D.A., Hickenbotham, M.T., Magrini, V., Mardis, E.R., Dangl, J.L., Jones, C.D., 2007. Extending assembly of short DNA sequences to handle error. Bioinformatics 23, 2942–2944. https://doi.org/10.1093/bioinformatics/btm451

Jenkins, G.M., Frohman, M.A., 2005. Phospholipase D: a lipid centric review. Cell. Mol. Life Sci. CMLS 62, 2305–2316. https://doi.org/10.1007/s00018-005-5195-z

Jenkins, T.L., Ellis, C.D., Stevens, J.R., 2019. SNP discovery in European lobster (Homarus gammarus) using RAD sequencing. Conserv. Genet. Resour. 11, 253–257. https://doi.org/10.1007/s12686-018-1001-8

Jeschke, J.M., Bacher, S., Blackburn, T.M., Dick, J.T.A., Essl, F., Evans, T., Gaertner, M., Hulme, P.E., Kühn, I., Mrugała, A., Pergl, J., Pyšek, P., Rabitsch, W., Ricciardi, A., Richardson, D.M., Sendek, A., VilÀ, M., Winter, M., Kumschick, S., 2014. Defining the Impact of Non-Native Species. Conserv. Biol. 28, 1188–1194. https://doi.org/10.1111/cobi.12299

Jimenez, A.G., Kinsey, S.T., Dillaman, R.M., Kapraun, D.F., 2010. Nuclear DNA content variation associated with muscle fiber hypertrophic growth in decapod crustaceans. Genome 53, 161–171. https://doi.org/10.1139/g09-095

Jin, S., Bian, C., Jiang, S., Han, K., Xiong, Y., Zhang, W., Shi, C., Qiao, H., Gao, Z., Li, R., Huang, Y., Gong, Y., You, X., Fan, G., Shi, Q., Fu, H., 2021. A chromosome-level genome assembly of the oriental river prawn, Macrobrachium nipponense. GigaScience 10, giaa160. https://doi.org/10.1093/gigascience/giaa160

Johnson, T., Seymour, R., Padgett, D., 2002. Biology and Systematics of the Saprolegniaceae.

Jones, A., Schwessinger, B., 2020. Sorbitol washing complex homogenate for improved DNA extractions. https://dx.doi.org/10.17504/protocols.io.beuvjew6

Jones, J., Razanabolana, J., Harvey, A., Toon, A., Oidtmann, B., Randrianarison, M., Raminosoa, N., Ravoahangimalala, O., 2008. The perfect invader: A parthenogenic crayfish poses a new threat to Madagascar's freshwater biodiversity. Biol. Invasions 11, 1475–1482. https://doi.org/10.1007/s10530-008-9334-y

Jünemann, S., Prior, K., Albersmeier, A., Albaum, S., Kalinowski, J., Goesmann, A., Stoye, J., Harmsen, D., 2014. GABenchToB: a genome assembly benchmark tuned on bacteria and benchtop sequencers. PloS One 9, e107014. https://doi.org/10.1371/journal.pone.0107014

Jung, H., Ventura, T., Chung, J.S., Kim, W.-J., Nam, B.-H., Kong, H.J., Kim, Y.-O., Jeon, M.-S., Eyun, S., 2020. Twelve quick steps for genome assembly and annotation in the classroom. PLOS Comput. Biol. 16, e1008325. https://doi.org/10.1371/journal.pcbi.1008325

Jurka, J., Kapitonov, V.V., Kohany, O., Jurka, M.V., 2007. Repetitive Sequences in Complex Genomes: Structure and Evolution. Annu. Rev. Genomics Hum. Genet. 8, 241–259. https://doi.org/10.1146/annurev.genom.8.080706.092416

Jussila, Japo, Edsman, L., Maguire, I., Diéguez-Uribeondo, J., Theissinger, K., 2021. Money Kills Native Ecosystems: European Crayfish as an Example. Front. Ecol. Evol. 9. https://doi.org/10.3389/fevo.2021.648495

Jussila, J., Francesconi, C., Theissinger, K., Kokko, H., Makkonen, J., 2021. Is Aphanomyces astaci Loosing its Stamina: A Latent Crayfish Plague Disease Agent From Lake Venesjärvi, Finland. http://dx.doi.org/10.5869/fc.2021.v26-2.139

Jussila, J., Vrezec, A., Jaklič, T., Kukkonen, H., Makkonen, J., Kokko, H., 2017. Aphanomyces astaci isolate from latently infected stone crayfish (Austropotamobius torrentium) population is virulent. J. Invertebr. Pathol. 149, 15–20. https://doi.org/10.1016/j.jip.2017.07.003

Kanehisa, M., Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30. https://doi.org/10.1093/nar/28.1.27

Katneni, V.K., Shekhar, M.S., Jangam, A.K., Krishnan, K., Prabhudas, S.K., Kaikkolante, N., Baghel, D.S., Koyadan, V.K., Jena, J., Mohapatra, T., 2022. A Superior Contiguous Whole Genome Assembly for Shrimp (Penaeus indicus). Front. Mar. Sci. 8.

Kawai, T., Cumberlidge, N. (Eds.), 2016. A Global Overview of the Conservation of Freshwater Decapod Crustaceans. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-42527-6

Kawato, S., Nishitsuji, K., Arimoto, A., Hisata, K., Kawamitsu, M., Nozaki, R., Kondo, H., Shinzato, C., Ohira, T., Satoh, N., Shoguchi, E., Hirono, I., 2021. Genome and transcriptome assemblies of the kuruma shrimp, Marsupenaeus japonicus. G3 GenesGenomesGenetics 11, jkab268. https://doi.org/10.1093/g3journal/jkab268

Keinath, M.C., Timoshevskiy, V.A., Timoshevskaya, N.Y., Tsonis, P.A., Voss, S.R., Smith, J.J., 2015. Initial characterization of the large genome of the salamander Ambystoma mexicanum using shotgun and laser capture chromosome sequencing. Sci. Rep. 5, 16413. https://doi.org/10.1038/srep16413

Kim, Y.-J., Lee, J., Han, K., 2012. Transposable Elements: No More "Junk DNA." Genomics Inform. 10, 226–233. https://doi.org/10.5808/GI.2012.10.4.226

Kimura, K., Itonori, S., Kajiwara, C., Hada, N., Takeda, T., Sugita, M., 2014. Structural Elucidation of the Neutral Glycosphingolipids, Mono-, Di-, Tri- and Tetraglycosylceramides from the Marine Crab Erimacrus isenbeckii. J. Oleo Sci. 63, 269–280. https://doi.org/10.5650/jos.ess13156

Ko, B.J., Lee, C., Kim, J., Rhie, A., Yoo, D.A., Howe, K., Wood, J., Cho, S., Brown, S., Formenti, G., Jarvis, E.D., Kim, H., 2022. Widespread false gene gains caused by duplication errors in genome assemblies. Genome Biol. 23, 205. https://doi.org/10.1186/s13059-022-02764-1

Koepfli, K.-P., Paten, B., Genome 10K Community of Scientists, O'Brien, S.J., 2015. The Genome 10K Project: a way forward. Annu. Rev. Anim. Biosci. 3, 57–111. https://doi.org/10.1146/annurev-animal-090414-014900

Kohany, O., Gentles, A.J., Hankus, L., Jurka, J., 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7, 474. https://doi.org/10.1186/1471-2105-7-474

Kojima, K.K., 2019. Structural and sequence diversity of eukaryotic transposable elements. Genes Genet. Syst. 94, 233–252. https://doi.org/10.1266/ggs.18-00024

Kolmogorov, M., Yuan, J., Lin, Y., Pevzner, P.A., 2019. Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37, 540–546. https://doi.org/10.1038/s41587-019-0072-8

Koning, A.P.J. de, Gu, W., Castoe, T.A., Batzer, M.A., Pollock, D.D., 2011. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. PLOS Genet. 7, e1002384. https://doi.org/10.1371/journal.pgen.1002384

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27, 722–736. https://doi.org/10.1101/gr.215087.116

Kouba, A., Petrusek, A., Kozák, P., 2014. Continental-wide distribution of crayfish species in Europe: update and maps. Knowl. Manag. Aquat. Ecosyst. 05. https://doi.org/10.1051/kmae/2014007

Kozák, P., Duris, Z., Petrusek, A., Buřič, M., Horká, I., Kouba, A., Kozubíková-Balcarová, E., Policar, T., 2015. Crayfish Biology and Culture.

Kozubíková, E., Viljamaa-Dirks, S., Heinikainen, S., Petrusek, A., 2011. Spiny-cheek crayfish Orconectes limosus carry a novel genotype of the crayfish plague pathogen Aphanomyces astaci. J. Invertebr. Pathol. 108, 214–216. https://doi.org/10.1016/j.jip.2011.08.002

Kuzmin, D.A., Feranchuk, S.I., Sharov, V.V., Cybin, A.N., Makolov, S.V., Putintseva, Y.A., Oreshkova, N.V., Krutovsky, K.V., 2019. Stepwise large genome assembly approach: a case of Siberian larch (Larix sibirica Ledeb). BMC Bioinformatics 20, 37. https://doi.org/10.1186/s12859-018-2570-y

Lanciano, S., Mirouze, M., 2018. Transposable elements: all mobile, all different, some stress responsive, some adaptive? Curr. Opin. Genet. Dev., Genome Architecture and Expression 49, 106–114. https://doi.org/10.1016/j.gde.2018.04.002

Lécher, P., Defaye, D., Noel, P., 1995. Chromosomes and nuclear DNA of Crustacea. Invertebr. Reprod. Dev. 27, 85–114. https://doi.org/10.1080/07924259.1995.9672440

Leplae, R., Hebrant, A., Wodak, S.J., Toussaint, A., 2004. ACLAME: A CLAssification of Mobile genetic Elements. Nucleic Acids Res. 32, D45–D49. https://doi.org/10.1093/nar/gkh084

Levene, M.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G., Webb, W.W., 2003. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. Science 299, 682–686. https://doi.org/10.1126/science.1079700

Lewin, H.A., Richards, S., Lieberman Aiden, E., Allende, M.L., Archibald, J.M., Bálint, M., Barker, K.B., Baumgartner, B., Belov, K., Bertorelle, G., Blaxter, M.L., Cai, J., Caperello, N.D., Carlson, K., Castilla-Rubio, J.C., Chaw, S.-M., Chen, L., Childers, A.K., Coddington, J.A., Conde, D.A., Corominas, M., Crandall, K.A., Crawford, A.J., DiPalma, F., Durbin, R., Ebenezer, T.E., Edwards, S.V., Fedrigo, O., Flicek, P., Formenti, G., Gibbs, R.A., Gilbert, M.T.P., Goldstein, M.M., Graves, J.M., Greely, H.T., Grigoriev, I.V., Hackett, K.J., Hall, N., Haussler, D., Helgen, K.M., Hogg, C.J., Isobe, S., Jakobsen, K.S., Janke, A., Jarvis, E.D., Johnson, W.E., Jones, S.J.M., Karlsson, E.K., Kersey, P.J., Kim, J.-H., Kress, W.J., Kuraku, S., Lawniczak, M.K.N., Leebens-Mack, J.H., Li, X., Lindblad-Toh, K., Liu, X., Lopez, J.V., Marques-Bonet, T., Mazard, S., Mazet, J.A.K., Mazzoni, C.J., Myers, E.W., O'Neill, R.J., Paez, S., Park, H., Robinson, G.E., Roquet, C., Ryder, O.A., Sabir, J.S.M., Shaffer, H.B., Shank, T.M., Sherkow, J.S., Soltis, P.S., Tang, B., Tedersoo, L., Uliano-Silva, M., Wang, K., Wei, X., Wetzer, R., Wilson, J.L., Xu, X., Yang, H., Yoder, A.D., Zhang, G., 2022. The Earth BioGenome Project 2020: Starting the clock. Proc. Natl. Acad. Sci. 119, e2115635118. https://doi.org/10.1073/pnas.2115635118

Lewis, S.H., Ross, L., Bain, S.A., Pahita, E., Smith, S.A., Cordaux, R., Miska, E.A., Lenhard, B., Jiggins, F.M., Sarkies, P., 2020. ------Widespread conservation and lineage-specific diversification of genome-wide DNA methylation patterns across arthropods. PLoS Genet. 16, e1008864. https://doi.org/10.1371/journal.pgen.1008864

Li, D., Liu, C.-M., Luo, R., Sadakane, K., Lam, T.-W., 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Li, H., 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32, 2103–2110. https://doi.org/10.1093/bioinformatics/btw152

Li, L., Stoeckert, C.J., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178–2189. https://doi.org/10.1101/gr.1224503

Li, S.-F., Su, T., Cheng, G.Q., Wang, B.-X., Li, X., Deng, C., Wujun, G., 2017. Chromosome Evolution in Connection with Repetitive Sequences and Epigenetics in Plants. Genes 8, 290. https://doi.org/10.3390/genes8100290

Liao, M., Xu, M., Hu, R., Xu, Z., Bonvillain, C., Li, Y., Li, X., Luo, X., Wang, Jianghua, Wang, Jie, Zhao, S., Gu, Z., 2024. The chromosome-level genome assembly of the red swamp crayfish Procambarus clarkii. Sci. Data 11, 885. https://doi.org/10.1038/s41597-024-03718-x

Liao, X., Zhu, W., Zhou, J., Li, H., Xu, X., Zhang, B., Gao, X., 2023. Repetitive DNA sequence detection and its role in the human genome. Commun. Biol. 6, 1–21. https://doi.org/10.1038/s42003-023-05322-y

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S., Dekker, J., 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326, 289–293. https://doi.org/10.1126/science.1181369

Lischer, H.E.L., Shimizu, K.K., 2017. Reference-guided de novo assembly approach improves genome reconstruction for related species. BMC Bioinformatics 18, 474. https://doi.org/10.1186/s12859-017-1911-6

Liu, L., Cui, Z., Song, C., Liu, Y., Hui, M., Wang, C., 2016. Flow cytometric analysis of DNA content for four commercially important crabs in China. Acta Oceanol. Sin. 35, 7–11. https://doi.org/10.1007/s13131-016-0876-z

Liu, M., Ge, S., Bhandari, S., Fan, C., Jiao, Y., Gai, C., Wang, Y., Liu, H., 2022. Genome characterization and comparative analysis among three swimming crab species. Front. Mar. Sci. 9. https://doi.org/10.3389/fmars.2022.895119100 of the World's Worst Invasive Alien Species: A Selection from the Global Invasive Species Database. T

Liu, N., Low, W.Y., Alinejad-Rokny, H., Pederson, S., Sadlon, T., Barry, S., Breen, J., 2021. Seeing the forest through the trees: prioritising potentially functional interactions from Hi-C. Epigenetics Chromatin 14, 41. https://doi.org/10.1186/s13072-021-00417-4

Liu, Z., Zheng, J., Li, H., Fang, K., Wang, S., He, Jian, Zhou, D., Weng, S., Chi, M., Gu, Z., He, Jianguo, Li, F., Wang, M., 2024. Genome assembly of redclaw crayfish (Cherax quadricarinatus) provides insights into its immune adaptation and hypoxia tolerance. BMC Genomics 25, 746. https://doi.org/10.1186/s12864-024-10673-9

Longshaw, M., 2011. Diseases of crayfish: A review. J. Invertebr. Pathol., Diseases of Edible Crustaceans 106, 54–70. https://doi.org/10.1016/j.jip.2010.09.013

Lovrenčić, L., Bonassin, L., Boštjančić, L.L., Podnar, M., Jelić, M., Klobučar, G., Jaklič, M., Slavevska-Stamenković, V., Hinić, J., Maguire, I., 2020. New insights into the genetic diversity of the stone crayfish: taxonomic and conservation implications. BMC Evol. Biol. 20, 146. https://doi.org/10.1186/s12862-020-01709-1

Lovrenčić, L., Temunović, M., Gross, R., Grgurev, M., Maguire, I., 2022. Integrating population genetics and species distribution modelling to guide conservation of the noble crayfish, Astacus astacus, in Croatia. Sci. Rep. 12, 2040. https://doi.org/10.1038/s41598-022-06027-8

Lowe, S.J., Browne, M., Boudjelas, S., De Poorter, M., 2004. 100 of the World's Worst Invasive Alien Species: A Selection from the Global Invasive Species Database. The Invasive Species Specialist Group (ISSG), a specialist group of the Species Survival Commission (SSC) of the IUCN, Gland, Switzerland.

Lu, H., Giordano, F., Ning, Z., 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics Proteomics Bioinformatics, SI: Big Data and Precision Medicine 14, 265–279. https://doi.org/10.1016/j.gpb.2016.05.004

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Yunjie, Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Yong, Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, Jian, Lam, T.-W., Wang, Jun, 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1, 18. https://doi.org/10.1186/2047-217X-1-18

Luo, X., Chen, S., Zhang, Y., 2022. PlantRep: a database of plant repetitive elements. Plant Cell Rep. 41, 1163–1166. https://doi.org/10.1007/s00299-021-02817-y

Lüscher, C., Malenka, R.C., 2012. NMDA Receptor-Dependent Long-Term Potentiation and Long-Term Depression (LTP/LTD). Cold Spring Harb. Perspect. Biol. 4, a005710. https://doi.org/10.1101/cshperspect.a005710

Lv, J., Li, R., Su, Z., Gao, B., Ti, X., Yan, D., Liu, G., Liu, P., Wang, C., Li, J., 2022. A chromosome-level genome of Portunus trituberculatus provides insights into its evolution, salinity adaptation and sex determination. Mol. Ecol. Resour. 22, 1606–1625. https://doi.org/10.1111/1755-0998.13564

Macas, J., Neumann, P., Navrátilová, A., 2007. Repetitive DNA in the pea (Pisum sativum L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and Medicago truncatula. BMC Genomics 8, 427. https://doi.org/10.1186/1471-2164-8-427

Maciaszek, R., Jabłońska, A., Prati, S., Wróblewski, P., Gruszczyńska, J., Świderek, W., 2022. Marbled crayfish Procambarus virginalis invades a nature reserve: how to stop further introductions? Eur. Zool. J. 89, 888–901. https://doi.org/10.1080/24750263.2022.2095046

Maguire, A., Jussila, J., Edsman, L., Maguire, I., Diguez-Uribéondo, J., Theissinger, K., 2022. The Crayfish Tale – A short & entertaining educational film, Available from: https://www.youtube.com/watch?v=oZLIaCJOwPo.

Makkonen, J., Jussila, J., Panteleit, J., Keller, N.S., Schrimpf, A., Theissinger, K., Kortet, R., Martín-Torrijos, L., Sandoval-Sierra, J.V., Diéguez-Uribeondo, J., Kokko, H., 2018. MtDNA allows the sensitive detection and haplotyping of the crayfish plague disease agent Aphanomyces astaci showing clues about its origin and migration. Parasitology 145, 1210–1218. https://doi.org/10.1017/S0031182018000227

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., Zdobnov, E.M., 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol. Biol. Evol. 38, 4647–4654. https://doi.org/10.1093/molbev/msab199

Mardis, E.R., 2008. The impact of next-generation sequencing technology on genetics. Trends Genet. 24, 133–141. https://doi.org/10.1016/j.tig.2007.12.007

Mariappan, P., Balasundaram, C., Schmitz, B., 2000. Decapod crustacean chelipeds: an overview. J. Biosci. 25, 301–313. https://doi.org/10.1007/BF02703939

Martin, F.J., Amode, M.R., Aneja, A., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S.K., Bignell, A., Boddu, S., Branco Lins, P.R.,

Brooks, L., Ramaraju, S.B., Charkhchi, M., Cockburn, A., Da Rin Fiorretto, L., Davidson, C., Dodiya, K., Donaldson, S., El Houdaigui, B., El Naboulsi, T., Fatima, R., Giron, C.G., Genez, T., Ghattaoraya, G.S., Martinez, J.G., Guijarro, C., Hardy, M., Hollis, Z., Hourlier, T., Hunt, T., Kay, M., Kaykala, V., Le, T., Lemos, D., Marques-Coelho, D., Marugán, J.C., Merino, G.A., Mirabueno, L.P., Mushtaq, A., Hossain, S.N., Ogeh, D.N., Sakthivel, M.P., Parker, A., Perry, M., Piližota, I., Prosovetskaia, I., Pérez-Silva, J.G., Salam, A.I.A., Saraiva-Agostinho, N., Schuilenburg, H., Sheppard, D., Sinha, S., Sipos, B., Stark, W., Steed, E., Sukumaran, R., Sumathipala, D., Suner, M.-M., Surapaneni, L., Sutinen, K., Szpak, M., Tricomi, F.F., Urbina-Gómez, D., Veidenberg, A., Walsh, T.A., Walts, B., Wass, E., Willhoft, N., Allen, J., Alvarez-Jarreta, J., Chakiachvili, M., Flint, B., Giorgetti, S., Haggerty, L., Ilsley, G.R., Loveland, J.E., Moore, B., Mudge, J.M., Tate, J., Thybert, D., Trevanion, S.J., Winterbottom, A., Frankish, A., Hunt, S.E., Ruffier, M., Cunningham, F., Dyer, S., Finn, R.D., Howe, K.L., Harrison, P.W., Yates, A.D., Flicek, P., 2023. Ensembl 2023. Nucleic Acids Res. 51, D933–D941. https://doi.org/10.1093/nar/gkac958

Martín-Torrijos, L., Martínez-Ríos, M., Casabella-Herrero, G., Adams, S.B., Jackson, C.R., Diéguez-Uribeondo, J., 2021. Tracing the origin of the crayfish plague pathogen, Aphanomyces astaci, to the Southeastern United States. Sci. Rep. 11, 9332. https://doi.org/10.1038/s41598-021-88704-8

Maxam, A.M., Gilbert, W., 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. U. S. A. 74, 560–564. https://doi.org/10.1073/pnas.74.2.560

Mazzoni, C.J., Ciofi, C., Waterhouse, R.M., 2023. Biodiversity: an atlas of European reference genomes. Nature 619, 252–252. https://doi.org/10.1038/d41586-023-02229-w

McClain, W.R., 2020. Crayfish Aquaculture, in: Lovrich, G., Thiel, M. (Eds.), Fisheries and Aquaculture: Volume 9. Oxford University Press, p. 0. https://doi.org/10.1093/oso/9780190865627.003.0011

McGrath, L.L., Vollmer, S.V., Kaluziak, S.T., Ayers, J., 2016. De novo transcriptome assembly for the lobster Homarus americanus and characterization of differential gene expression across nervous system tissues. BMC Genomics 17, 63. https://doi.org/10.1186/s12864-016-2373-3

Meng, X., Fu, Q., Luan, S., Luo, K., Sui, J., Kong, J., 2021. Genome survey and high-resolution genetic map provide valuable genetic resources for Fenneropenaeus chinensis. Sci. Rep. 11, 7533. https://doi.org/10.1038/s41598-021-87237-4

Metzker, M.L., 2010. Sequencing technologies — the next generation. Nat. Rev. Genet. 11, 31–46. https://doi.org/10.1038/nrg2626

Meyerson, L.A., Pyšek, P., Lučanová, M., Wigginton, S., Tran, C.-T., Cronin, J.T., 2020. Plant genome size influences stress tolerance of invasive and native plants via plasticity. Ecosphere 11, e03145. https://doi.org/10.1002/ecs2.3145

Michaeli, Y., Ebenstein, Y., 2012. Channeling DNA for optical mapping. Nat. Biotechnol. 30, 762–763. https://doi.org/10.1038/nbt.2324

Miga, K.H., 2015. Completing the human genome: the progress and challenge of satellite DNA assembly. Chromosome Res. 23, 421–426. https://doi.org/10.1007/s10577-015-9488-2

Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A.J., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T.R., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L.S., Kawahara, A.Y., Krogmann, L., Kubiak, M., Lanfear,

R., Letsch, H., Li, Yiyuan, Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D.D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J.L., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von Reumont, B.M., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N.U., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M.G., Wiegmann, B.M., Wilbrandt, J., Wipfler, B., Wong, T.K.F., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D.K., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Yunhui, Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Yingrui, Xu, X., Zhang, Yong, Yang, H., Wang, Jian, Wang, Jun, Kjer, K.M., Zhou, X., 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346, 763–767. https://doi.org/10.1126/science.1257570

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D., Bateman, A., 2021. Pfam: The protein families database in 2021. Nucleic Acids Res. 49, D412–D419. https://doi.org/10.1093/nar/gkaa913

Mlinarec, J., M, M., Pavlica, M., Šrut, M., Klobučar, G., Maguire, I., 2011. Comparative Karyotype Investigations in the European Crayfish Astacus astacus and A. leptodactylus (Decapoda, Astacidae). Crustaceana 84. https://doi.org/10.2307/23065234

Momot, W.T., 1995. Redefining the role of crayfish in aquatic ecosystems. Rev. Fish. Sci. 3, 33–63. https://doi.org/10.1080/10641269509388566

Mravinac, B., Plohl, M., Ugarković, Đ., 2005. Preservation and High Sequence Conservation of Satellite DNAs Suggest Functional Constraints. J. Mol. Evol. 61, 542–550. https://doi.org/10.1007/s00239-004-0342-y

Muffato, M., Louis, A., Nguyen, N.T.T., Lucas, J., Berthelot, C., Roest Crollius, H., 2023. Reconstruction of hundreds of reference ancestral genomes across the eukaryotic kingdom. Nat. Ecol. Evol. 7, 355–366. https://doi.org/10.1038/s41559-022-01956-z

Nagasawa, H., 2012. The crustacean cuticle: Structure, composition and mineralization. Front. Biosci. Elite Ed. 4, 711–20. https://doi.org/10.2741/E412

Nature Biotechnology, 2018. A reference standard for genome biology. Nat. Biotechnol. 36, 1121–1121. https://doi.org/10.1038/nbt.4318

Neale, D.B., Zimin, A.V., Zaman, S., Scott, A.D., Shrestha, B., Workman, R.E., Puiu, D., Allen, B.J., Moore, Z.J., Sekhwal, M.K., De La Torre, A.R., McGuire, P.E., Burns, E., Timp, W., Wegrzyn, J.L., Salzberg, S.L., 2022. Assembled and annotated 26.5 Gbp coast redwood genome: a resource for estimating evolutionary adaptive potential and investigating hexaploid origin. G3 GenesGenomesGenetics 12, jkab380. https://doi.org/10.1093/g3journal/jkab380

Nevers, Y., Defosset, A., Lecompte, O., 2020. Orthology: Promises and Challenges, in: Pontarotti, P. (Ed.), Evolutionary Biology—A Transdisciplinary Approach. Springer International Publishing, Cham, pp. 203–228. https://doi.org/10.1007/978-3-030-57246-4_9

Nevers, Y., Jones, T.E.M., Jyothi, D., Yates, B., Ferret, M., Portell-Silva, L., Codo, L., Cosentino, S., Marcet-Houben, M., Vlasova, A., Poidevin, L., Kress, A., Hickman, M., Persson, E., Piližota, I., Guijarro-Clarke, C., the OpenEBench team the Quest for Orthologs Consortium, Iwasaki, W., Lecompte, O., Sonnhammer, E., Roos, D.S., Gabaldón, T., Thybert, D., Thomas, P.D., Hu, Y., Emms, D.M., Bruford, E., Capella-Gutierrez, S., Martin, M.J., Dessimoz, C., Altenhoff, A., 2022. The Quest for Orthologs orthology

benchmark service in 2022. Nucleic Acids Res. 50, W623–W632. https://doi.org/10.1093/nar/gkac330

Nevers, Y., Kress, A., Defosset, A., Ripp, R., Linard, B., Thompson, J.D., Poch, O., Lecompte, O., 2019. OrthoInspector 3.0: open portal for comparative genomics. Nucleic Acids Res. 47, D411–D418. https://doi.org/10.1093/nar/gky1068

Nie, Y., Liu, X., Zhao, L., Huang, Y., 2024. Repetitive element expansions contribute to genome size gigantism in Pamphagidae: A comparative study (Orthoptera, Acridoidea). Genomics 116, 110896. https://doi.org/10.1016/j.ygeno.2024.110896

Niiyama, H., 1962. On the Unprecedentedly Large Number of Chromosomes of the Crayfish, Astacus trowbridgii Stimpson. Annot Zool Jpn.

Nishijima, S., Nishikawa, C., Miyashita, T., 2017. Habitat modification by invasive crayfish can facilitate its growth through enhanced food accessibility. BMC Ecol. 17, 37. https://doi.org/10.1186/s12898-017-0147-7

Novák, P., Ávila Robledillo, L., Koblížková, A., Vrbová, I., Neumann, P., Macas, J., 2017. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. Nucleic Acids Res. 45, e111. https://doi.org/10.1093/nar/gkx257

Novák, P., Neumann, P., Pech, J., Steinhaisl, J., Macas, J., 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29, 792–793. https://doi.org/10.1093/bioinformatics/btt054

Nowoshilow, S., Schloissnig, S., Fei, J.-F., Dahl, A., Pang, A.W.C., Pippel, M., Winkler, S., Hastie, A.R., Young, G., Roscito, J.G., Falcon, F., Knapp, D., Powell, S., Cruz, A., Cao, H., Habermann, B., Hiller, M., Tanaka, E.M., Myers, E.W., 2018. The axolotl genome and the evolution of key tissue formation regulators. Nature 554, 50–55. https://doi.org/10.1038/nature25458

Ojeda-Martinez, D., Diaz, I., Santamaria, M.E., Ortego, F., 2024. Comparative genomics reveals carbohydrate enzymatic fluctuations and herbivorous adaptations in arthropods. Comput. Struct. Biotechnol. J. 23, 3744–3758. https://doi.org/10.1016/j.csbj.2024.10.027

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733-745. https://doi.org/10.1093/nar/gkv1189

PacBio Revio | Long-read sequencing at scale [WWW Document], n.d. . PacBio. URL https://www.pacb.com/revio/ (accessed 9.21.23).

Paine, R. t., 1995. A Conversation on Refining the Concept of Keystone Species. Conserv. Biol. 9, 962–964. https://doi.org/10.1046/j.1523-1739.1995.09040962.x

Paine, R.T., 1969. The Pisaster-Tegula Interaction: Prey Patches, Predator Food Preference, and Intertidal Community Structure. Ecology 50, 950–961. https://doi.org/10.2307/1936888

Pârvulescu, L., 2019. Introducing a new Austropotamobius crayfish species (Crustacea, Decapoda, Astacidae): A Miocene endemism of the Apuseni Mountains, Romania. Zool. Anz. 279, 94–102. https://doi.org/10.1016/j.jcz.2019.01.006

Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B.L., Salazar, G.A., Bileschi, M.L., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D.H., Letunić, I., Marchler-Bauer, A., Mi, H., Natale, D.A., Orengo, C.A., Pandurangan, A.P., Rivoire, C., Sigrist, C.J.A., Sillitoe, I., Thanki, N., Thomas, P.D., Tosatto, S.C.E., Wu, C.H., Bateman, A., 2023. InterPro in 2022. Nucleic Acids Res. 51, D418–D427. https://doi.org/10.1093/nar/gkac993

Pedro, D.L.F., Amorim, T.S., Varani, A., Guyot, R., Domingues, D.S., Paschoal, A.R., 2021. An Atlas of Plant Transposable Elements. F1000Research 10, 1194. https://doi.org/10.12688/f1000research.74524.1

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O., 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc. Natl. Acad. Sci. 96, 4285. https://doi.org/10.1073/pnas.96.8.4285

Peltola, H., Söderlund, H., Ukkonen, E., 1984. SEQAID: a DNA sequence assembling program based on a mathematical model. Nucleic Acids Res. 12, 307–321. https://doi.org/10.1093/nar/12.1Part1.307

Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., Cruaud, C., Labadie, K., Aury, J.-M., Istace, B., Lebrigand, K., Barbry, P., Engelen, S., Lemainque, A., Wincker, P., Liti, G., Schacherer, J., 2018. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. Nature 556, 339–344. https://doi.org/10.1038/s41586-018-0030-5

Petersen, M., Armisén, D., Gibbs, R.A., Hering, L., Khila, A., Mayer, G., Richards, S., Niehuis, O., Misof, B., 2019. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. BMC Ecol. Evol. 19, 11. https://doi.org/10.1186/s12862-018-1324-9

Pezer, Ž., Brajković, J., Feliciello, I., Ugarković, Đ., 2012. Satellite DNA-Mediated Effects on Genome Regulation, in: Genome Dynamics. pp. 153–169. https://doi.org/10.1159/000337116

Platt, R.N., Vandewege, M.W., Ray, D.A., 2018. Mammalian transposable elements and their impacts on genome evolution. Chromosome Res. 26, 25–43. https://doi.org/10.1007/s10577-017-9570-z

Plohl, M., Luchetti, A., Meštrović, N., Mantovani, B., 2008. Satellite DNAs between selfishness and functionality: Structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. Gene 409, 72–82. https://doi.org/10.1016/j.gene.2007.11.013

Plohl, M., Meštrović, N., Mravinac, B., 2012. Satellite DNA Evolution. Repetitive DNA 7, 126–152. https://doi.org/10.1159/000337122

Polinski, J.M., Zimin, A.V., Clark, K.F., Kohn, A.B., Sadowski, N., Timp, W., Ptitsyn, A., Khanna, P., Romanova, D.Y., Williams, P., Greenwood, S.J., Moroz, L.L., Walt, D.R., Bodnar, A.G., 2021. The American lobster genome reveals insights on longevity, neural, and immune adaptations. Sci. Adv. 7, eabe8290. https://doi.org/10.1126/sciadv.abe8290

Poore, G.C.B., 2004. Marine Decapod Crustacea of Southern Australia: A Guide to Identification. Csiro Publishing.

Pop, M., 2009. Genome assembly reborn: recent computational challenges. Brief. Bioinform. 10, 354–366. https://doi.org/10.1093/bib/bbp026

Porter, M.L., Pérez-Losada, M., Crandall, K.A., 2005. Model-based multi-locus estimation of decapod phylogeny and divergence times. Mol. Phylogenet. Evol. 37, 355–369. https://doi.org/10.1016/j.ympev.2005.06.021

Procedure & Checklist - Preparing HiFi SMRTbell Libraries from Ultra-Low DNA Input, 2021.

Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J.W., Cunningham, C.W., 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463, 1079–1083. https://doi.org/10.1038/nature08742

Ren, X., Lv, J., Liu, M., Wang, Q., Shao, H., Liu, P., Li, J., 2022. A chromosome-level genome of the kuruma shrimp (Marsupenaeus japonicus) provides insights into its evolution and cold-resistance mechanism. Genomics 114, 110373. https://doi.org/10.1016/j.ygeno.2022.110373

Reuter, J.A., Spacek, D., Snyder, M.P., 2015. High-Throughput Sequencing Technologies. Mol. Cell 58, 586–597. https://doi.org/10.1016/j.molcel.2015.05.004

Rey, O., Danchin, E., Mirouze, M., Loot, C., Blanchet, S., 2016. Adaptation to Global Change: A Transposable Element–Epigenetics Perspective. Trends Ecol. Evol. 31, 514–526. https://doi.org/10.1016/j.tree.2016.03.013

Rey, O., Eizaguirre, C., Angers, B., Baltazar-Soares, M., Sagonas, K., Prunier, J.G., Blanchet, S., 2020. Linking epigenetics and biological conservation: Towards a conservation epigenetics perspective. Funct. Ecol. 34, 414–427. https://doi.org/10.1111/1365-2435.13429

Reynolds, J., Souty-Grosset, C., Richardson, A., 2013. Ecological Roles of Crayfish in Freshwater and Terrestrial Habitats. Freshw. Crayfish 19, 197–218.

Reynolds, J.D., 2011. A review of ecological interactions between crayfish and fish, indigenous and introduced. Knowl. Manag. Aquat. Ecosyst. 10. https://doi.org/10.1051/kmae/2011024

Rezinciuc, S., Sandoval-Sierra, J., Oidtmann, B., Diéguez-Uribeondo, J., 2015. The Biology of Crayfish Plague Pathogen Aphanomyces astaci: Current Answers to Most Frequent Questions. pp. 182–204. https://doi.org/10.1201/b18723-12

Rheinsmith, E.L., Hinegardner, R., Bachmann, K., 1974. Nuclear DNA amounts in crustacea. Comp. Biochem. Physiol. Part B Comp. Biochem. 48, 343–348. https://doi.org/10.1016/0305-0491(74)90269-7

Rich, P.R., 2003. The molecular machinery of Keilin's respiratory chain. Biochem. Soc. Trans. 31, 1095–1105. https://doi.org/10.1042/bst0311095

Richards, S., 2019. Arthropod Genome Sequencing and Assembly Strategies, in: Brown, S.J., Pfrender, M.E. (Eds.), Insect Genomics: Methods and Protocols, Methods in Molecular Biology. Springer, New York, NY, pp. 1–14. https://doi.org/10.1007/978-1-4939-8775-7_1

Roach, M.J., Schmidt, S.A., Borneman, A.R., 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 19, 460. https://doi.org/10.1186/s12859-018-2485-7

Ruan, J., Li, H., 2020. Fast and accurate long-read assembly with wtdbg2. Nat. Methods 17, 155–158. https://doi.org/10.1038/s41592-019-0669-3

Ruiz-Ruano, F.J., López-León, M.D., Cabrero, J., Camacho, J.P.M., 2016. High-throughput analysis of the satellitome illuminates satellite DNA evolution. Sci. Rep. 6, 28333. https://doi.org/10.1038/srep28333

Rutz, C., Bonassin, L., Kress, A., Francesconi, C., Boštjančić, L.L., Merlat, D., Theissinger, K., Lecompte, O., 2023. Abundance and Diversification of Repetitive Elements in Decapoda Genomes. Genes 14, 1627. https://doi.org/10.3390/genes14081627

Ryan Gregory, T., 2002. Genome size and developmental complexity. Genetica 115, 131–146. https://doi.org/10.1023/A:1016032400147

Sakoparnig, T., Field, C., van Nimwegen, E., 2021. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. eLife 10, e65366. https://doi.org/10.7554/eLife.65366

Sambrook, J., Russell, D.W., 2006. Purification of nucleic acids by extraction with phenol:chloroform. CSH Protoc. 2006, pdb.prot4455. https://doi.org/10.1101/pdb.prot4455

Sánchez-Roncancio, C., García, B., Gallardo-Hidalgo, J., Yáñez, J.M., 2022. GWAS on Imputed Whole-Genome Sequence Variants Reveal Genes Associated with Resistance to Piscirickettsia salmonis in Rainbow Trout (Oncorhynchus mykiss). Genes 14, 114. https://doi.org/10.3390/genes14010114

Sanger, F., Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. 94, 441–448. https://doi.org/10.1016/0022-2836(75)90213-2

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74, 5463–5467.

SanMiguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., Bennetzen, J.L., 1996. Nested Retrotransposons in the Intergenic Regions of the Maize Genome. Science 274, 765–768. https://doi.org/10.1126/science.274.5288.765

Sarwat, M., Yamdagni, M.M., 2016. DNA barcoding, microarrays and next generation sequencing: recent tools for genetic diversity estimation and authentication of medicinal plants. Crit. Rev. Biotechnol. 36, 191–203. https://doi.org/10.3109/07388551.2014.947563

Scalzitti, N., Jeannin-Girardon, A., Collet, P., Poch, O., Thompson, J., 2020. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. BMC Genomics 21, 293. https://doi.org/10.1186/s12864-020-6707-9

Schram, F.R., 2009. On the Origin of Decapoda, in: Decapod Crustacean Phylogenetics. CRC Press.

Schram, F.R., Feldmann, R.M., Copeland, M.J., 1978. The Late Devonian Palaeopalaemonidae and the Earliest Decapod Crustaceans. J. Paleontol. 52, 1375–1387.

Schrimpf, A., 2014. Invasive Chinese mitten crab (Eriocheir sinensis) transmits crayfish plague pathogen (Aphanomyces astaci). Aquat. Invasions 9, 203–209. https://doi.org/10.3391/ai.2014.9.2.09

Schrimpf, A., Piscione, M., Cammaerts, R., Collas, M., Herman, D., Jung, A., Ottburg, F., Roessink, I., Rollin, X., Schulz, R., Theissinger, K., 2017. Genetic characterization of Western European noble crayfish populations (Astacus astacus) for advanced conservation management strategies. Conserv. Genet. 18, 1299–1315. https://doi.org/10.1007/s10592-017-0981-3

Schrimpf, A., Theissinger, K., Dahlem, J., Maguire, I., Pârvulescu, L., Schulz, H.K., Schulz, R., 2014. Phylogeography of noble crayfish (Astacus astacus) reveals multiple refugia. Freshw. Biol. 59, 761–776. https://doi.org/10.1111/fwb.12302

Schulz, R., 2000. Status of the noble crayfish Astacus astacus (L.) in Germany : monitoring protocol and the use of RAPD markers to assess the genetic structure of populations. Bull. Fr. Pêche Piscic. 123–138. https://doi.org/10.1051/kmae:2000007

Schumann, G.G., Gogvadze, E.V., Osanai-Futahashi, M., Kuroki, A., Münk, C., Fujiwara, H., Ivics, Z., Buzdin, A.A., 2010. Chapter Three - Unique Functions of Repetitive Transcriptomes, in: Jeon, K.W. (Ed.), International Review of Cell and Molecular Biology. Academic Press, pp. 115–188. https://doi.org/10.1016/B978-0-12-381047-2.00003-7

Schwentner, M., Combosch, D.J., Nelson, J.P., Giribet, G., 2017. A Phylogenomic Solution to the Origin of Insects by Resolving Crustacean-Hexapod Relationships. Curr. Biol. 27, 1818-1824.e5. https://doi.org/10.1016/j.cub.2017.05.040

Secomandi, S., Gallo, G.R., Sozzoni, M., Iannucci, A., Galati, E., Abueg, L., Balacco, J., Caprioli, M., Chow, W., Ciofi, C., Collins, J., Fedrigo, O., Ferretti, L., Fungtammasan, A., Haase, B., Howe, K., Kwak, W., Lombardo, G., Masterson, P., Messina, G., Møller, A.P., Mountcastle, J., Mousseau, T.A., Ferrer Obiol, J., Olivieri, A., Rhie, A., Rubolini, D., Saclier, M., Stanyon, R., Stucki, D., Thibaud-Nissen, F., Torrance, J., Torroni, A., Weber, K., Ambrosini, R., Bonisoli-Alquati, A., Jarvis, E.D., Gianfranceschi, L., Formenti, G., 2023. A chromosome-level reference genome and pangenome for barn swallow population genomics. Cell Rep. 42, 111992. https://doi.org/10.1016/j.celrep.2023.111992

Seppey, M., Manni, M., Zdobnov, E.M., 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol. Biol. Clifton NJ 1962, 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14

Shao, C., Sun, S., Liu, K., Wang, Jiahao, Li, S., Liu, Q., Deagle, B.E., Seim, I., Biscontin, A., Wang, Q., Liu, X., Kawaguchi, S., Liu, Yalin, Jarman, S., Wang, Yue, Wang, H.-Y., Huang, G., Hu, J., Feng, B., Pittà, C.D., Liu, Shanshan, Wang, R., Ma, K., Ying, Y., Sales, G., Sun, T., Wang, X., Zhang, Y., Zhao, Y., Pan, S., Hao, X., Wang, Yang, Xu, J., Yue, B., Sun, Y., Zhang, H., Xu, M., Liu, Yuyan, Jia, X., Zhu, J., Liu, Shufang, Ruan, J., Zhang, G., Yang, H., Xu, X., Wang, Jun, Zhao, X., Meyer, B., Fan, G., 2023. The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights. Cell 186, 1279-1294.e19. https://doi.org/10.1016/j.cell.2023.02.005

Shapiro, J.A., Sternberg, R. von, 2005. Why repetitive DNA is essential to genome function. Biol. Rev. 80, 227–250. https://doi.org/10.1017/S1464793104006657

Shi, L., Yi, S., Li, Y., 2018. Genome survey sequencing of red swamp crayfish Procambarus clarkii. Mol. Biol. Rep. 45, 799–806. https://doi.org/10.1007/s11033-018-4219-3

Simpson, J.T., Pop, M., 2015. The Theory and Practice of Genome Sequence Assembly. Annu. Rev. Genomics Hum. Genet. 16, 153–172. https://doi.org/10.1146/annurev-genom-090314-050032

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M., Birol, İ., 2009. ABySS: A parallel assembler for short read sequence data. Genome Res. 19, 1117–1123. https://doi.org/10.1101/gr.089532.108

Sinclair, E., Fetzner Jr, J., Buhay, J., Crandall, K., 2004. Proposal to complete a phylogenetic taxonomy and systematic revision for freshwater crayfish (Astacida). Freshw. Crayfish 14, 21–29. https://doi.org/10.5869/fc.2004.v14.021

Skurdal, J., Taugbøl, T., 2002. Astacus. Biol. Freshw. Crayfish 467–510.

Smit, A.F., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9, 657–663. https://doi.org/10.1016/s0959-437x(99)00031-3

Smit, AFA, Hubley, R & Green, P, 2013. RepeatMasker Open-4.0. <http://www.repeatmasker.org>.

Söderhäll, K., Cerenius, L., 1999. The crayfish plague fungus: History and recent advances. Freshw. Crayfish 12, 11–35.

Sohn, J., Nam, J.-W., 2018. The present and future of de novo whole-genome assembly. Brief. Bioinform. 19, 23–40. https://doi.org/10.1093/bib/bbw096

Sotero-Caio, C.G., Platt, R.N., II, Suh, A., Ray, D.A., 2017. Evolution and Diversity of Transposable Elements in Vertebrate Genomes. Genome Biol. Evol. 9, 161–177. https://doi.org/10.1093/gbe/evw264

Souty-Grosset, C., Anastácio, P.M., Aquiloni, L., Banha, F., Choquer, J., Chucholl, C., Tricarico, E., 2016. The red swamp crayfish Procambarus clarkii in Europe: Impacts on aquatic ecosystems and human well-being. Limnologica 58, 78–93. https://doi.org/10.1016/j.limno.2016.03.003

Souty-Grosset, C., Holdich, D.D.M., Noël, P.Y., Reynolds, J., Haffner, P., 2006. Atlas of crayfish in Europe. [WWW Document].

Souty-Grosset, C., Reynolds, J.D., 2009. Current ideas on methodological approaches in European crayfish conservation and restocking procedures. Knowl. Manag. Aquat. Ecosyst. 01. https://doi.org/10.1051/kmae/2009021

Sproul, J.S., Hotaling, S., Heckenhauer, J., Powell, A., Larracuente, A.M., Kelley, J.L., Pauls, S.U., Frandsen, P.B., 2022. Repetitive elements in the era of biodiversity genomics: insights from 600+ insect genomes (preprint). Genomics. https://doi.org/10.1101/2022.06.02.494618

Sproul, J.S., Hotaling, S., Heckenhauer, J., Powell, A., Marshall, D., Larracuente, A.M., Kelley, J.L., Pauls, S.U., Frandsen, P.B., 2023. Analyses of 600+ insect genomes reveal repetitive element dynamics and highlight biodiversity-scale repeat annotation challenges. Genome Res. 33, 1708–1717. https://doi.org/10.1101/gr.277387.122

Statzner, B., Peltret, O., Tomanová, S., 2003. Crayfish as geomorphic agents and ecosystem engineers: effect of a biomass gradient on baseflow and flood-induced transport of gravel and sand in experimental streams. https://doi.org/10.1046/J.1365-2427.2003.00984.X

Stoler, N., Nekrutenko, A., 2021. Sequencing error profiles of Illumina sequencing instruments. NAR Genomics Bioinforma. 3, lqab019. https://doi.org/10.1093/nargab/lqab019

Storer, J., Hubley, R., Rosen, J., Wheeler, T.J., Smit, A.F., 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob. DNA 12, 2. https://doi.org/10.1186/s13100-020-00230-y

Svoboda, J., Mrugała, A., Kozubíková-Balcarová, E., Kouba, A., Diéguez-Uribeondo, J., Petrusek, A., 2014a. Resistance to the crayfish plague pathogen, Aphanomyces astaci, in two freshwater shrimps. J. Invertebr. Pathol. 121, 97–104. https://doi.org/10.1016/j.jip.2014.07.004

Svoboda, J., Mrugała, A., Kozubíková-Balcarová, E., Petrusek, A., 2017. Hosts and transmission of the crayfish plague pathogen Aphanomyces astaci: a review. J. Fish Dis. 40, 127–140. https://doi.org/10.1111/jfd.12472

Svoboda, J., Strand, D.A., Vrålstad, T., Grandjean, F., Edsman, L., Kozák, P., Kouba, A., Fristad, R.F., Bahadir Koca, S., Petrusek, A., 2014b. The crayfish plague pathogen can infect

freshwater-inhabiting crabs. Freshw. Biol. 59, 918–929. https://doi.org/10.1111/fwb.12315

Swathi, A., Shekhar, M.S., Katneni, V.K., Vijayan, K.K., 2018. Genome size estimation of brackishwater fishes and penaeid shrimps by flow cytometry. Mol. Biol. Rep. 45, 951–960. https://doi.org/10.1007/s11033-018-4243-3

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L., Fang, T., Doncheva, N.T., Pyysalo, S., Bork, P., Jensen, L.J., von Mering, C., 2023. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 51, D638–D646. https://doi.org/10.1093/nar/gkac1000

Tan, M.H., Gan, H.M., Lee, Y.P., Grandjean, F., Croft, L.J., Austin, C.M., 2020. A Giant Genome for a Giant Crayfish (Cherax quadricarinatus) With Insights Into cox1 Pseudogenes in Decapod Genomes. Front. Genet. 11. https://doi.org/10.3389/fgene.2020.00201

Tang, B., Wang, Zhongkai, Liu, Q., Wang, Zhengfei, Ren, Y., Guo, H., Qi, T., Li, Yuetian, Zhang, H., Jiang, S., Ge, B., Xuan, F., Sun, Y., She, S., Yam Chan, T., Sha, Z., Jiang, H., Li, H., Jiang, W., Qin, Y., Wang, K., Qiu, Q., Wang, W., Li, X., Ng, N.K., Zhang, D., Li, Yongxin, 2021. Chromosome-level genome assembly of Paralithodes platypus provides insights into evolution and adaptation of king crabs. Mol. Ecol. Resour. 21, 511–525. https://doi.org/10.1111/1755-0998.13266

Tang, B., Wang, Zhongkai, Liu, Q., Zhang, H., Jiang, S., Li, X., Wang, Zhengfei, Sun, Y., Sha, Z., Jiang, H., Wu, X., Ren, Y., Li, H., Xuan, F., Ge, B., Jiang, W., She, S., Sun, H., Qiu, Q., Wang, W., Wang, Q., Qiu, G., Zhang, D., Li, Y., 2020a. High-Quality Genome Assembly of Eriocheir japonica sinensis Reveals Its Unique Genome Evolution. Front. Genet. 10. https://doi.org/10.3389/fgene.2019.01340

Tang, B., Zhang, D., Li, H., Jiang, S., Zhang, H., Xuan, F., Ge, B., Wang, Zhengfei, Liu, Y., Sha, Z., Cheng, Y., Jiang, W., Jiang, H., Wang, Zhongkai, Wang, K., Li, C., Sun, Y., She, S., Qiu, Q., Wang, W., Li, X., Li, Y., Liu, Q., Ren, Y., 2020b. Chromosome-level genome assembly reveals the unique genome evolution of the swimming crab (Portunus trituberculatus). GigaScience 9, giz161. https://doi.org/10.1093/gigascience/giz161

Tenaillon, M.I., Hufford, M.B., Gaut, B.S., Ross-Ibarra, J., 2011. Genome size and transposable element content as determined by high-throughput sequencing in maize and Zea luxurians. Genome Biol. Evol. 3, 219–229. https://doi.org/10.1093/gbe/evr008

The Gene Ontology Consortium, Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N.L., Hill, D.P., Lee, R., Mi, H., Moxon, S., Mungall, C.J., Muruganugan, A., Mushayahama, T., Sternberg, P.W., Thomas, P.D., Van Auken, K., Ramsey, J., Siegele, D.A., Chisholm, R.L., Fey, P., Aspromonte, M.C., Nugnes, M.V., Quaglia, F., Tosatto, S., Giglio, M., Nadendla, S., Antonazzo, G., Attrill, H., dos Santos, G., Marygold, S., Strelets, V., Tabone, C.J., Thurmond, J., Zhou, P., Ahmed, S.H., Asanitthong, P., Luna Buitrago, D., Erdol, M.N., Gage, M.C., Ali Kadhum, M., Li, K.Y.C., Long, M., Michalak, A., Pesala, A., Pritazahra, A., Saverimuttu, S.C.C., Su, R., Thurlow, K.E., Lovering, R.C., Logie, C., Oliferenko, S., Blake, J., Christie, K., Corbani, L., Dolan, M.E., Drabkin, H.J., Hill, D.P., Ni, L., Sitnikov, D., Smith, C., Cuzick, A., Seager, J., Cooper, L., Elser, J., Jaiswal, P., Gupta, P., Jaiswal, P., Naithani, S., Lera-Ramirez, M., Rutherford, K., Wood, V., De Pons, J.L., Dwinell, M.R., Hayman, G.T., Kaldunski, M.L., Kwitek, A.E., Lauledерkind, S.J.F., Tutaj, M.A., Vedi, M., Wang, S.-J., D'Eustachio, P., Aimo, L., Axelsen, K., Bridge, A., Hyka-Nouspikel, N., Morgat, A., Aleksander, S.A., Cherry, J.M., Engel, S.R., Karra, K., Miyasato, S.R., Nash, R.S., Skrzypek, M.S., Weng, S.,

Wong, E.D., Bakker, E., Berardini, T.Z., Reiser, L., Auchincloss, A., Axelsen, K., Argoud-Puy, G., Blatter, M.-C., Boutet, E., Breuza, L., Bridge, A., Casals-Casas, C., Coudert, E., Estreicher, A., Livia Famiglietti, M., Feuermann, M., Gos, A., Gruaz-Gumowski, N., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Le Mercier, P., Lieberherr, D., Masson, P., Morgat, A., Pedruzzi, I., Pourcel, L., Poux, S., Rivoire, C., Sundaram, S., Bateman, A., Bowler-Barnett, E., Bye-A-Jee, H., Denny, P., Ignatchenko, A., Ishtiaq, R., Lock, A., Lussi, Y., Magrane, M., Martin, M.J., Orchard, S., Raposo, P., Speretta, E., Tyagi, N., Warner, K., Zaru, R., Diehl, A.D., Lee, R., Chan, J., Diamantakis, S., Raciti, D., Zarowiecki, M., Fisher, M., James-Zorn, C., Ponferrada, V., Zorn, A., Ramachandran, S., Ruzicka, L., Westerfield, M., 2023. The Gene Ontology knowledgebase in 2023. Genetics 224, iyad031. https://doi.org/10.1093/genetics/iyad031

The UniProt Consortium, 2023. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 51, D523–D531. https://doi.org/10.1093/nar/gkac1052

Theissinger, K., Falckenhayn, C., Blande, D., Toljamo, A., Gutekunst, J., Makkonen, J., Jussila, J., Lyko, F., Schrimpf, A., Schulz, R., Kokko, H., 2016. De Novo assembly and annotation of the freshwater crayfish Astacus astacus transcriptome. Mar. Genomics 28, 7–10. https://doi.org/10.1016/j.margen.2016.02.006

Theissinger, K., Fernandes, C., Formenti, G., Bista, I., Berg, P.R., Bleidorn, C., Bombarely, A., Crottini, A., Gallo, G.R., Godoy, J.A., Jentoft, S., Malukiewicz, J., Mouton, A., Oomen, R.A., Paez, S., Palsbøll, P.J., Pampoulie, C., Ruiz-López, M.J., Secomandi, S., Svardal, H., Theofanopoulou, C., Vries, J. de, Waldvogel, A.-M., Zhang, G., Jarvis, E.D., Bálint, M., Ciofi, C., Waterhouse, R.M., Mazzoni, C.J., Höglund, J., Aghayan, S.A., Alioto, T.S., Almudi, I., Alvarez, N., Alves, P.C., Rosario, I.R.A. do, Antunes, A., Arribas, P., Baldrian, P., Bertorelle, G., Böhne, A., Bonisoli-Alquati, A., Boštjančić, L.L., Boussau, B., Breton, C.M., Buzan, E., Campos, P.F., Carreras, C., Castro, L.Fi.C., Chueca, L.J., Čiampor, F., Conti, E., Cook-Deegan, R., Croll, D., Cunha, M.V., Delsuc, F., Dennis, A.B., Dimitrov, D., Faria, R., Favre, A., Fedrigo, O.D., Fernández, R., Ficetola, G.F., Flot, J.-F., Gabaldón, T., Agius, D.R., Giani, A.M., Gilbert, M.T.P., Grebenc, T., Guschanski, K., Guyot, R., Hausdorf, B., Hawlitschek, O., Heintzman, P.D., Heinze, B., Hiller, M., Husemann, M., Iannucci, A., Irisarri, I., Jakobsen, K.S., Klinga, P., Kloch, A., Kratochwil, C.F., Kusche, H., Layton, K.K.S., Leonard, J.A., Lerat, E., Liti, G., Manousaki, T., Marques-Bonet, T., Matos-Maraví, P., Matschiner, M., Maumus, F., Cartney, A.M.M., Meiri, S., Melo-Ferreira, J., Mengual, X., Monaghan, M.T., Montagna, M., Mysłajek, R.W., Neiber, M.T., Nicolas, V., Novo, M., Ozretić, P., Palero, F., Pârvulescu, L., Pascual, M., Paulo, O.S., Pavlek, M., Pegueroles, C., Pellissier, L., Pesole, G., Primmer, C.R., Riesgo, A., Rüber, L., Rubolini, D., Salvi, D., Seehausen, O., Seidel, M., Studer, B., Theodoridis, S., Thines, M., Urban, L., Vasemägi, A., Vella, A., Vella, N., Vernes, S.C., Vernesi, C., Vieites, D.R., Wheat, C.W., Wörheide, G., Wurm, Y., Zammit, G., 2023. How genomics can help biodiversity conservation. Trends Genet. 39, 545–559. https://doi.org/10.1016/j.tig.2023.01.005

Thomas, G.W.C., Dohmen, E., Hughes, D.S.T., Murali, S.C., Poelchau, M., Glastad, K., Anstead, C.A., Ayoub, N.A., Batterham, P., Bellair, M., Binford, G.J., Chao, H., Chen, Y.H., Childers, C., Dinh, H., Doddapaneni, H.V., Duan, J.J., Dugan, S., Esposito, L.A., Friedrich, M., Garb, J., Gasser, R.B., Goodisman, M.A.D., Gundersen-Rindal, D.E., Han, Y., Handler, A.M., Hatakeyama, M., Hering, L., Hunter, W.B., Ioannidis, P., Jayaseelan, J.C., Kalra, D., Khila, A., Korhonen, P.K., Lee, C.E., Lee, S.L., Li, Y., Lindsey, A.R.I., Mayer, G., McGregor, A.P., McKenna, D.D., Misof, B., Munidasa, M., Munoz-Torres, M., Muzny, D.M., Niehuis, O.,

Osuji-Lacy, N., Palli, S.R., Panfilio, K.A., Pechmann, M., Perry, T., Peters, R.S., Poynton, H.C., Prpic, N.-M., Qu, J., Rotenberg, D., Schal, C., Schoville, S.D., Scully, E.D., Skinner, E., Sloan, D.B., Stouthamer, R., Strand, M.R., Szucsich, N.U., Wijeratne, A., Young, N.D., Zattara, E.E., Benoit, J.B., Zdobnov, E.M., Pfrender, M.E., Hackett, K.J., Werren, J.H., Worley, K.C., Gibbs, R.A., Chipman, A.D., Waterhouse, R.M., Bornberg-Bauer, E., Hahn, M.W., Richards, S., 2020. Gene content evolution in the arthropods. Genome Biol. 21, 15. https://doi.org/10.1186/s13059-019-1925-7

Tielens, A.G.M., Van Hellemond, J.J., 1998. The electron transport chain in anaerobically functioning eukaryotes. Biochim. Biophys. Acta BBA - Bioenerg., 10th European Bioenergetics Conference 1365, 71–78. https://doi.org/10.1016/S0005-2728(98)00045-0

Tørresen, O.K., Star, B., Mier, P., Andrade-Navarro, M.A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A.V., Promponas, V.J., Anisimova, M., Jakobsen, K.S., Linke, D., 2019. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. 47, 10994–11006. https://doi.org/10.1093/nar/gkz841

Treangen, T.J., Salzberg, S.L., 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13, 36–46. https://doi.org/10.1038/nrg3117

Tree of Life Web Project [WWW Document], n.d. URL http://tolweb.org/tree/ (accessed 9.21.23).

Uengwetwanit, T., Pootakham, W., Nookaew, I., Sonthirod, C., Angthong, P., Sittikankaew, K., Rungrassamee, W., Arayamethakorn, S., Wongsurawat, T., Jenjaroenpun, P., Sangsrakru, D., Leelatanawit, R., Khudet, J., Koehorst, J.J., Schaap, P.J., Martins dos Santos, V., Tangy, F., Karoonuthaisiri, N., 2021. A chromosome-level assembly of the black tiger shrimp (Penaeus monodon) genome facilitates the identification of growth-associated genes. Mol. Ecol. Resour. 21, 1620–1640. https://doi.org/10.1111/1755-0998.13357

Uffelmann, E., Huang, Q.Q., Munung, N.S., de Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T., Posthuma, D., 2021. Genome-wide association studies. Nat. Rev. Methods Primer 1, 1–21. https://doi.org/10.1038/s43586-021-00056-9

Unneberg, P., Larsson, M., Olsson, A., Wallerman, O., Petri, A., Bunikis, I., Vinnere Pettersson, O., Papetti, C., Gislason, A., Glenner, H., Cartes, J.E., Blanco-Bercial, L., Eriksen, E., Meyer, B., Wallberg, A., 2024. Ecological genomics in the Northern krill uncovers loci for local adaptation across ocean basins. Nat. Commun. 15, 1–29. https://doi.org/10.1038/s41467-024-50239-7

Usio, N., Townsend, C.R., 2004. Roles of Crayfish: Consequences of Predation and Bioturbation for Stream Invertebrates. Ecology 85, 807–822. https://doi.org/10.1890/02-0618

van Dijk, E.L., Auger, H., Jaszczyszyn, Y., Thermes, C., 2014. Ten years of next-generation sequencing technology. Trends Genet. 30, 418–426. https://doi.org/10.1016/j.tig.2014.07.001

VanHook, A.M., Patel, N.H., 2008. Crustaceans. Curr. Biol. CB 18, R547-550. https://doi.org/10.1016/j.cub.2008.05.021

Veldsman, W.P., Ma, K.Y., Hui, J.H.L., Chan, T.F., Baeza, J.A., Qin, J., Chu, K.H., 2021. Comparative genomics of the coconut crab and other decapod crustaceans: exploring the molecular basis of terrestrial adaptation. BMC Genomics 22, 1–15. https://doi.org/10.1186/s12864-021-07636-9

Verma, A.K., 2017. Genetic Diversity as Buffer in Biodiversity. Indian J. Biol. 4, 61–63. https://doi.org/10.21088/ijb.2394.1391.4117.9

Vigneux, E., 1997. Les introductions de crustacés décapodes d'eau douce en France. Peut-on parler de gestion ? Bull. Fr. Pêche Piscic. 357–370. https://doi.org/10.1051/kmae:1997035

Vitales, D., Garcia, S., Dodsworth, S., 2020. Reconstructing phylogenetic relationships based on repeat sequence similarities. Mol. Phylogenet. Evol. 147, 106766. https://doi.org/10.1016/j.ympev.2020.106766

Voelkerding, K.V., Dames, S.A., Durtschi, J.D., 2009. Next-Generation Sequencing: From Basic Research to Diagnostics. Clin. Chem. 55, 641–658. https://doi.org/10.1373/clinchem.2008.112789

von Reumont, B.M., Jenner, R.A., Wills, M.A., Dell'Ampio, E., Pass, G., Ebersberger, I., Meyer, B., Koenemann, S., Iliffe, T.M., Stamatakis, A., Niehuis, O., Meusemann, K., Misof, B., 2012. Pancrustacean Phylogeny in the Light of New Phylogenomic Data: Support for Remipedia as the Possible Sister Group of Hexapoda. Mol. Biol. Evol. 29, 1031–1045. https://doi.org/10.1093/molbev/msr270

von Thaden, A., Nowak, C., Tiesmeyer, A., Reiners, T.E., Alves, P.C., Lyons, L.A., Mattucci, F., Randi, E., Cragnolini, M., Galián, J., Hegyeli, Z., Kitchener, A.C., Lambinet, C., Lucas, J.M., Mölich, T., Ramos, L., Schockert, V., Cocchiararo, B., 2020. Applying genomic data in wildlife monitoring: development guidelines for genotyping degraded samples with reduced single nucleotide polymorphism (SNP) panels. Mol. Ecol. Resour. 20, 10.1111/1755-0998.13136. https://doi.org/10.1111/1755-0998.13136

Wajid, B., Serpedin, E., 2016. Do it yourself guide to genome assembly. Brief. Funct. Genomics 15, 1–9. https://doi.org/10.1093/bfgp/elu042

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., Earl, A.M., 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE 9, e112963. https://doi.org/10.1371/journal.pone.0112963

Waltling, L., Thiel, M., 2013. Functional Morphology and Diversity, The Natural History of the Crustacea. Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780195398038.001.0001

Wang, J., Yang, W., Zhang, S., Hu, H., Yuan, Y., Dong, J., Chen, L., Ma, Y., Yang, T., Zhou, L., Chen, J., Liu, B., Li, C., Edwards, D., Zhao, J., 2023. A pangenome analysis pipeline provides insights into functional gene identification in rice. Genome Biol. 24, 19. https://doi.org/10.1186/s13059-023-02861-9

Wang, J., Chen, X., Hou, X., Wang, Jingan, Yue, W., Huang, S., Xu, G., Yan, J., Lu, G., Hofreiter, M., Li, C., Wang, C., 2022. "Omics" data unveil early molecular response underlying limb regeneration in the Chinese mitten crab, Eriocheir sinensis. Sci. Adv. 8, eabl4642. https://doi.org/10.1126/sciadv.abl4642

Wang, P., Meng, F., Moore, B.M., Shiu, S.-H., 2021. Impact of short-read sequencing on the misassembly of a plant genome. BMC Genomics 22, 99. https://doi.org/10.1186/s12864-021-07397-5

Wang, Q., Ren, X., Liu, P., Li, Jitao, Lv, J., Wang, J., Zhang, H., Wei, W., Zhou, Y., He, Y., Li, Jian, 2022. Improved genome assembly of Chinese shrimp (Fenneropenaeus chinensis) suggests adaptation to the environment during evolution and domestication. Mol. Ecol. Resour. 22, 334–344. https://doi.org/10.1111/1755-0998.13463

Wang, Q., Ren, X., Liu, P., Li, Jitao, Lv, J., Wang, J., Zhang, H., Wei, W., Zhou, Y., He, Y., Li, Jian, 2021. High-quality genome assembly of Chinese shrimp (Fenneropenaeus chinensis) suggests genome contraction and adaptation to the environment (preprint). Preprints. https://doi.org/10.22541/au.161798993.31069269/v1

Wang, X., Zhang, B., 2014. Integrating Genomic, Transcriptomic, and Interactome Data to Improve Peptide and Protein Identification in Shotgun Proteomics. J. Proteome Res. 13, 2715–2723. https://doi.org/10.1021/pr500194t

Wanichthanarak, K., Fahrmann, J.F., Grapov, D., 2015. Genomic, Proteomic, and Metabolomic Data Integration Strategies. Biomark. Insights 10s4, BMI.S29511. https://doi.org/10.4137/BMI.S29511

Waskom, M.L., 2021. seaborn: statistical data visualization. J. Open Source Softw. 6, 3021. https://doi.org/10.21105/joss.03021

Watson-Zink, V.M., 2021. Making the grade: Physiological adaptations to terrestrial environments in decapod crabs. Arthropod Struct. Dev. 64, 101089. https://doi.org/10.1016/j.asd.2021.101089

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A.M., Schatz, M.C., Myers, G., DePristo, M.A., Ruan, J., Marschall, T., Sedlazeck, F.J., Zook, J.M., Li, H., Koren, S., Carroll, A., Rank, D.R., Hunkapiller, M.W., 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. 37, 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., Schulman, A.H., 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8, 973–982. https://doi.org/10.1038/nrg2165

Wingett, S.W., Andrews, S., 2018. FastQ Screen: A tool for multi-genome mapping and quality control. F1000Research 7, 1338. https://doi.org/10.12688/f1000research.15931.2

Wolfe, J.M., Breinholt, J.W., Crandall, K.A., Lemmon, A.R., Lemmon, E.M., Timm, L.E., Siddall, M.E., Bracken-Grissom, H.D., 2019. A phylogenomic framework, evolutionary timeline and genomic resources for comparative studies of decapod crustaceans. Proc. R. Soc. B Biol. Sci. 286, 20190079. https://doi.org/10.1098/rspb.2019.0079

World Health Organization, 2020. Genomics [WWW Document]. URL https://www.who.int/news-room/questions-and-answers/item/genomics (accessed 9.21.23).

Wu, C., Lu, J., 2019. Diversification of Transposable Elements in Arthropods and Its Impact on Genome Evolution. Genes 10, 338. https://doi.org/10.3390/genes10050338

Xu, Z., Gao, T., Xu, Y., Li, X., Li, J., Lin, H., Yan, W., Pan, J., Tang, J., 2021. A chromosome-level reference genome of red swamp crayfish Procambarus clarkii provides insights into the gene families regarding growth or development in crustaceans. Genomics 113, 3274–3284. https://doi.org/10.1016/j.ygeno.2021.07.017

Xue, W., Li, J.-T., Zhu, Y.-P., Hou, G.-Y., Kong, X.-F., Kuang, Y.-Y., Sun, X.-W., 2013. L_RNA_scaffolder: scaffolding genomes with transcripts. BMC Genomics 14, 604. https://doi.org/10.1186/1471-2164-14-604

Yandell, M., Ence, D., 2012. A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. 13, 329–342. https://doi.org/10.1038/nrg3174

Yang, G., Nagel, D.H., Feschotte, C., Hancock, C.N., Wessler, S.R., 2009. Tuned for Transposition: Molecular Determinants Underlying the Hyperactivity of a Stowaway MITE. Science 325, 1391–1394. https://doi.org/10.1126/science.1175688

Ye, C., Hill, C.M., Wu, S., Ruan, J., Ma, Z. (Sam), 2016. DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. Sci. Rep. 6, 31900. https://doi.org/10.1038/srep31900

Young, A.M., Elliott, J.A., 2020. Life History and Population Dynamics of Green Crabs (Carcinus maenas). Fishes 5, 4. https://doi.org/10.3390/fishes5010004

Yuan, H., Liu, X.-J., Liu, X.-Z., Zhao, L.-N., Mao, S.-L., Huang, Y., 2024. The evolutionary dynamics of genome sizes and repetitive elements in Ensifera (Insecta: Orthoptera). BMC Genomics 25, 1041. https://doi.org/10.1186/s12864-024-10949-0

Yuan, J., Gao, Y., Zhang, X., Wei, J., Liu, C., Li, F., Xiang, J., 2017. Genome Sequences of Marine Shrimp Exopalaemon carinicauda Holthuis Provide Insights into Genome Size Evolution of Caridea. Mar. Drugs 15, 213. https://doi.org/10.3390/md15070213

Yuan, J., Yu, Y., Zhang, X., Li, S., Xiang, J., Li, F., 2023. Recent advances in crustacean genomics and their potential application in aquaculture. Rev. Aquac. 15, 1501–1521. https://doi.org/10.1111/raq.12791

Yuan, J., Zhang, X., Li, F., Xiang, J., 2021a. Genome Sequencing and Assembly Strategies and a Comparative Analysis of the Genomic Characteristics in Penaeid Shrimp Species. Front. Genet. 12. https://doi.org/10.3389/fgene.2021.658619

Yuan, J., Zhang, Xiaojun, Wang, M., Sun, Y., Liu, C., Li, S., Yu, Y., Gao, Y., Liu, F., Zhang, Xiaoxi, Kong, J., Fan, G., Zhang, C., Feng, L., Xiang, J., Li, F., 2021b. Simple sequence repeats drive genome plasticity and promote adaptive evolution in penaeid shrimp. Commun. Biol. 4, 1–14. https://doi.org/10.1038/s42003-021-01716-y

Zhang, B., Kuster, B., 2019. Proteomics Is Not an Island: Multi-omics Integration Is the Key to Understanding Biological Systems. Mol. Cell. Proteomics 18, S1–S4. https://doi.org/10.1074/mcp.E119.001693

Zhang, Q.-L., Xu, B., Wang, X.-Q., Yuan, M.-L., Chen, J.-Y., 2017. Genome-wide comparison of the protein-coding repertoire reveals fast evolution of immune-related genes in cephalochordates and Osteichthyes superclass. Oncotarget 9, 83–95. https://doi.org/10.18632/oncotarget.22749

Zhang, R.-G., Li, G.-Y., Wang, X.-L., Dainat, J., Wang, Z.-X., Ou, S., Ma, Y., 2022. TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. Hortic. Res. 9, uhac017. https://doi.org/10.1093/hr/uhac017

Zhang, Xiaojun, Yuan, J., Sun, Y., Li, S., Gao, Y., Yu, Y., Liu, C., Wang, Q., Lv, X., Zhang, Xiaoxi, Ma, K.Y., Wang, X., Lin, W., Wang, Long, Zhu, X., Zhang, C., Zhang, J., Jin, S., Yu, K., Kong, J., Xu, P., Chen, J., Zhang, H., Sorgeloos, P., Sagi, A., Alcivar-Warren, A., Liu, Z., Wang, Lei, Ruan, J., Chu, K.H., Liu, B., Li, F., Xiang, J., 2019. Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. Nat. Commun. 10, 356. https://doi.org/10.1038/s41467-018-08197-4

Zhang, Z.-Q., 2011. Phylum Arthropoda von Siebold, 1848 In: Zhang, Z.-Q. (Ed.) Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. Zootaxa 3148, 99. https://doi.org/10.11646/zootaxa.3148.1.14

Zhong, Y., Zhao, W., Tang, Z., Huang, L., Zhu, X., Liang, X., Yan, A., Lu, Zhifa, Yu, Y., Tang, D., Wang, D., Lu, Zhuanling, 2021. Comparative transcriptomic analysis of the different developmental stages of ovary in red swamp crayfish Procambarus clarkii. BMC Genomics 22, 199. https://doi.org/10.1186/s12864-021-07537-x

Zhou, S., Herschleb, J., Schwartz, D.C., 2007. Chapter 9 A Single Molecule System for Whole Genome Analysis, in: Mitchelson, K.R. (Ed.), Perspectives in Bioanalysis, New High Throughput Technologies for DNA Sequencing and Genomics. Elsevier, pp. 265–300. https://doi.org/10.1016/S1871-0069(06)02009-X

Zhou, Z., Mo, L., Li, D., Zeng, W., Wu, H., Wu, Z., Huang, J., 2022. Comparative transcriptomics analyses of chemosensory genes of antenna in male red swamp crayfish Procambarus clarkii. Front. Ecol. Evol. 10. https://doi.org/10.3389/fevo.2022.976448

Zhu, B., Pennack, J.A., McQuilton, P., Forero, M.G., Mizuguchi, K., Sutcliffe, B., Gu, C.-J., Fenton, J.C., Hidalgo, A., 2008. Drosophila neurotrophins reveal a common mechanism for nervous system formation. PLoS Biol. 6, e284. https://doi.org/10.1371/journal.pbio.0060284

Zimin, A.V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L., Yorke, J.A., 2013. The MaSuRCA genome assembler. Bioinforma. Oxf. Engl. 29, 2669–2677. https://doi.org/10.1093/bioinformatics/btt476

Zrzavý, J., Štys, P., 1997. The basic body plan of arthropods: insights from evolutionary morphology and developmental biology. J. Evol. Biol. 10, 353–367. https://doi.org/10.1046/j.1420-9101.1997.10030353.x

Annexes

Supplementary 1: The MetaInvert soil invertebrate genome resource provides insights into below-ground biodiversity and evolution. (Collins et al., 2023)

communications biology

ARTICLE

https://doi.org/10.1038/s42003-023-05621-4 OPEN

The MetaInvert soil invertebrate genome resource provides insights into below-ground biodiversity and evolution

Gemma Collins[1,2], Clément Schneider[2,3], Ljudevit Luka Boštjančić[1,4,5], Ulrich Burkhardt[6], Axel Christian[3], Peter Decker[3], Ingo Ebersberger[1,2,7], Karin Hohberg[3], Odile Lecompte[4], Dominik Merges[8], Hannah Muelbaier[7], Juliane Romahn[1,2], Jörg Römbke[9], Christelle Rutz[4], Rüdiger Schmelz[10], Alexandra Schmidt[1,11], Kathrin Theissinger[1,2,5], Robert Veres[1,12], Ricarda Lehmitz[3], Markus Pfenninger[1,2,13] & Miklós Bálint[1,2,14]✉

Soil invertebrates are among the least understood metazoans on Earth. Thus far, the lack of taxonomically broad and dense genomic resources has made it hard to thoroughly investigate their evolution and ecology. With MetaInvert we provide draft genome assemblies for 232 soil invertebrate species, representing 14 common groups and 94 families. We show that this data substantially extends the taxonomic scope of DNA- or RNA-based taxonomic identification. Moreover, we confirm that theories of genome evolution cannot be generalised across evolutionarily distinct invertebrate groups. The soil invertebrate genomes presented here will support the management of soil biodiversity through molecular monitoring of community composition and function, and the discovery of evolutionary adaptations to the challenges of soil conditions.

[1] Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany. [2] LOEWE Centre for Translational Biodiversity Genomics, Frankfurt am Main, Germany. [3] Soil Zoology, Senckenberg Museum of Natural History, Görlitz, Germany. [4] Department of Computer Science, ICube, UMR 7357, University of Strasbourg, CNRS, Centre de Recherche en Biomédecine de Strasbourg, Strasbourg, France. [5] Department of Molecular Ecology, Institute for Environmental Sciences, Rhineland-Palatinate Technical University Kaiserslautern Landau, Landau, Germany. [6] Soil Organism Research, Görlitz, Germany. [7] Institute of Cell Biology and Neuroscience, Goethe University, Frankfurt am Main, Germany. [8] Department of Forest Mycology and Plant Pathology, Swedish University of Agricultural Sciences, Uppsala, Sweden. [9] ECT Oekotoxikologie GmbH, Flörsheim, Germany. [10] Freelance Biologist, A Coruña, Spain. [11] Limnological Institute, University of Konstanz, Konstanz, Germany. [12] Institute of Biology and Geology, Babeș-Bolyai University, Cluj-Napoca, Romania. [13] Johannes Gutenberg University, Mainz, Germany. [14] Department of Insect Biotechnology, Justus-Liebig University, Gießen, Germany. ✉email: miklos.balint@senckenberg.de

Soils and soil biodiversity are becoming increasingly valued and protected at the policy level[1]. Soil invertebrates are major components of soil biodiversity, and their activity is important for almost all soil ecosystem services[2]. For example, soil invertebrates are responsible for up to 50% of the litter decomposition[3]. They contribute to functional services crucial to humans, such as nutrient cycling, water storage and support above-ground food production through the integration of nutrients in food webs[4–6]. Furthermore, soil invertebrates play major roles in regulating microbial activity along the plant-soil continuum[7]. Consistent with their importance in soil ecosystems, they are actively promoted in soil biodiversity conservation frameworks[8].

However, soil invertebrates are inherently difficult to study morphologically due to their incredible diversity, huge abundances, and small body size with microscopic morphological details. Though generally tiny, they show a ~100-fold variation in body weight, which ranges from nanograms to grams[9]. There are potentially hundreds of thousands of undescribed species globally[10]. Moreover, taxonomic expertise is declining[11] and this is particularly problematic for groups where experts have always been rare.

DNA- and RNA-based methods are long promoted to support traditional taxonomy and ecological studies in difficult organism groups. Shotgun metagenomics randomly sequences DNA fragments from a sample, instead of relying on PCR-amplified taxonomic marker genes. Metagenomics is an increasingly feasible approach to record the presence of higher eukaryotes in a diverse range of samples[12–14]. Since metagenomics can utilise all genomic information for taxonomic identification, it has improved sensitivity and specificity compared to metabarcoding[15], and it promises superior quantification of species' biomass[16]. Metatranscriptomics in turn records genes which are actively transcribed into RNA and thus drive ongoing biological processes[17], informing about the metabolic activity of soil community members[18], and functional changes in these communities[17].

Comprehensive genome collections are the backbone for metagenomics and metatranscriptomics. If genome databases are available, shotgun metagenomics and metatranscriptomics have shown to provide unprecedented insights[17,19], e.g., into vegetation change over glacial cycles[15], historic population genomic processes[20,21], and kingdom-spanning processes of ecosystem functioning[22]. Large genome sequencing initiatives like the Earth Biogenome Project[23] will provide this data ultimately, but progress so far mainly focused on large, prominent organisms, such as mammals[24], birds[25], insects[26] and plants[15]. In addition to serving taxonomic identification, broad (many distinct groups) and dense (many species from a group) sequencing of genomes additionally allows identifying common patterns of gene evolution and test the taxonomic generality of hypotheses on genome evolution.

Results and discussion

A genome resource for soil invertebrates. Here, we have generated a large genomic resource to support insights into the structure, activity and functioning of soil invertebrate communities (Fig. 1). We had two aims. First, we wanted to provide a large number of soil invertebrate genomes to aid species identifications through metagenomics or metatranscriptomics. Second, we intended to explore patterns of genome evolution across taxa, which needs a taxonomically broad and dense sampling of genomes. We sequenced and assembled the genomes of 232 species, representing 14 common soil invertebrate groups (hereafter referred to as "groups") encompassing 94 families, most of which were lacking whole-genome data thus far (Fig. 2, Table 1), including Collembola ($n = 87$ species), Oribatida ($n = 62$), two

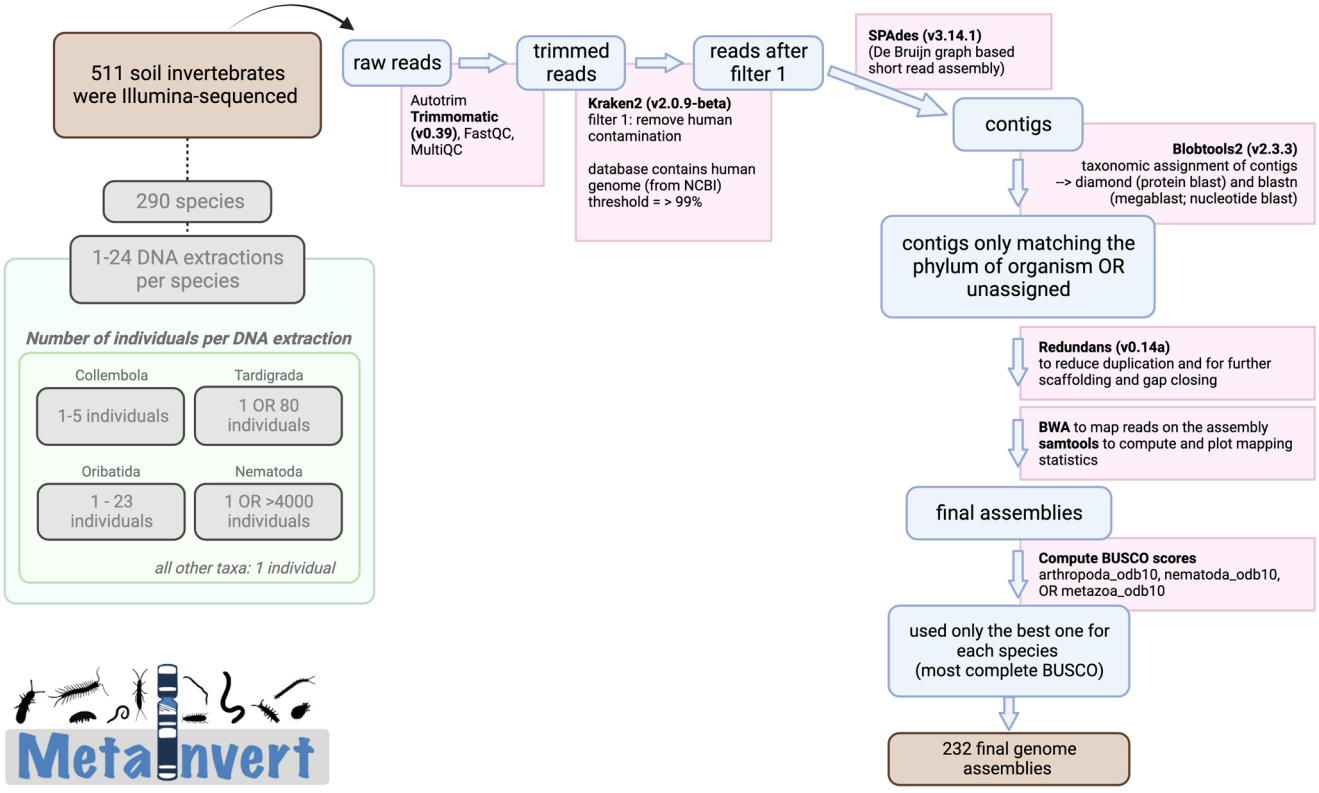

Fig. 1 Overview of the bioinformatic pipeline for genome assembly and quality control. The genome assembly pipeline consists of a read quality filtering step, short read assembly and several steps for removing non-target DNA reads, co-sequenced along the genomes of the targeted species. The MetaInvert logo was created by the first author. Animal silhouettes originate from phylopic.org, and they can be reused under Creative Common licences.

classes of Myriapoda ($n = 23$ Diplopoda; $n = 19$ Chilopoda) and Nematoda ($n = 18$). Genome completeness estimated with benchmarking universal single copy orthologs (BUSCO)[27] was 59.78% on average (median: 69.2%), with an average contig N50 of 6080 bases (median 4039), and with an average L50 of 28,375 (median: 11503, Supplementary Data 1).

Improved taxonomic assignment of metazoan environmental sequence data. To demonstrate the relevance of this genomic resource, we first used the 232 genomes to improve the taxonomic assignment metatranscriptomic sequences generated from a 2-year sampling of soil environmental RNA (eRNA) along an elevational gradient[28]. Such assignments of soil eRNA were previously limited in scope due to a general lack of soil invertebrate genome data. Briefly, we assigned eRNA reads with bacterial, fungal, plant and soil invertebrate genomes, with and without including the MetaInvert genomes presented here. We found that about 2.45% (854,409 reads) of the classified metatranscriptomic reads (40,265,768) could be assigned to soil invertebrates, in comparison to bacteria (77,1%, 31,063,088), fungi (20.1%, 8,078,679), and plants (0.33%, 134,852)[29]. Previous metatranscriptomic studies reported a similar microbial eukaryote to bacteria ratio[29,30]. The inclusion of the MetaInvert genomes significantly increased reads assigned to soil invertebrates (Kruskal-Wallis $X^2 = 9.14$, $df = 1$, $p = 0.002$, Fig. 3a). We recorded 11 soil invertebrate classes (Fig. 3b), of which the most abundant were nematodes of the class Chromadorea followed by clitellates (comprising both earthworms and enchytraeids). Linear regression showed a marked dip in soil invertebrate richness along the elevation gradient (ANOVA, $F_{elevation} = 0.22$, $p_{elevation} = 0.65$, $F_{elevation\text{^}2} = 9.1$, $p_{elevation\text{^}2} = 0.02$, Fig. 3c). This is in contrast with findings of hump-shaped elevation - richness relationships in soil invertebrates[31]. The pattern observed by us might be driven by distinct vegetation covers, although the confirmation of this needs a better sampling resolution. The community composition of soil invertebrates showed no statistically significant changes along the elevation gradient (analysis of deviance of multivariate generalised linear model fits, $df = 8$, $dev = 434.60$, $p = 0.13$), marginally significant differences across habitats study years ($df = 65$, $dev = 806.03929.87$, $p = 0.085$), and statistically significant differences between the two study years ($df = 5$, $dev = 1066.09$, $p = 0.04$).

No change in community composition along the elevation gradient is consistent with observed high abundances of soil invertebrates at high altitudes[30]. Differences in vegetation are known to influence soil invertebrate community composition, although our analysis may lack power to equivocally detect these. Differences in community composition between the study years may reflect year-specific environmental differences. However, we caution not to over interpret these results. The power of an analysis of drivers of community composition and richness on this gradient should be increased with more extensive sampling. The analyses nonetheless demonstrate the value of a dedicated soil invertebrate genome database for the identification of shotgun-sequenced environmental nucleotide samples from soils.

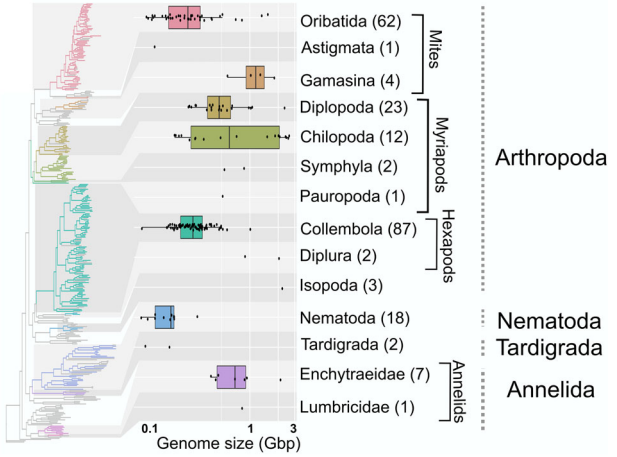

Fig. 2 Maximum likelihood phylogenetic tree of soil invertebrate genomes. The tree is based on an alignment of 141 metazoan BUSCO genes of the 232 soil invertebrates sequenced in this work (coloured branches), and 118 NCBI RefSeq (grey branches), representing four phyla. A high-resolution, annotated version of the tree is available as Supplementary Fig. 1. A more detailed tree and the alignment are available on FigShare[30]. Boxplots reflect the genome size distribution of the taxa subsumed in the corresponding clades in gigabases (Gb). Numbers of sequenced genomes with genome size estimates are indicated for each group. Genome size estimation was not possible for some of the assemblies. Genome size estimates can be found in Supplementary Data 1. Center line: median; box limits: upper and lower quartiles; whiskers: 1.5× interquartile range; points: outliers.

Table 1 Overview of 232 soil invertebrate genome assemblies.

Phylum	Taxon group [rank]	Common name	n known species (soil or terrrestrial)	n species (published genomes)	n species (genomes contributed here)
Annelida	Lumbricidae [family]	Earthworms	7000	2	1
Annelida	Enchytraeidae [family]	Potworms	700	1	7
Nematoda	Nematoda [phylum]	Nematodes	25000	73	18
Tardigrada	Tardigrada [phylum]	Tardigrades	1150	4	2
Arthropoda	Gamasina [infraorder]	Predatory mites	40000	1	4
Arthropoda	Astigmata [suborder]	Mites [not soil]		7	1
Arthropoda	Oribatida [suborder]	Box mites		7	62
Arthropoda	Chilopoda [class]	Centipedes	3000	2	19
Arthropoda	Diplopoda [class]	Millipedes	12000	3	23
Arthropoda	Symphyla [class]	Symphylans	200	0	2
Arthropoda	Pauropoda [class]	Pauropods	800	0	1
Arthropoda	Isopoda [order]	Pill bugs	3637	5	3
Arthropoda	Diplura [order]	Diplurans	1000	2	2
Arthropoda	Collembola [class]	Springtails	8500	35	87

For each taxonomic group we also list the number of species with publicly available genome assemblies (as of June 2022).

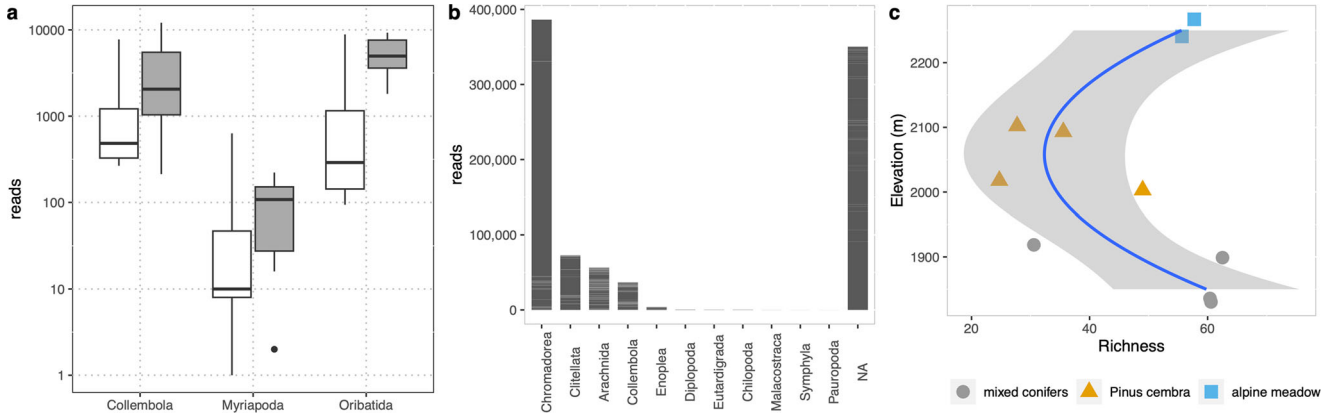

Fig. 3 Taxonomic assignments of soil metatranscriptomes using soil invertebrate genomes. a Assignment success of soil metatranscriptomic reads using genomes available in NCBI RefSeq (white) and MetaInvert genomes in addition to NCBI genomes (grey). Please note the log-scale of the y-axis (taxon observations with RefSeq genomes: Collembola $n = 290$, Myriapoda $n = 80$, Oribatida $n = 90$; independent taxon observations with RefSeq + MetaInvert genomes: Collembola $n = 810$, Myriapoda $n = 270$, Oribatida $n = 630$), center line: median; box limits: upper and lower quartiles; whiskers: 1.5× interquartile range; points: outliers; **b** reads assigned to common soil invertebrate classes, with NA marking metazoan reads not assigned to soil invertebrates at the class level; **c** soil invertebrate richness trend along an elevation gradient (grey area marks standard error of the trendline). Assignments are available as Supplementary Data 2.

Insights into genome size evolution. As a second example, we addressed hypotheses concerning genome size evolution. We estimated the genome size for 191 species using the assembly-based approach ModEst[31]. We found a 30-fold range of genome sizes across the groups (Fig. 2), from 79 Mb (the nematode *Discolaimus major*) to 2.9 Gb (the chilopod *Lithobius crassipesoides*). Nematoda and Tardigrada had typically small genomes, whereas the genomes of Enchytraeidae were remarkably larger. In addition to between-group variation, some groups also had a wide range of genome sizes among member species. For example, Chilopoda (centipedes) genomes ranged in size from 0.178 to 2.90 Gb, while Oribatida genomes ranged from 0.09 to 1.72 Gb. Repeat content and GC content also varied widely both within and between soil invertebrate groups (Supplementary Fig. 2).

Classic theory predicts that a few basic factors, in particular effective population size, should lead to causal relationships between genome properties and functional traits (Fig. 4a)[32–34]. However, recent studies have shown that taxon-specific processes might be more important for genome size than demography[35,36]. We used our taxonomically broad data set to test the classical hypothesis of a few factors generally influencing genome size evolution vs. a more lineage-specific view with a series of structural equation models (SEMs, Fig. 4). We used genomes with at least 50% BUSCO completeness and 8× mode coverage. To parametrize the SEM and connect the 143 new genome assemblies with ecological traits, we first gathered trait data from original literature. Information about habitat preferences was added from the Edaphobase data warehouse for soil biodiversity (https://portal.edaphobase.org/). We focussed on three traits: (a) body length as a proxy for body size (minimum female adult body length for nematodes, and mean adult body length for all other taxa), (b) reproduction mode, and c) the number of known habitat types where a species occurs, as a proxy of habitat generality (based on CORINE—Coordination of Information on the Environment[37]). We annotated repetitive elements with species-specific repeat libraries. We estimated effective population size (theta) directly from the genome data by making use of the genome-wide heterozygosity in the reference individual. This proxy measure of effective population size was calculated individually for each genome assembly with at least 8X coverage. Genomic and ecological traits are accessible in Supplementary Data 1.

The variables tested have complex interactions that need to be modelled in the SEMs. Effective population size should be influenced by habitat generalism, with the expectation that species able to thrive in a wide range of habitats should have larger population sizes and therefore also larger effective population sizes (N_e)[38]. N_e should be inversely related to body size, as larger populations of small-bodied organisms can be maintained by the same amount of resources in comparison to large-bodied species[32]. The reproductive mode is known to impact N_e, because the higher the degree of inbreeding, the smaller the expected N_e[39]. High N_e is frequently hypothesised to contribute to reducing repeats as evolutionary burdens from genomes, as selection is more efficient in larger populations[33,34]. Repeats are frequently considered to increase genome size[40,41]. If the repeats themselves are biassed in base composition, this should reflect in the overall GC content. Interestingly, GC content is also linked to resource availability[42], which may be linked to habitat generalism via higher metabolic flexibility[43]. Even though most of these observations originate from bacterial studies, ample evidence exists that the environment may influence base composition also in metazoans[42,44–46].

When modelling all soil invertebrate groups together, most hypothesised causal relationships were either statistically insignificant or pointed to the opposite directions than classical theory predicted (Supplementary Fig. 3). Most strikingly, high N_e size was linked to higher repeat content which in turn implies larger genome size. This suggests that efficient selection does not universally reduce the evolutionary burden of large genomes and repeat content[47]. The SEMs supported only two of the hypothesised causal relationships when these were modelled for all taxa together (Fig. 4b): a positive link between repeat content and genome size, and a negative link between repeat content and GC content. Genome size is frequently considered to be driven by repeat content[48,49], but with variation in the relationship among higher taxa of vertebrates[50]. Such variation might be due to epigenetic regulation via repetitive elements, maintenance of chromosome structure[51], and modification of gene expression and transcript diversification[52]. Higher GC content is linked to smaller genome size in many but not all eukaryotic groups[49]. This link might also originate from the expansion of repeats with low GC content.

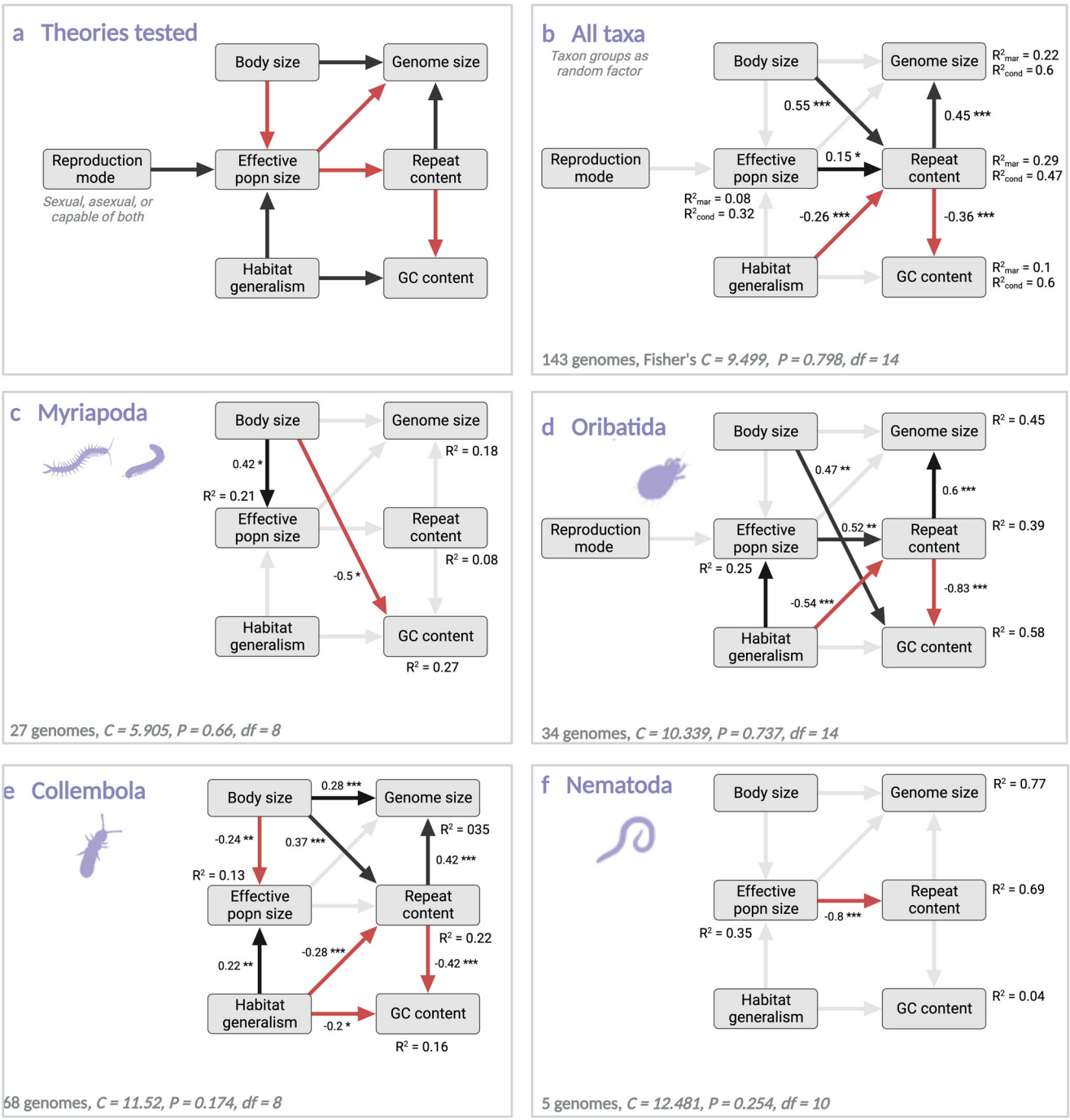

Fig. 4 Structural equation models (SEMs) of hypothesised causal relationships among genomic traits and their ecological drivers. a Initial SEM with hypothesised links; **b–f** SEMs fitted to all taxa, and to major taxonomic groups. Arrows indicate hypothesised or modelled relationships, positive (black) or negative (red). Links marked with grey arrows were not statistically significant in the SEM. Fisher's *C* evaluates conditional independence claims among nodes and indicates model fit, with *p*-values showing whether discrepancies between the model and the data are statistically significant. Degrees of freedom are marked with *df*. Values next to arrows show standardised estimates, with asterisk indicating the statistical significance of the relationship (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). Animal silhouettes originate from phylopic.org, and they can be reused under Creative Common licences.

Our results confirm that the strength and direction of relationships among genome size, repeat content, GC content, and their ecological drivers vary among higher taxa of invertebrates (Supplementary Fig. 3). SEMs fitted separately to higher taxa (myriapods, oribatid mites, springtails, nematodes) showed marked group-specific differences in the support of causal hypotheses between genomic and ecological traits (Fig. 4c–f, Supplementary Fig. 3). The assumed positive link between body size and genome size[32] received statistical support only in Collembola, but with an opposite sign as predicted by the nucleotypic theory[32]. The effects of body size on genomes are often difficult to disentangle from other co-variables[53,54]. This indicates lineage-specific expansion or contraction of genomes, reported for diverse eukaryotes[50,55]. The expected negative relationship between N_e and repeat content[40,41] was confirmed only in nematodes. However, the relationship was positive in oribatid mites, and missing altogether in the other taxa. Habitat generalism was positively linked to effective population size only

in Collembola and oribatids, but not in myriapods and nematodes. This suggests that generalists may not be as fit as specialists in any particular habitats[56,57], and their evolution might depend on differential rates of population evolution compared to rates of environmental change[58]. Interestingly, models of oribatids and Collembola suggested that higher habitat generality might be linked to lower repeat contents. Altogether, our analysis supports a more nuanced, lineage-specific view of factors driving genome size evolution rather than the classical view of only a few general factors governing the C-value enigma.

Gene loss patterns in springtails and oribatid mites. As a third example, we explored whether shared gene loss might be related to repeated adaptations of phylogenetically distant metazoans to soil conditions. Gene loss is a key process in evolution[59,60]. Here, the dense taxon sampling for individual groups allowed to differentiate between consistent gene absence across several taxa, which likely indicates gene loss, and the sporadic absence of a gene in individual taxa, which likely represents noise introduced by assembly incompleteness. To further reduce the risk that incomplete gene annotations generate a spurious signal of gene loss, we used a targeted search for orthologs in the un-annotated genome assemblies to determine the presence/absence patterns of genes across taxa. We analysed the presence of 1482 core metazoan gene orthologs. Notably, this revealed that 50 core genes are missing in springtails ($n = 78$ species), and 97 core genes were not found in the oribatid mites ($n = 54$ species) (Fig. 5). Given the large number of investigated taxa in the two groups, it is unlikely that these genes have been accidentally missed. Instead, their absence indicates gene losses early during diversification of the respective groups, similar to what has been seen for other animals[61]. Overall, fifteen gene ontology terms were significantly enriched (testFisher < 0.05) among the genes lost involving biological processes such as tubulin metabolism and cellular and subcellular movement (Oribatida). There was a significant loss of genes involved in pyridine-containing compound metabolic processes in springtails (Fig. 5; Supplementary Data 3, 4). Pyridine-containing molecules have a considerable spectrum of antimicrobial and antiviral activities[62], and associated gene loss might be related to the gain of endogenous antibiotic synthesis ability by many springtail species[63]. We also manually screened the UniProt database (accessed on 28.6.2023) for putative gene functions associated with genes missing from Collembola and Oribatida assemblies. We aimed to identify functional or other relevant commonalities among the genes which might be missed by an algorithmic GO enrichment analysis. We could not detect patterns in gene functions. It was noteworthy that all existing annotations originated from only two species: *Drosophila melanogaster* or *Strigamia maritima*. This highlights the general difficulties with transferring annotations gained from a few model taxa to the breadth of biodiversity, with targeted annotation of specific genes being a solution.

In summary, our large collection of soil invertebrate genomes is a first major step towards a comprehensive DNA- or RNA-based identification of the entire soil biodiversity: they extend the scope of metagenomic or metatranscriptomic studies from microorganisms to metazoans. An important limitation of the study is the quality of the genomes, which precludes deeper analyses, such as structural comparisons. Genome quality is currently restrained by the qualitative and quantitative requirements of the current sequencing techniques with respect to genomic DNA. Although it is already possible to generate highly contiguous and complete genomes of soil invertebrates from single specimens[64], the minute amounts of genomic DNA (often fragmented because of field preservation) does not yet allow for the generation of better quality genomes on scale. Nonetheless, the genomes are of sufficiently high contiguity or completeness to considerably improve metagenomic and metatranscriptomic sequence assignments[65]. Further, the taxonomically broad and dense sampling of genomes provides unique insights into genome evolution, although clearly not into structural differences. Here we could show that no single theory of genome evolution fits all taxa: there are probably no simple overarching explanations for observed variations in genome properties, but interactions of multiple drivers result in divergent genome evolution patterns in different groups, reflecting their unique evolutionary history. Broad genome sampling allows for the identification of group-specific gene loss patterns, highlighting issues and future directions around the functional annotation of genomes from non-model taxa in diverse habitats. Overall, the 232 soil invertebrate genomes demonstrate the importance of genome sequencing efforts for understanding the ecology and evolution of the full scale of eukaryotic biodiversity, and project a future when maximum taxonomic and functional information will be gained from every environmental DNA or RNA fragment.

Methods

Specimen sampling and species-level identification. Specimens were collected in the field or obtained from cultures, supplemented with existing soil invertebrate specimens from Senckenberg museum collections (Supplementary Data 1). Sampling occurred between 2011 and 2020, mostly in Germany, but in some cases also from countries in Europe. Soil macrofauna was mainly collected by hand, whereas meso- and microfauna were obtained from soil samples with MacFadyen[66] or Baermann extraction[67]. DNA was extracted from over 500 single specimens, or occasionally from multiple individuals (single-species cultures of tardigrades and smaller-bodied nematodes, Supplementary Data 5). A non-destructive DNA extraction method[68] was preferred and used where possible. Otherwise, the MagAttract High Molecular Weight DNA Kit (Qiagen, Hilden, Germany) was used, mostly for cultured specimens. Voucher specimens are deposited in the Senckenberg museum collection in Görlitz.

For larger taxa such as Chilopoda, Diplopoda, Isopoda, Enchytraeidae and Lumbricidae, the species-level morphological identification was possible before DNA extraction, and only a single leg, a few body segments or musculature of mouthparts were used for DNA extraction, the rest of the body was kept as a voucher. For medium sized taxa like Acari and Collembola that normally would require clearing in lactic acid prior to species identification, the specimens were presorted on family or genus level, the whole specimens were used for non-destructive DNA extraction, and finally species-level identifications were carried out with recovered vouchers. In cases where non-destructive DNA extraction did not deliver sufficient amounts of DNA or the voucher was lost during extraction, identification was validated by aligning species markers (28 S, COI) from the whole-genome sequence data with existing species markers in GenBank or generated by us. For small, soft-skinned taxa (Nematoda, Tardigrada), where non-destructive DNA extraction is not possible, two different sources/techniques were used: (1) for most species, specimens were derived from own established cultures with known taxon and strain names, or (2) where such cultures did not exist, we freshly Baermann-extracted specimens from soil samples and identified morphospecies with at least 6 specimens at 400x magnification under an inverted microscope. We then extracted DNA from half of the specimens and prepared permanent slides of the other half (vouchers). We assigned species identity to the genome-sequenced specimens, if all vouchers were identified as the same species[69].

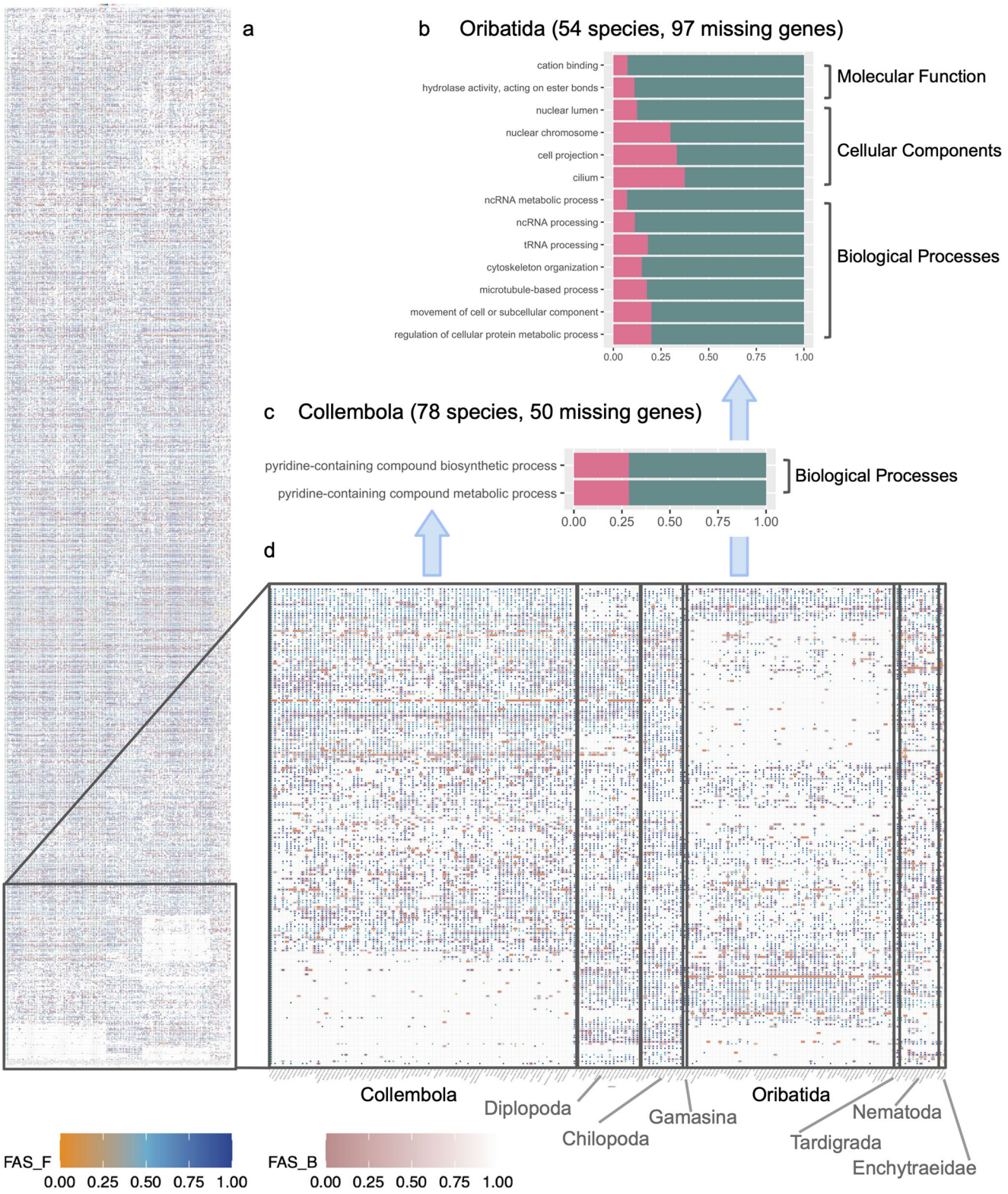

Fig. 5 Loss of metazoan core genes in soil invertebrate species. phylogenetic profiles of 1482 metazoan core genes across 177 soil invertebrate species; fraction of genes annotated with GO terms in the loss set (red) and in the background set (green) in oribatid mites **b** and springtails **c**; **d** genes consistently missing in springtails or oribatid mites. Colours in **a**, **d** represent feature architecture similarity among the identified orthologs and the reference gene, with a score between 1 (same architecture) and 0 (dissimilar architecture, or no features in the reference protein). The score is computed once by comparing the reference gene with the identified ortholog (FAS_F, dots on the graphic), and once by comparing the identified ortholog with the reference gene (FAS_B, background colour to dots). Data underlying the GO enrichment analysis are available as Supplementary Data 4.

Illumina sequencing. Sequencing libraries for each specimen, or pool of specimens, were prepared in-house at Senckenberg, Frankfurt, Germany with the BEST protocol[70] or with the NEBnext ULTRA II DNA Library Prep Kit, according to the manufacturer's protocol. Short-read Illumina sequencing (300-bp paired-end) was carried out at Novogene Europe (Cambridge, UK) using the NovaSeq 6000 platform, with unique dual indexing as the library tagging strategy for multiplexing on the lanes. Our central aim with the genome database was to improve species identifications. As this can be achieved with low sequencing coverage[65], our initial sequencing efforts targeted 2 gigabase (Gb) per species. We increased efforts to 10 Gb per species as sequencing became more affordable. For most of the reported genomes we obtained ~10 Gb per species.

Genome assembly pipeline. We established a pipeline to assemble reads into draft genomes (Fig. 1). First, the sequencing adapters were trimmed using Trimmomatic (v0.39; paramaters: ILLUMINACLIP:adapters.fasta:2:30:10:8:true SLIDINGWINDOW:4:20 MINLEN:50 TOPHRED33[71]). The trimmed reads were queried against the human genome (GRCh38 assembly on NCBI) using Kraken2 (v2.0.9-beta; --confidence set to 0.2, other parameters default[72]), and all 'human' positive reads, if any, were discarded. The remaining reads were then assembled using SPAdes (v3.14.1; default settings[73]). The resulting contigs were then queried against the NCBI non-redundant nucleotide database using blastn (megablast mode, -max_target_seqs 10, -max_hsps 1, -evalue 1e-25), and against the NCBI non-redundant protein database using Diamond (blastx mode, --sensitive --max-target-seqs 1, --evalue 1e-25[74]). NCBI databases were downloaded on 27-Oct-2020. Blobtools2 (v2.3.3[7]) was used to perform a taxonomic assignment based on the Blast and Diamond results, using the 'bestsumorder' rule. The contigs assigned to the phylum of the target organism as well as the unassigned contigs were kept (i.e., contigs assigned to other phyla were considered obvious contaminants and removed). Redundans (v0.14a[75]) was used to reduce the amount of duplication in the retained contigs, as well as further scaffolding and gap closing (default parameters were used). The resulting scaffolds were used as the final assembly draft for subsequent analyses. The Burrows-Wheeler Aligner (BWA) was then used to map the reads on the assembly and samtools[76] (v1.11-2-g26d7c73) to compute and plot the mapping statistics (e.g., GC content).

Quality assessment of assemblies using BUSCO. Benchmarking Universal Single-Copy Orthologs (BUSCO) databases[27] are sets of genes for specific taxon groups, where every gene in the BUSCO set is expected to be present once in each member species. We searched for BUSCO genes in our final assemblies as a quality indicator of genome assembly completeness, we used the most specific BUSCO database that was available for each of the invertebrate groups (nematoda_odb10 BUSCO genes for nematode assemblies, arthropoda_odb10 for arthropods, metazoa_odb10 for tardigrades, enchytraeids and earthworms). We selected the genome assembly with the highest percentage of complete BUSCO genes as the species representative if more than a single replicate per species was available. This resulted in a total of 232 genome assembly drafts used for downstream analyses.

Improving metatranscriptomic assignments. Metatranscriptomic reads were generated from soil samples collected along an elevation gradient spanning 400 m of elevation in the Alps[28,77,78]. Briefly, short soil cores were taken and preserved in LifeGuard (Qiagen, Hilden, Germany) in 2015 and 2017. RNA was extracted with an RNeasy PowerSoil Total RNA Kit (Qiagen) from ten cores. RNA sequencing libraries were prepared of each RNA extracts with a NEBNext Ultra RNA Library Prep Kit (Frankfurt am Main, Germany), and 8 gigabases of each library were sequenced at Novogene (UK) on an Illumina NovaSeq6000 sequencer in a 150 bp paired-end reaction. Reads were trimmed of adapters with Trimmomatic[71]. Reads were taxonomically assigned with kraken2[72] in a three-step process. First, we screened the metatranscriptomes against the human genome for eventual human contamination. Second, we assigned remaining reads with a custom database containing all bacterial, plant and fungal reference genomes from NCBI (accessed on 15.1.2023). Third, we then tested the impact of a dedicated genome database for soil invertebrate detection: unassigned reads from the second step were mapped against all springtail (57), oribatid mite (9) and myriapod genomes (8) available in NCBI RefSeq as of 20.6.2023, with and without including the 232 MetaInvert genomes (Supplementary Data 2). We visualised the richness of soil invertebrates along the elevation gradient at the genus level. As nucleotide sequence counts are not normally distributed and they are frequently overdispersed[79], we evaluated differences in community composition among the study years and habitats, and along the elevation with a model-based analysis of multivariate abundance data[80]. Community analyses were performed in R v4.2.2[81].

Building the phylogeny using metazoan BUSCO genes. We searched for BUSCO genes with the metazoan_obd10 database (v4.1.4) to generate a single phylogeny of the 232 soil invertebrate genomes and a selection of 118 publicly available invertebrate RefSeq genomes from NCBI (downloaded on 16.09.2021). The RefSeq genomes were included if they a) were from the same taxon group as our specimens, b) served to shorten the evolutionary distance between taxa in the tree. More specifically, we included any chromosome-level Protostomia genomes (excluding Insecta), genomes of any assembly quality for species within our 14 taxonomic groups of interest, and some additional specific outgroups (two Echinodermata, three Rotifera, a Priapulida, *Machilis hrabei* and *Drosophila albomicans*). We found 141 metazoan BUSCO genes which were present in at least 75% of the genome assemblies (Supplementary Data 1, 6). The phylogenetic approach is based on the https://github.com/mag-wolf/BUSCO-to-Phylogeny pipeline. We aligned these with Mafft (v7.481[82]) with 1000 iterative refinements. These gene alignments were then concatenated into a supermatrix using FASCONCAT (v1.04[83]) and trimmed using clipkit (v1.1.5[84]), keeping only parsimony-informative and completely conserved sites. We used IQ-TREE (v2.0.3[85]) to build four separate maximum likelihood trees (each with 1000 bootstrap replicates), selecting the best one based on the -log Likelihood value closest to zero[86]. We used R to visualise the phylogeny, using the packages ggtree (v3.1.5.900[87]), tidyverse, treeio (v1.17.2[88]) and colorspace[89].

We note the placement of Tardigrada in our phylogeny is next to Nematoda which is in disagreement with the currently accepted view that they should be closer to Arthropoda[90]. This is likely an artefact due to lack of public outgroup data[91,92], and has no downstream consequences for our analyses.

Estimating genome size. To estimate genome size, we used ModEst[31] which yields results comparable in accuracy to flow cytometry, the main non-sequencing method of genome size estimation, even from incomplete genomes. Briefly, we first plotted the distribution of sequencing coverage across each genome and visually inspected each plot for the mode coverage (the highest point of the peak). If a genome assembly did not have a clearly discernible peak in sequencing coverage then genome size was not estimated for this species. Otherwise, genome size was

estimated by dividing the total mapped bases by the mode coverage.

Estimating effective population size. Using mlRho (v2.9) we estimated theta directly from the genome data by making use of the genome-wide heterozygosity in the reference individual. This proxy measure of effective population size was calculated individually for each genome assembly with at least 8X coverage, twice as high as recommended by Haubold et al.[93].

Annotating repeat content. In addition to investigating several genome properties (i.e., GC content, BUSCO gene content, genome size and effective population size), and because repeat content is particularly relevant for explaining genome size variation among species, we also annotated the repetitive elements. Species-specific repeat libraries were constructed using the automated RepeatModeler (v2.0.1) pipeline with LTR Structural discovery pipeline activated[94]. For each genome, the resulting repeat libraries were merged with the RepBase (v26.05) Arthropoda-specific section[95] and subsequently used for the annotation and estimation of proportion of repetitive elements with RepeatMasker (v4.1.2-P1[96]).

Ecological trait data. To connect the 232 new genome assemblies with ecological traits of the respective species, we first gathered existing functional trait data from Edaphobase (https://portal.edaphobase.org/) and from literature. We focussed on a) body length (minimum female adult body length for nematodes, and mean body length for all other taxa) as a proxy for body size, b) reproduction mode, and c) known occurrences in different soil habitat types (based on level 2 hierarchies described by the Coordination of information on the environment (CORINE)[37]). We provide this collected information as an additional database resource in Supplementary Data 1.

Structural equation models. We tested established or hypothesised causal relations among genomic, life-cycle and ecological variables through a series of structural equation models, with the aim of resolving multivariate relationships from the many interrelated variables. We selected only genomes with at least 50% BUSCO completeness and 8X mode coverage. Log transformations were applied to body size variables (due to non-normal distribution as determined by a two-sided Kolmogorov-Smirnov test ($p < 0.01$). We fitted the SEMs with piecewiseSEM (v2.1.0[97]). We performed the path analyses for all taxa together (linear mixed effect models, with soil invertebrate groups as random variable), and separately for each of the more densely sampled taxa (Collembola, Oribatida, combined Chilopoda and Diplopoda, and Nematoda, linear models). Reproduction mode was included only into the models of all taxa and of oribatids, as this data were limited in the other groups.

Searching for core metazoan genes. As a first-look into the functional capacities of the soil invertebrates in our study, we searched the genomes for potential loss of protein-coding genes. To make this analysis robust, we decided to focus on evolutionarily old genes that were present already in the last common ancestor of the animals. Using 11 species from across the Metazoa tree of life (Supplementary Data 7) which were part of the Orthologous MAtrix database (OMA[98]), we computed a list of 1482 core metazoan genes which were common to at least 9 of these 11 species using DCC (https://github.com/BIONF/dcc2) and pre-computed ortholog groups from the OMA DB. Given the evolutionary age of these genes and their conserved presence throughout the animal evolution, it appears likely that their loss has a substantial functional impact. We preferred to use a custom core gene set over the standard BUSCO Metazoa ODB10 data set mainly for two reasons First, the BUSCO set with only 954 core genes is considerably smaller than the set computed by us. This gives us more power to detect differences in the presence/absence pattern of genes in the analysed taxa. Second, OMA groups represent cliques of orthologous proteins, i.e., all members within a group identify each other as pair-wise orthologs. As a consequence, OMA groups reconstruct orthologous relationships across proteins from many taxa with the highest precision among all available tools[98]. We then searched for orthologs of these 1482 metazoan core genes among the more complete (>50% BUSCO completeness) soil invertebrate genomes ($n = 177$) with fDOG-Assembly (https://github.com/BIONF/fDOG/tree/fdog_goes_assembly). fDOG-Assembly performs targeted, feature-aware ortholog search without the need for annotated genomes as the starting point. Due to the taxonomic breadth of our dataset, six separate ortholog searches were performed, each using the three most closely-related reference species with protein annotations available (Supplementary Data 8). Genes without orthologs in all investigated species were excluded from the following analyses. The resulting phylogenetic ortholog profiles were visualised with PhyloProfile (v1.8.6[99]) and clustered according to the euclidean distance of the presence and absence patterns of the ortholog groups. Hence, after visual inspection of the ortholog profiles, we were able to identify patches of core metazoan genes which were missing from certain groups.

We tested for gene ontology (GO) enrichment of the potentially missing genes using the InterProScan database[100] and the function runTest from the topGO package (v2.42.0[101]). For this GO-enrichment analysis, 1482 core metazoan genes were assigned to their ontology group(s), where GO annotation data were available. Using this list as a comparison, the two gene lists of interest (50 genes missing from the 78 Collembola species; 97 genes missing from the 54 Oribatida species) were separately tested for any significant enrichment of genes belonging to any of the three gene ontology groups (biological process, cellular component, or molecular function). Significant enrichment of a gene ontology term in the missing genes was stated when the category was represented by more than five genes in the list of 1482 core metazoan genes and with a significant over-representation in a Fisher's exact test ($p < 0.05$). Further, we manually screened putative functions associated with genes missing from Collembola and Oribatida assemblies in the UniProt database (accessed on 28.6.2023), aiming to identify functional or other relevant commonalities which might be missed by an algorithmic GO enrichment analysis.

The mean empirical probability of not being able to detect a particular gene in a taxon was 0.22 for all OMA genes, excluding those missing in springtails and oribatids. So this is also the probability of not finding a particular OMA gene in the genome of a new taxon. The probability that it is actually present in the majority of the taxa if it is also not found in the second sequenced species drops to 0.05; already in the third species in which the gene is not found, the probability that the gene is actually present in the majority of the species of the taxon is below the significance level.

Statistics and reproducibility. Genome analyses are based on Illumina genomes of 232 soil invertebrate species. Genome sizes could be estimated for 191 species. Metatranscripomic assignment was performed on 10 soil RNA samples. Structural equation models were fitted on genome properties of 143 taxa, including 27 myriapods, 34 oribatids, 68 springtails, 5 nematodes. Genomes of 177 species were assessed for the presence of core metazoan genes.Tests of normal distribution were performed to ensure that

assumptions of regression are fulfilled for the structural equation models.

Reporting summary. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Vouchers are deposited in the collections of the Senckenberg Museum of Natural History Görlitz (SMNG), Germany. Raw sequence files and draft assemblies accessible through the ENA/NCBI project PRJNA758215. 28 S and COI barcodes are publicly available at dx.doi.org/10.5883/DS-TBGMI. Genome metadata can be accessed at the Genomes on a Tree (https://goat.genomehubs.org/projects/METAINVERT). Repeat elements can be accessed in the Dfam database (https://www.dfam.org/). Alignment of BUSCO genes and the resulting phylogenetic tree are available in FigShare (https://doi.org/10.6084/m9.figshare.24435052)[29]. Source data for Fig. 2 are part of Supplementary Data 1. Source data for Fig. 3 are provided in Supplementary Data 2. Source data for Fig. 5c, d are provided as Supplementary Data 4.

Code availability

No custom code or mathematical algorithms are central for the conclusions of the paper. R commands for metatranscriptome analysis and structural equation models are deposited in FigShare[29]. A list of used software with versions are deposited in FigShare[29].

Received: 16 August 2023; Accepted: 21 November 2023;
Published online: 08 December 2023

References

1. FAO, ITPS, GSBI, CBD & EC. *State of knowledge of soil biodiversity - Status, challenges and potentialities, Report 2020*. (FAO). https://doi.org/10.4060/cb1928en. 2020.
2. Potapov, A. M. et al. Feeding habits and multifunctional classification of soil-associated consumers from protists to vertebrates. *Biol. Rev.* **97**, 1057–1117 (2022).
3. García-Palacios, P., Maestre, F. T., Kattge, J. & Wall, D. H. Climate and litter quality differently modulate the effects of soil fauna on litter decomposition across biomes. *Ecol. Lett.* **16**, 1045–1053 (2013).
4. Bardgett, R. D. & van der Putten, W. H. Belowground biodiversity and ecosystem functioning. *Nature* **515**, 505–511 (2014).
5. de Vries, F. T. & Wallenstein, M. D. Below-ground connections underlying above-ground food production: a framework for optimising ecological connections in the rhizosphere. *J. Ecol.* **105**, 913–920 (2017).
6. Lavelle, P. et al. Soil invertebrates and ecosystem services. *Eur. J. Soil Biol.* **42**, S3–S15 (2006).
7. Challis, R., Richards, E., Rajan, J., Cochrane, G. & Blaxter, M. BlobToolKit – interactive quality assessment of genome assemblies. *G3 Genes Genomes Genet.* **10**, 1361–1374 (2020).
8. Guerra, C. A. et al. Tracking, targeting, and conserving soil biodiversity. *Science* **371**, 239–241 (2021).
9. Potapov, A. M. et al. Size compartmentalization of energy channeling in terrestrial belowground food webs. *Ecology* **102**, e03421 (2021).
10. Stork, N. E. How many species of insects and other terrestrial arthropods are there on earth? *Annu. Rev. Entomol.* **63**, 31–45 (2018).
11. Pearson, D. L., Hamilton, A. L. & Erwin, T. L. Recovery plan for the endangered taxonomy profession. *BioScience* **61**, 58–63 (2011).
12. Greshake Tzovaras, B. et al. What is in umbilicaria pustulata? A metagenomic approach to reconstruct the holo-genome of a lichen. *Genome Biol. Evol.* **12**, 309–324 (2020).
13. Pedersen, M. W. et al. Supplement: postglacial viability and colonization in North America's ice-free corridor. *Nature* **537**, 45–49 (2016).
14. Schmidt, A. et al. Shotgun metagenomics of soil invertebrate communities reflects taxonomy, biomass, and reference genome properties. *Ecol. Evol.* **12**, e8991 (2022).
15. Wang, Y. et al. Late Quaternary dynamics of Arctic biota from ancient environmental genomics. *Nature* **600**, 86–92 (2021).
16. Bista, I. et al. Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. *Mol. Ecol. Resour.* **18**, 1020–1034 (2018).
17. Yates, M. C., Derry, A. M. & Cristescu, M. E. Environmental RNA: a revolution in ecological resolution? *Trends Ecol. Evol.* **36**, 601–609 (2021).
18. Shakya, M., Lo, C.-C. & Chain, P. S. G. Advances and challenges in metatranscriptomic analysis. *Front. Genet.* **10**, 904 (2019).
19. Seeber, P. A. & Epp, L. S. Environmental DNA and metagenomics of terrestrial mammals as keystone taxa of recent and past ecosystems. *Mammal. Rev.* **52**, 538–553 (2022).
20. Bálint, M. et al. Environmental DNA time series in ecology. *Trends Ecol. Evol.* **33**, 945–957 (2018).
21. Pedersen, M. W. et al. Environmental genomics of Late Pleistocene black bears and giant short-faced bears. *Curr. Biol.* https://doi.org/10.1016/j.cub.2021.04.027. (2021).
22. Law, S. R. et al. Metatranscriptomics captures dynamic shifts in mycorrhizal coordination in boreal forests. *Proc. Natl Acad. Sci. USA* **119**, e2118852119 (2022).
23. Lewin, H. A. et al. The Earth BioGenome Project 2020: starting the clock. *Proc. Natl Acad. Sci. USA* **119**, e2115635118 (2022).
24. Genereux, D. P. et al. A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).
25. Feng, S. et al. Dense sampling of bird diversity increases power of comparative genomics. *Nature* **587**, 252–257 (2020).
26. Hotaling, S. et al. Long reads are revolutionizing 20 years of insect genome sequencing. *Genome Biol. Evol.* **13**, evab138 (2021).
27. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
28. Merges, D. et al. Metatranscriptomics reveals contrasting effects of elevation on the activity of bacteria and bacterial viruses in soil. *Mol. Ecol.* https://doi.org/10.1111/mec.16756. (2022).
29. Collins, G. et al. Supplementary Data to MetaInvert. https://doi.org/10.6084/m9.figshare.24435052.v1. (2023).
30. Winkler, M. et al. Side by side? Vascular plant, invertebrate, and microorganism distribution patterns along an alpine to nival elevation gradient. *Arct. Antarct. Alp. Res.* **50**, e1475951 (2018).
31. Pfenninger, M., Schönnenbeck, P. & Schell, T. ModEst: accurate estimation of genome size from next generation sequencing data. *Mol. Ecol. Resour.* **22**, 1454–1464 (2022).
32. Gregory, T. R. Coincidence, coevolution, or causation? DNA content, cellsize, and the C-value enigma. *Biol. Rev.* **76**, 65–101 (2001).
33. Lynch, M. Evolution of the mutation rate. *Trends Genet.* **26**, 345–352 (2010).
34. Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G. & Lynch, M. Drift-barrier hypothesis and mutation-rate evolution. *Proc. Natl Acad. Sci. USA* **109**, 18488–18492 (2012).
35. Blommaert, J. Genome size evolution: towards new model systems for old questions. *Proc. R. Soc. B Biol. Sci.* **287**, 20201441 (2020).
36. Pasquesi, G. I. M. et al. Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. *Nat. Commun.* **9**, 2774 (2018).
37. Steemans, C. Coordination of Information on the Environment (CORINE). *Encycl. Geogr. Inf. Sci. Ed. Kemp K Sage Publ. Inc Thousand Oaks CA* 49–50 (2008).
38. Hubbell, S. P. *The Unified Neutral Theory of Biodiversity and Biogeography*. (Princeton University Press, 2001).
39. Wang, J., Santiago, E. & Caballero, A. Prediction and estimation of effective population size. *Heredity* **117**, 193–206 (2016).
40. Chénais, B., Caruso, A., Hiard, S. & Casse, N. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene* **509**, 7–15 (2012).
41. Hawkins, J. S., Grover, C. E. & Wendel, J. F. Repeated big bangs and the expanding universe: directionality in plant genome size evolution. *Plant Sci.* **174**, 557–562 (2008).
42. Rocha, E. P. C. & Danchin, A. Base composition bias might result from competition for metabolic resources. *Trends Genet.* **18**, 291–294 (2002).
43. Chen, Y.-J. et al. Metabolic flexibility allows bacterial habitat generalists to become dominant in a frequently disturbed ecosystem. *ISME J.* **15**, 2986–3004 (2021).
44. Chaurasia, A., Uliano, E., Berná, L., Agnisola, C. & D'Onofrio, G. Does Habitat Affect the Genomic GC Content? A Lesson from Teleostean Fish: A Mini Review. In *Fish Ecology* 61–80 (Nova Science Publishers, 2011).
45. Foerstner, K. U., von Mering, C., Hooper, S. D. & Bork, P. Environments shape the nucleotide composition of genomes. *EMBO Rep.* **6**, 1208–1213 (2005).
46. Moura, A., Savageau, M. A. & Alves, R. Relative amino acid composition signatures of organisms and environments. *PLoS ONE* **8**, e77319 (2013).
47. Charlesworth, B. & Barton, N. Genome size: does bigger mean worse? *Curr. Biol.* **14**, R233–R235 (2004).
48. Canapa, A., Barucca, M., Biscotti, M. A., Forconi, M. & Olmo, E. Transposons, genome size, and evolutionary insights in animals. *Cytogenet. Genome Res.* **147**, 217–239 (2015).
49. Elliott, T. A. & Gregory, T. R. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20140331 (2015).
50. Kapusta, A., Suh, A. & Feschotte, C. Dynamics of genome size evolution in birds and mammals. *Proc. Natl Acad. Sci. USA* **114**, E1460–E1469 (2017).

51. Plohl, M., Luchetti, A., Meštrović, N. & Mantovani, B. Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. *Gene* **409**, 72–82 (2008).
52. Meštrović, N. et al. Structural and functional liaisons between transposable elements and satellite DNAs. *Chromosome Res.* **23**, 583–596 (2015).
53. Hultgren, K. M., Jeffery, N. W., Moran, A. & Gregory, T. R. Latitudinal variation in genome size in crustaceans. *Biol. J. Linn. Soc.* **123**, 348–359 (2018).
54. Yu, J. P., Liu, W., Mai, C. L. & Liao, W. B. Genome size variation is associated with life-history traits in birds. *J. Zool.* **310**, 255–260 (2020).
55. Raffaele, S. & Kamoun, S. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat. Rev. Microbiol.* **10**, 417–430 (2012).
56. Bono, L. M., Draghi, J. A. & Turner, P. E. Evolvability costs of niche expansion. *Trends Genet. TIG* **36**, 14–23 (2020).
57. MacArthur, R. H. *Geographical Ecology*. (Harper & Row Publishers Inc., 1972).
58. Sachdeva, V., Husain, K., Sheng, J., Wang, S. & Murugan, A. Tuning environmental timescales to evolve and maintain generalists. *Proc. Natl Acad. Sci. USA* **117**, 12693–12699 (2020).
59. Albalat, R. & Cañestro, C. Evolution by gene loss. *Nat. Rev. Genet.* **17**, 379–391 (2016).
60. Sharma, V. et al. A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nat. Commun.* **9**, 1215 (2018).
61. Guijarro-Clarke, C., Holland, P. W. H. & Paps, J. Widespread patterns of gene loss in the evolution of the animal kingdom. *Nat. Ecol. Evol.* **4**, 519–523 (2020).
62. De, S. et al. Pyridine: the scaffolds with significant clinical diversity. *RSC Adv.* **12**, 15385–15406 (2022).
63. Suring, W. et al. Evolutionary ecology of beta-lactam gene clusters in animals. *Mol. Ecol.* **26**, 3217–3229 (2017).
64. Schneider, C. et al. Two high-quality de novo genomes from single ethanol-preserved specimens of tiny metazoans (Collembola). *GigaScience* **10**, 5 (2021).
65. Bohmann, K., Mirarab, S., Bafna, V. & Gilbert, M. T. P. Beyond DNA barcoding: the unrealized potential of genome skim data in sample identification. *Mol. Ecol.* **29**, 2521–2534 (2020).
66. Macfadyen, A. Improved funnel-type extractors for soil arthropods. *J. Anim. Ecol.* **30**, 171–184 (1961).
67. Decker, H. *Phytonematologie*. (Deutscher Landwirtschftsverlag, 1969).
68. Gilbert, M. T. P., Moore, W., Melchior, L. & Worobey, M. DNA extraction from dry museum beetles without conferring external morphological damage. *PLoS ONE* **2**, e272 (2007).
69. Schenk, J., Hohberg, K., Helder, J., Ristau, K. & Traunspurger, W. The D3-D5 region of large subunit ribosomal DNA provides good resolution of German limnic and terrestrial nematode communities. *Nematology* **19**, 821–837 (2017).
70. Carøe, C. et al. Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.* **9**, 410–419 (2017).
71. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
72. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
73. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
74. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
75. Pryszcz, L. P. & Gabaldón, T. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* **44**, e113 (2016).
76. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
77. Merges, D., Bálint, M., Schmitt, I., Böhning-Gaese, K. & Neuschulz, E. L. Spatial patterns of pathogenic and mutualistic fungi across the elevational range of a host plant. *J. Ecol.* **106**, 1545–1557 (2018).
78. Merges, D., Bálint, M., Schmitt, I., Manning, P. & Neuschulz, E. L. High throughput sequencing combined with null model tests reveals specific plant-fungi associations linked to seedling establishment and survival. *J. Ecol.* **108**, 574–585 (2020).
79. Bálint, M. et al. Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. *FEMS Microbiol. Rev.* **40**, 686–700 (2016).
80. Wang, Y., Naumann, U., Wright, S. T. & Warton, D. I. mvabund– an R package for model-based analysis of multivariate abundance data. *Methods Ecol. Evol.* **3**, 471–474 (2012).
81. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2022).
82. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
83. Kück, P. & Meusemann, K. FASconCAT: convenient handling of data matrices. *Mol. Phylogenet. Evol.* **56**, 1115–1118 (2010).
84. Steenwyk, J. L., Iii, T. J. B., Li, Y., Shen, X.-X. & Rokas, A. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol.* **18**, e3001007 (2020).
85. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
86. Misof, B. et al. Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–767 (2014).
87. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
88. Wang, L.-G. et al. Treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. *Mol. Biol. Evol.* **37**, 599–603 (2020).
89. Zeileis, A. et al. Colorspace: a toolbox for manipulating and assessing colors and palettes. *J. Stat. Softw.* **96**, 1–49 (2020).
90. Treffkorn, S., Mayer, G. & Janssen, R. Review of extra-embryonic tissues in the closest arthropod relatives, onychophorans and tardigrades. *Philos. Trans. R. Soc. B Biol. Sci.* **377**, 20210270 (2022).
91. Giribet, G. & Edgecombe, G. D. The phylogeny and evolutionary history of arthropods. *Curr. Biol.* **29**, R592–R602 (2019).
92. Telford, M., Rota-Stabelli, O. & Pisani, D. Phylo-evo-devo, tardigrades and insights into the evolution of segmentation. in (Padova University Press, 2018).
93. Haubold, B., Pfaffelhuber, P. & Lynch, M. mlRho - a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Mol. Ecol.* **19**, 277–284 (2010).
94. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
95. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
96. Smit, A., Hubbley, R. & Green, P. RepeatMasker Open 4.0. http://www.repeatmasker.org (2015).
97. Lefcheck, J. S. piecewiseSEM: piecewise structural equation modelling in r for ecology, evolution, and systematics. *Methods Ecol. Evol.* **7**, 573–579 (2016).
98. Altenhoff, A. M. et al. OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res.* **49**, D373–D379 (2021).
99. Tran, N.-V., Greshake Tzovaras, B. & Ebersberger, I. PhyloProfile: dynamic visualization and exploration of multi-layered phylogenetic profiles. *Bioinf. Oxf. Engl.* **34**, 3041–3043 (2018).
100. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
101. Alexa, A. & Rahnenfuhrer, J. topGO: Enrichment Analysis for Gene Ontology v2. https://doi.org/10.18129/B9.bioc.topGO (2022).

Acknowledgements

This work is a result of the LOEWE Centre for Translational Biodiversity Genomics funded by the Hessen State Ministry of Higher Education, Research and the Arts (HMWK). This project contributes to the Soil Invertebrate Genome Initiative (https://tbg.senckenberg.de/sigi/), affiliated with the Earth BioGenome Project. Special thanks to Damian Baranski, Jürgen Otte and Jörg Müller for DNA extractions, to Astrid König for extraction and preparation of nematodes and tardigrades from Senckenberg cultures and soils, to Lena Bonassin and Jade Tessien for help with organising the datasets and barcodes, to Prof. Dr. Florian Grundler (INRES Molekulare Phytomedizin, Bonn University, Germany) for thousands of J2 juveniles of *Heterodera schachtii* and *Meloidogyne incognita* in ethanol from their cultures, and to Magnus Wolf for the BUSCO-to-phylogeny pipeline. Animal silhouettes originate from PhyloPic and they can be reused under Creative Commons licenses (http://www.phylopic.org).

Author contributions

P.D., I.E., K.H., O.L., R.L., M.P., M.B. conceived and designed the experiments. P.D., K.H., H.M., R.L., performed the experiments. G.C., C.S., L.B., I.E., O.L., H.M., J.u.R., C.R., R.V., M.P., M.B. analysed the data. C.S., L.B., U.B., A.C., P.D., I.E., K.H., D.M., H.M., J.ö.R., J.u.R., C.R., R.S., A.S., K.T. contributed materials/analysis tools. G.C., M.P., M.B. wrote the paper.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-023-05621-4.

Correspondence and requests for materials should be addressed to Miklós. Bálint.

Peer review information This manuscript was previously reviewed at another Nature Portfolio journal. *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: George Inglis. A peer review file is available.

Reprints and permission information is available at http://www.nature.com/reprints

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

Supplementary 2: Decapoda conserved proteins across crustacean taxonomic groups. CNS = Conserved Non-Specific. CS = Conserved Specific. Crus = Crustacea. Deca = Decapoda. Pleo = Pleocyemata. Dend = Dendrobranchiata. Asta = Astacidea. Brac = Brachyura. R = Remaining. SS = Species specific

Species Name	Species Group	CNS Crus	CS Crus	CNS Deca	CS Deca	CNS Pleo	CS Pleo	CNS Dend	CS Dend	CNS Asta	CS Asta	CNS Brac	CS Brac	R	SS
P. monodon	Dendrobranchiata	3011	4	3037	625			2569	1761					10498	2506
P. japonicus		3199	4	3260	572			2713	1813					9210	1530
P. vannamei		3062	3	3060	589			2614	1898					10974	2787
P. chinensis		3207	2	3139	559			2646	1617					7718	1188
P. indicus		3102	1	3220	554			2692	1724					9068	1463
H. rubra	Caridea	2911	4	2947	604	750	190							10482	7453
H. americanus	Astacidea	3141	5	3245	586	824	188			2295	1185			7990	2909
P. clarkii		3209	2	3316	619	833	189			2387	1194			11798	2870
C. quadricarinatus		3034	3	3162	628	800	212			2267	1055			5545	1446
P. manimaculis	Anomura	3502	8	3408	652	864	230							17074	14558
P. cinctipes		3553	1	3664	711	888	301							18927	16466
P. trituberculatus	Brachyura	3214	2	3163	565	837	185					734	260	7052	1280
E. sinensis		3307	6	3343	638	859	242					747	323	8337	1813
C. opilio		2871	1	2288	516	469	127					668	404	6926	8156
S. paramamosain		3198	3	3229	575	823	188					829	313	8554	2127

Supplementary 3: Genome IDs used in addition to original databases implemented in FasQScreen.

Category	Genome ID(s)
Alpha-Proteobacteria	GCF_002549835.1, GCF_000015985.1, GCF_000697965.2, GCF_000166055.1, GCF_003071405.1, GCF_013415845.1, GCF_002847445.1, GCF_001402875.1, GCF_000284415.1, GCF_000499665.2
Beta-Proteobacteria	GCF_002752675.1, GCF_002762215.1, GCF_001465545.3, GCF_001676725.1, GCF_000015565.1, GCF_001040945.1, GCF_000176855.2, GCF_001267925.1, GCF_001761385.1, GCF_000934605.2
Delta-Proteobacteria	GCF_000014965.1, GCF_000022265.1, GCF_000280925.3, GCF_001553625.1, GCF_001628815.1, GCF_001263175.1, GCF_001263205.1, GCF_003258315.1, GCF_001278055.1, GCF_000022145.1
Epsilon-Proteobacteria	GCF_013283835.1, GCF_013201665.1, GCF_000092245.1, GCF_000568815.1, GCF_013201825.1, GCF_001723605.1, GCF_013201725.1, GCF_013177675.1, GCF_006459125.1
Gamma-Proteobacteria	GCF_002215215.1, GCF_000012985.1
Zeta-Proteobacteria	GCF_013387475.1, GCF_000379405.1, GCF_013387455.1, GCF_000153765.1
Aphanomyces astaci	GCF_000520075.1
Crustaceans	GCA_013387185.1, GCA_013436485.1, GCA_013167095.1, GCA_013283005.1, GCA_010645155.1, GCA_007210705.1, GCA_003990815.1, GCA_002872375.1, GCA_012959195.1, GCA_009805615.1, GCA_900092285.2, GCA_000764305.3, GCA_011947565.1, GCA_003789085.1, GCA_003724045.1, GCA_009830355.1, GCA_009176605.1, GCA_000591075.2, GCA_014673585.1, GCA_900157175.1, GCA_004104545.1, GCA_006783055.1, GCA_014220935.1, GCA_001005205.1, GCA_003723985.1, GCA_000981345.1, GCA_001587735.2, GCA_900659605.1, GCA_009761615.1, GCA_900241095.1, GCA_900607525.1, GCA_007890405.1, GCA_002838885.1,

Supplementary 4: Noble crayfish DNA extraction.

All extractions were mainly conducted by Lena Bonassin.

Nanopore

All sequencing runs were performed with the HMW DNA isolated from two male noble crayfish individuals. The first individual being the same as for Illumina sequencing. All library preps were performed with the SQK-LSK109 kit from ONT. DNA extraction from muscle tissue was performed using a phenol-chloroform extraction with homogenisation buffer H1, following the protocol by Sambrook and Russell 2006 (Sambrook and Russell, 2006).

PacBio

One male crayfish individual was obtained from breeder Flusskrebszucht Frömel (Kavelstorf, Germany). Muscle tissue was dissected, and flash frozen with liquid nitrogen, and stored at -80 °C until DNA extraction. High-molecular weight genomic DNA extraction was first performed using the MagAttract HMW DNA kit (Qiagen, Germany) according to the manufacturer's protocol for purification of DNA from fresh or frozen tissue with the following modifications. Lysis was performed overnight, and all incubation steps were performed at 1200 rpm. Finally, DNA was eluted in 60 µL of AE buffer preheated to 37 °C. DNA was quantified using the QuantiFluor® dsDNA System on the Quantus™ Fluorometer (Promega, USA). The fragment size distribution was assessed using the Femto Pulse System Genomic DNA 165 kb Kit (Agilent, USA). DNA was sheared to 20 kb using the Megaruptor 2 (Diagenode, USA).

Low input PacBio HiFi library was prepared according to the protocol Preparing HiFi Libraries from Low DNA Input Using SMRTbell® Express Template Prep Kit 2.0 (PacBio, Version 06, August 2020). Library preparation was started with 80 ng genomic DNA in 50 µL. After adapter ligation, nuclease treatment of the libraries was performed according to the protocol. The fragment size distribution was assessed using the Femto Pulse System (Agilent, USA). The library was sequenced using the PacBio SequelIIe at Novogene (UK). The sequencing of the first library produced 124 Mb of HiFi data and 15 211 HiFi reads. The read number is lower than the expected 4 000 000 reads that can be produced on a SequelIIe system (https://www.pacb.com/technology/hifi-sequencing/sequel-system/). To improve the

sequencing yield we tested different extraction methods and library preparation protocols (data not shown). The second DNA extraction was performed using a combined protocol of sorbitol wash and salting out protocol. Tissue was grinded and sorbitol wash was added following the protocol (Jones and Schwessinger, 2020). DNA extraction was continued using the salting out protocol (Jenkins et al., 2019) with the following modifications: the digestion of the tissue was performed for 3 h at 65 °C and 400 rpm, to remove the proteins and cellular debris the samples were centrifuged at 5000 g for 10 min, and to precipitate the DNA the samples were centrifuged at 5000 g for 5 min. Finally, the DNA pellet was resuspended in 60 µL nuclease-free water. DNA was quantified using the QuantiFluor® dsDNA System on the Quantus™ Fluorometer (Promega, USA). The DNA purity was estimated using the Nanophotometer P300 (Implen, Germany). The fragment size distribution was assessed using the Femto Pulse System (Agilent, USA).

For ultra-low PacBio sequencing, library preparation and sequencing were performed at the West German Genome Centre (WGGC, Düsseldorf, Germany). Libraries were prepared with ultra-low input WGS workflow using the Express Template Prep Kit 2.0 (PacBio) according to protocol Preparing HiFi SMRTbell Libraries from Ultra-Low DNA Input (PacBio, Version 02, August 2020). Library preparation was started with 20 ng genomic DNA in 50 µL. The DNA was sheared using gTubes (Covaris, UK) aiming for a mean fragment size of 15kb. Centrifugation speed was increased gradually until the whole sample had passed through the membrane. Fragment size distribution was assessed using the Femto Pulse System (Agilent, USA) with FP-1002 run protocol (165 kb). The sheared sample was used as input for library preparation according to the manufacturer's instructions. Whole genome amplification was performed with the gDNA Sample Amplification Kit (PacBio) according to the above-mentioned protocol with the following modifications: Reaction mix A was replaced with LA Taq (Takara, Japan), the PCR mix was prepared with 0,75µl Takara LA Taq, 37,5µl 2x GC buffer I, 12µl dNTP mixture and 2µl PacBio amplification PCR primers. Both PCR programs were run with initially 15 cycles. Additional 2-3 cycles were performed after the initial 15 cycles if the amount of PCR product was less than 250 ng per reaction. For reaction mix A a 10 minutes extension time was used. Concentration of the amplified library was measured with Qubit fluorometer (Thermo Fischer Scientific, USA) and size distribution was analyzed on a Fragment Analyzer (Agilent) with DNF-464 run protocol using a 1 ng/µL dilution in TE. For each sample the amplified library from the

two reaction mixes were pooled and a second library preparation was performed with the SPK3 kit (PacBio) according to the protocol version 02 (Procedure-checklist-Preparing whole genome and metagenome libraries using SMRTbell prep kit 3.0, PacBio; REV02, March 2023) including nuclease treatment. The library was size selected on a Blue Pippin (Sage Science, USA) with the following settings: 0,75% DF Marker S1 High-Pass 6-10kb vs3. Cutoff 6kb. Library concentration was measured with Qubit dsDNA HS (Thermo Fischer Scientific, USA) and size distribution was analyzed with a Fragment Analyzer (Agilent, USA). Sequencing primer (v3.2) and polymerase were bound to the library using the Sequel® II Binding Kit 3.2 (PacBio) and the library was sequenced on a Sequel II/SequellIIe instrument using 20 8M SequelII SMRT Cell with a final on plate loading concentration of 85pM, 2h pre-extension and 30h movie time. Circular consensus (CCS) reads were generated with SMRT Link version 11 with min. predicted accuracy of 0.99 and min. 3 passes.

Résumé

Introduction

Les écrevisses sont des crustacés décapodes, un ordre extrêmement diversifié au sein de l'embranchement des arthropodes. Les décapodes comptent plus de 15 000 espèces vivantes, notamment les Dendrobranchiata (crevettes grises), et les Pleocyemata avec les Caridea (crevettes roses), les Achelata (langoustes et cigales de mer), les Astacidea (homards et écrevisses), les Anomura (bernard l'ermite, crabes royaux) et les Brachyura (crabes). Les décapodes se caractérisent par la présence de 10 pattes servant au déplacement incluant une paire possédant une pince. Ils sont majoritairement marins, bien que certains vivent en eau douce et quelques rares espèces sont semi-terrestres. Parmi les espèces d'eau douce, on retrouve les écrevisses qui présentent plus de 650 espèces identifiées. Les écrevisses, appartenant à la l'infra-ordre des Astacidea, peuvent être divisées en 3 familles : Parastacidae, Astacidae et Cambaridae. En Europe on retrouve 5 espèces natives, toutes appartenant exclusivement à la famille Astacidae. Parmi elles, la plus répandue est l'écrevisse noble, *Astacus astacus*, espèce emblématique en Europe.

Les écrevisses se nourrissent de manière opportuniste, s'adaptant aux ressources disponibles. Elles contribuent à améliorer la propreté de l'eau, impactent la structure et la fonction de l'écosystème, régulent les populations d'animaux et servent également de proie à plusieurs organismes faisant des écrevisses une ressource alimentaire vitale au sein du réseau trophique. Ces contributions écologiques font des écrevisses des espèces clé maintenant l'équilibre au sein de leur écosystème. De plus, les écrevisses creusent des galeries afin de créer leur habitat et trouver de la nourriture, ce qui peut physiquement façonner et influencer le milieu environnant, faisant des écrevisses des ingénieurs de l'écosystème. Leur disparition pourrait causer des changements importants, voire menacerait d'extinction l'ensemble de l'écosystème.

Les écrevisses ont également une grande valeur commerciale, étant considérées à la fois comme un mets raffiné et comme animal de compagnie. C'est à des fins commerciales que des écrevisses ont été importées en Europe. Ces espèces non-européennes, souvent plus agressives et se reproduisant plus rapidement, peuvent supplanter les espèces natives, menaçant la biodiversité et perturbant l'écosystème. On compte aujourd'hui plus de 10

espèces invasives sur le sol européen. L'importation d'espèces non-européennes s'accompagne de l'introduction de pathogènes qui affectent les populations natives. Ainsi, la peste de l'écrevisse, causée par l'oomycète *Aphanomyces astaci*, a été introduite accidentellement en Europe lors de l'importation d'écrevisses Nord-Américaines. Grâce à une longue coexistence et à la coévolution hôte-pathogène, les espèces nord-américaines ont développé des mécanismes de résistance, tandis que la peste de l'écrevisse est généralement mortelle pour les espèces européennes. Il a cependant été décelé dans certaines population d'*A. astacus* une plus longue tolérance au pathogène que dans d'autres. Cette résistance reste insuffisante pour permettre la survie de la population. Cette maladie a entraîné des pertes massives au sein des populations d'écrevisses européennes, avec des déclins allant jusqu'à 95 % observés dans certaines régions au cours des 150 dernières années, conduisant *A. astacus* à être classée comme espèce vulnérable sur la liste rouge des espèces en danger.

La situation critique des espèces d'écrevisses européennes souligne l'urgence de mettre en œuvre des stratégies efficaces de gestion et de prévention des maladies, tout en luttant contre la prolifération des espèces d'écrevisses non-européennes. Il devient de plus en plus évident que la génomique de la conservation est essentielle pour relever ces défis et améliorer notre compréhension de la diversité génétique au sein de l'espèce. Dans ce contexte, disposer d'un génome de référence pour *A. astacus* est essentiel pour mener des études de populations et mettre en œuvre des approches de conservation spécifiques à cette espèce. De plus, ce génome de 17 Gb est crucial afin de mieux comprendre l'évolution des génomes d'écrevisses qui présentent des tailles extrêmement variables. Cependant, en raison de la variabilité et de la grande taille des génomes de décapodes, pouvant atteindre 40 Gb, le séquençage des génomes s'avère souvent très difficile. Ainsi, on ne dispose à l'heure actuelle que de quatre génomes d'écrevisse et aucun ne représente une espèce européenne. Parmi les génomes de décapodes disponibles on observe une variabilité de la taille des génomes aussi bien entre les infra-ordres qu'au sein des infra-ordre. Cette taille peut varier entre 1 et 8 Gb avec les Astacidae et les Anomoures qui présentent les estimations de tailles de génomes les plus élevés. Les génomes d'Astacidae sont parmi les plus grands et les plus complexes des arthropodes, atteignant une taille estimée à 17 Gb pour *A. astacus*.

Les avancées en séquençage génomique ont transformé l'étude des espèces non-modèles, comme les écrevisses. Deux principales approches de séquençage sont utilisées : les lectures courtes et les lectures longues. Le séquençage basé sur les lectures courtes, comme Illumina, génère des fragments d'ADN de 150 pb à 300 pb. Cette méthode offre une grande précision et un faible coût. Le séquençage basé sur les lectures longues, utilisant des technologies comme Pacific Bioscience (PacBio) et Oxford Nanopore (ONT), produit des fragments d'ADN beaucoup plus longs, allant de 15 kb jusqu'à 25 kb pour PacBio, et 100 kb pour Nanopore. Cependant, cette approche souffre d'un taux d'erreur de séquençage plus élevé, bien que PacBio ait considérablement réduit ces erreurs grâce à ses lectures haute-fidélité (HiFi). Une combinaison des deux techniques est souvent utilisée pour pallier leurs limites respectives : les lectures courtes peuvent corriger les erreurs des lectures longues, et ensemble, elles permettent de produire des assemblages plus fiables.

Trois approches principales sont utilisées pour assembler les génomes : l'assemblage basé sur les lectures courtes, l'assemblage basé sur les lectures longues, et l'assemblage hybride. L'assemblage de lectures courtes rencontre souvent des difficultés dans les génomes complexes, notamment ceux riches en éléments répétés. L'assemblage de lectures longues permet de mieux assembler les régions complexes, cependant le taux d'erreur peut empêcher l'assemblage de certaines lectures et introduire des biais de séquençage dans l'assemblage finale. L'assemblage hybride, en combinant les données des lectures courtes et longues, génère des résultats plus complets et précis. Les étapes clés de l'assemblage d'un génome comprennent le regroupement des lectures chevauchantes pour former des séquences continues (*contigs*), et le *scaffolding*, qui relie les *contigs* en utilisant des informations supplémentaires. Une étape de correction des lectures longues peut être fait à l'aide de lectures courtes avant l'assemblage des lectures longues corrigées. Le *scaffolding* peut s'appuyer sur des lectures longues dans le cas d'un assemblage de lectures courte, permettant notamment de résoudre les régions riches en éléments répétés et d'organiser les contigs. Le *scaffolding* peut aussi s'appuyer sur des données de séquençages Hi-C ou de cartographie optique. Une fois l'assemblage effectué, des outils comme BUSCO (Benchmarking Universal Single-Copy Orthologs) sont utilisés pour évaluer la qualité et la complétude du génome assemblé. Ensuite, l'annotation génomique identifie les gènes et autres éléments fonctionnels, une étape cruciale pour comprendre les mécanismes biologiques sous-jacents.

Compte tenu de la taille estimée (17Gb) du génome d'*A. astacus*, une présence importante d'éléments répétés est attendue. Les éléments répétés peuvent être catégorisé en tant qu'éléments transposables ou ADN satellites. Les ADN satellites sont des séquences courtes répétées en tandem, souvent localisées dans des régions spécifiques comme les centromères et les télomères. Parmi les éléments transposables on retrouve les transposons à ADN et les rétrotransposons. Ces derniers comprennent notamment les LINEs (*Long Interspersed Nuclear Elements*) et SINEs (*Short Interspersed Nuclear Elements*). Les éléments répétés jouent un rôle central dans l'évolution et la plasticité adaptative des espèces. Leur présence contribue à l'évolution structurale, par la duplication et la divergence des séquences, l'innovation génétique, par l'introduction de nouvelles séquences fonctionnelles via les transpositions et l'adaptation environnementale, certains éléments transposables sont activés en réponse au stress, favorisant des mutations adaptatives. Cependant, ces mêmes éléments peuvent aussi provoquer des mutations délétères et une instabilité génomique, par exemple par des réarrangements chromosomiques.

Ces éléments répétés peuvent compliquent considérablement le séquençage et l'assemblage du génome. Les technologies traditionnelles de séquençage basées sur des lectures courtes s'avèrent insuffisantes pour capturer les structures complexes du génome. Les lectures longues, bien que prometteuses, demandent des ressources significatives en termes de calcul et de stockage pour assembler correctement ces génomes. L'absence de génomes de référence pour des espèces proches constitue également un obstacle majeur. En l'absence de données comparatives, l'annotation des gènes devient laborieuse, et l'identification des éléments spécifiques aux écrevisses reste limitée.

Contributions

Etude comparative des éléments répétés

Les génomes de décapodes sont encore peu explorés au niveau génomique avec seulement 22 génomes disponibles en 2022. Les assemblages disponibles ont une taille qui varie entre 1,6 Gb et 8,5 Gb tandis que les estimations de tailles peuvent aller jusqu'à 40 Gb. Ces variations suggèrent une contribution déterminante des éléments répétés dans l'évolution des génomes de Décapodes. Des études chez les arthropodes (principalement chez les insectes) ont révélé

que les éléments répétés peuvent représenter jusqu'à 80% du génome. Aucune étude comparative n'a cependant été réalisée chez les décapodes. J'ai donc analysé les éléments répétés chez les 20 génomes de décapodes disponibles et de qualité suffisante en y intégrant 6 génomes de crustacés non-décapodes à titre de comparaison. Pour ce faire, j'ai développé un pipeline permettant une annotation standardisée plus exhaustive des éléments répétés en utilisant les programmes RepeatModeler2, TAREAN et RepeatMasker ainsi que la base de données d'éléments répétés RepBase. L'utilisation de notre pipeline nous a permis d'annoter en moyenne 10% d'éléments répétés supplémentaires comparé aux précédentes annotations. De plus, la proportion d'éléments répétés non catégorisés dans chaque génome a diminué de 7,8 % en moyenne. Les résultats montrent que les éléments répétés représentent en moyenne entre 40 % et 80% des génomes des Décapodes analysés, plus important que chez les autres crustacés (25% à 50% avec une exception à 75%).

Nous avons également démontré que le nombre de copies d'éléments répétés est corrélé à la taille de l'assemblage du génome. Le pourcentage d'éléments répétés présente une corrélation moins significative avec la taille de l'assemblage, et ainsi démontre la difficulté à assembler ces éléments répétés. Les génomes plus grands, comme ceux des écrevisses, affichent une proportion encore plus élevée de familles de satellites à ADN et un nombre plus élevé de familles largement répétés, montrant l'expansion des satellites à ADN dans ces grands génomes.

L'analyse comparative des éléments transposables au sein des Décapodes révèle chez les Dendrobranchiata une présence plus élevée de transposons à ADN, tandis que chez les Pleocyemata on retrouve principalement des LINEs et des LTRs. Ce fort signal phylogénétique, à la fois au regard des autres crustacés analysés et au sein même des décapodes pourrait refléter des pressions sélectives uniques ou des événements adaptatifs liés aux environnements distincts occupés par ces espèces. Nos travaux révèlent en outre une dynamique évolutive des éléments transposables distincte entre les deux sous-ordres de décapodes. Le sous-ordre Pleocyemata présente des expansions plus récentes, voire des éléments transposables encore très actifs comparé au sous-ordre des Dendrobranchiata.

Les éléments répétés ont un impact significatif sur les propriétés des génomes des Décapodes. Leur abondance explique les complications lors de l'assemblage des séquences en augmentant le risque de fragmentation et d'erreurs lors des reconstructions génomiques. Cependant, ces éléments jouent également un rôle dans l'évolution en induisant des réarrangements chromosomiques, des duplications de gènes et des innovations fonctionnelles et présentent un signal phylogénétique participant à l'expliquer l'évolution des décapodes.

Comparaison des protéomes de décapodes

Les génomes de décapodes présentent en moyenne 25 000 protéines. Afin d'étudier les protéines de décapodes ayant des homologues parmi les décapodes mais également aux seins des eucaryotes, une sélection de 15 protéomes de Décapodes a été réalisée. Chacun des protéomes présentant un score de complétude supérieur à 60 % selon l'outil BUSCO. Ces protéomes ont été comparés entre eux pour identifier les relations d'orthologie et explorer les protéines conservées et spécifiques dans ce groupe. Les protéines de Décapodes ont également été comparées aux protéomes de la base de données OrthoInspector, contenant 1 472 espèces eucaryotes, afin de comprendre les relations entre les protéines des Décapodes et celles d'autres grands groupes taxonomiques.

L'analyse a révélé un *core-protéome* des Décapodes comprenant environ 7 000 protéines, représentant un quart du protéome moyen d'un Décapode. Ce *core-protéome* inclut des protéines essentielles impliquées dans des processus métaboliques centraux et des fonctions cellulaires de base. On y retrouve entre autres des protéines liées au métabolisme énergétique tel que des enzymes impliquées dans les voies glycolytiques et la chaîne respiratoire, des protéines essentielles pour le cycle cellulaire et la réparation de l'ADN, et des protéines liées à la signalisation cellulaire avec récepteurs et molécules associées à la réponse aux signaux environnementaux.

Les espèces appartenant à l'ordre des Pleocyemata présentent une conservation limitée de leurs protéines, avec seulement 795 protéines communes identifiées. Cela suggère que des événements évolutifs, tels que des duplications ou des pertes de gènes, ont considérablement

influencé la diversité génétique au sein de cet ordre. Chez les Dendrobranchiata, l'analyse est limitée à des espèces du genre *Penaeus*. Bien que ces espèces montrent des similarités dans leurs répertoires protéiques, les conclusions générales sont limitées par le faible nombre de génomes disponibles pour ce sous-ordre.

La comparaison du protéome de *Procambarus clarkii* aux 1472 protéomes eucaryotes par profilage phylogénétique a mis en évidence des modules évolutifs spécifiques et des clusters enrichis dans des voies telles que la détoxification cellulaire, associées à la neutralisation des métaux lourds et des contaminants organiques, et la perception des stimuli chimiques, impliquées dans la détection et la réponse aux signaux environnementaux. Ces adaptations pourraient expliquer le succès écologique de *P. clarkii*, espèce envahissante en Europe, et leur résilience dans des environnements variés. D'autres clusters, partagés par plus de groupes d'espèces sont liés à la réparation de l'ADN, les réponses immunitaires ou plus généralement à des voies métaboliques tel que les enzymes liées à la dégradation des lipides et des protéines.

L'étude comparative des protéomes des Décapodes offre une première vue d'ensemble des protéines conservées et spécifiques de ce groupe. Elle met en lumière les adaptations évolutives et fonctionnelles qui sous-tendent la diversité biologique des Décapodes. Les données générées constituent une base précieuse pour de futures études fonctionnelles. Une exploration plus approfondie des clusters de gènes spécifiques et de leurs rôles biologiques pourrait révéler les mécanismes sous-jacents à des traits phénotypiques uniques. À l'avenir, l'intégration de nouveaux génomes de Décapodes pourraient affiner ces résultats et ouvrir de nouvelles perspectives en biologie évolutive et en écologie.

Assemblage du génome de l'écrevisse noble (*Astacus astacus*)

Etant donné que seul 4 génomes d'écrevisses sont disponibles, provenant de familles éloignées d'*A. astacus*, l'assemblage *de novo* est la seule option. La grande taille du génome d'*A. astacus* estimée à 17 Gb ainsi que la proportion élevée d'éléments répétés attendue ont compliqué considérablement aussi bien le séquençage que l'assemblage du génome et nous ont conduit à tester différentes approches.

La stratégie initiale consistait à réaliser un séquençage de profondeur 20 X en lectures longues ainsi que 50 X en lectures courtes Illumina. Le séquençage Illumina a permis de d'obtenir une couverture de 45X. Au niveau des lectures longues, nous avons testé la plateforme Nanopore MinIon, mais avons obtenu un très faible rendement lié à un problème de saturation des nanopores. La plateforme Pacific Bioscience Sequel II, avec les modes CLR et HiFi, a conduit à un meilleur rendement, tout en restant insuffisant, avec au maximum 1.7 Gb par run en lectures haute-fidélité (HiFi).

Face aux difficultés rencontrées pour obtenir des lectures longues, nous avons décidé de réaliser un assemblage des lectures courtes représentant un total de 771 Gb. De nombreux outils ne peuvent gérer une aussi grande quantité de données, ou alors demandent une quantité de mémoire vive excessive. Nous avons donc testé deux stratégies. La première consiste à réaliser un assemblage dit « par étape » qui consiste à réaliser des assemblages de sous-ensemble de données de lectures courtes, puis d'assembler les contigs obtenus des différents assemblages en y incluant les lectures longues. La seconde stratégie repose sur l'utilisation de Megahit, assembleur développé pour l'assemblage de métagénomes offrant l'avantage de gérer très bien la mémoire. L'assemblage par étape des lectures courtes a permis de générer six assemblages composer d'environ 8x de couverture présentant des statistiques similaires. La taille totale de chaque assemblage est de 3,3 Gb avec un N50 égale pour chaque assemblage à 1 025 bp. Un septième assemblage, utilisant toutes les lectures mais sans les informations sur les paires de lectures a permis de générer un assemblage de 1,8 Gb avec un N50 de1 229 bp. L'assemblage de Megahit combiné à SOAPdenovo-fusion a généré un assemblage de 6,8 Gb qui reste cependant très fragmenté avec un N50 de 2 kb et un score Compleasm complet de 18,16 %.

De nouvelles tentatives de séquençage de lectures longues ont été menées. Finalement, l'utilisation de la technologie Pacific Bioscience avec la méthode « ultra-low DNA input » qui repose sur l'amplification de l'ADN génomique a permis d'obtenir un rendement convenable d'environ 20 Gb à 31 Gb par run conduisant à une couverture totale de 38 X.

L'étape de *scaffolding* utilisant les lectures longues sur l'assemblage réalisé à l'aide de Megahit a augmenté la taille de l'assemblage à 13,8 Gb, et le score de complétude à 28,14 %, cependant le N50 a diminué à 723 bp. La finalisation de l'assemblage par étape, combinant les contigs des différents assemblages ainsi que les lectures longues a permis de générer un assemblage de 8.6 Gb avec un N50 de 30 kb. Le score de complétude de Compleasm est de 34,16 %. L'assemblage par étape, bien que présentant une meilleure complétude, aboutit à une taille d'assemblage inférieure et peut également présenter plus de biais d'assemblage lié aux différents assemblages de lectures courtes de faible couverture.

Les lectures longues ont été assemblées en utilisant le programme hifiasm dédié à l'assemblage des lectures longues d'haute-fidélité (HiFi) de PacBio. L'assemblage produit atteint une taille de 21,8 Gb et s'avère nettement moins fragmenté que l'assemblage des lectures courtes avec un N50 de 128 kb et le score Compleasm présente une complétude de 42,54 %. Cette taille supérieure à la taille estimée du génome peut être attribuée aux éléments répétés non résolus ou au fait que les haplotypes n'ont pas été résolus. Bien que l'assemblage réalisé à l'aide d'hifiasm reste incomplet, les statistiques aussi bien de complétude que de contiguïté restent comparables à la qualité des génomes d'arthropodes déjà publiés, malgré une taille de génome supérieure.

L'assemblage du génome de l'écrevisse noble représente une avancée significative dans l'étude des génomes géants. Bien qu'il reste des défis à surmonter, notamment pour atteindre une meilleure complétude, ces travaux posent les bases pour une compréhension plus approfondie des mécanismes évolutifs et adaptatifs des écrevisses. Ils offrent également des outils précieux pour orienter les efforts de conservation de l'espèce.

Conclusions et Perspectives

Cette thèse représente une contribution majeure à l'étude des génomes des Décapodes, un groupe encore peu exploré malgré son importance économique, évolutive et écologique. En réalisant la première analyse comparative approfondie des génomes de ces espèces, ce travail met en lumière des conservations fonctionnelles essentielles ainsi que des spécificités évolutives marquantes. Ces travaux révèlent également le rôle crucial des éléments répétés dans l'évolution et l'organisation des génomes, et identifient un *core-protéome* partagé par

les Décapodes. L'assemblage du génome d *A. astacus*, l'un des plus complexes chez les invertébrés, constitue une base essentielle pour de futures recherches, mais doit encore être amélioré. Plusieurs défis techniques ont émergé au cours de ce travail, notamment la résolution des régions riches en répétitions des génomes de grande taille. Ces difficultés soulignent l'importance d'un investissement continu dans les technologies de séquençage et les outils bioinformatiques pour surmonter ces obstacles. L'intégration de nouvelles approches comme le séquençage Hi-C pour organiser les génomes à l'échelle chromosomique, offrent des opportunités prometteuses.

Les résultats de cette thèse ouvrent plusieurs perspectives pour les recherches futures. Ils invitent à approfondir les études sur les mécanismes adaptatifs, en particulier ceux liés aux expansions de gènes et aux diversifications protéiques, ainsi qu'à explorer les bases génétiques de l'immunité des espèces. Ces données constituent également une ressource clé pour la conservation génomique, notamment en développant des marqueurs génétiques pour des programmes de gestion et de préservation des populations menacées.

En synthèse, ce travail fournit une base essentielle pour la compréhension des génomes des Décapodes. Il illustre comment la combinaison de technologies avancées et d'analyses comparatives peut relever les défis scientifiques liés à l'étude des espèces non-modèles, tout en ayant des retombées concrètes en biologie évolutive et en conservation des espèces.

Christelle RUTZ

Assembly of the giant genome of the noble crayfish and genome evolution of Decapoda

Résumé

Les écrevisses sont des crustacés décapodes jouant un rôle déterminant dans les écosystèmes d'eau douce mais beaucoup d'espèces européennes, en particulier l'écrevisse noble *Astacus astacus*, sont menacées par la peste de l'écrevisse. J'ai relevé le défi de l'assemblage du génome géant d'*A. astacus* (17 Gb) et obtenu un génome préliminaire de 21,8 Gb. Ce premier génome d'écrevisse européenne est l'un des plus grands génomes d'invertébrés disponibles. J'ai également exploré l'évolution des génomes de décapodes. J'ai pu montrer la contribution centrale des éléments répétés à l'expansion et la diversification des génomes de décapodes. La comparaison des protéomes a mis en évidence un core-proteome de 7 000 protéines et la diversité des histoires évolutives de leurs gènes. Mes travaux ouvrent la voie au séquençage et à la comparaison de nouveaux génomes d'écrevisses pour identifier des marqueurs de résistance à la peste et soutenir des stratégies de réintroduction et d'aquaculture durable.

Mots clés :
Génomique de la conservation, écosystèmes d'eau douce, éléments répétés, génomique comparative.

Abstract

Crayfish are decapod crustaceans that play a decisive role in freshwater ecosystems, but many European species, particularly the noble crayfish *Astacus astacus*, are threatened by the crayfish plague. I took on the challenge of assembling the giant *A. astacus* genome (17 Gb) and obtained a preliminary genome of 21.8 Gb. This first European crayfish genome is one of the largest invertebrate genomes available. I also explored the evolution of decapod genomes. I was able to show the central contribution of repeated elements to the expansion and diversification of decapod genomes. A comparison of proteomes revealed a core-proteome of 7,000 proteins and the diversity of the evolutionary histories of their genes. My work paves the way for the sequencing and comparison of new crayfish genomes to identify markers of resistance to the plague and support strategies for reintroduction and sustainable aquaculture.

Keywords:
Conservation genomics, freshwater ecosystems, repetitive elements, comparative genomics.