Université de Strasbourg

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

RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE

ICube – UMR 7357

RHEINLAND-PFÄLZISCHE TECHNISCHE UNIVERSITÄT KAISERSLAUTERN-LANDAU

Fachbereich Natur- und Umweltwissenschaften

Accepted DISSERTATION presented by**:**

Lena BONASSIN

Defended on: **19 November 2025**

to obtain the grade of**:**

Docteur de l'Université de Strasbourg, Doctor of Natural Sciences (RPTU)

Discipline (Specialty):

Doctorat Sciences de la vie & de la santé, Bioinformatique et biologie des systèmes

Genome characterisation of European freshwater crayfish

DISSERTATION supervisor:

Prof. Odile Lecompte	University of Strasbourg, France
Dr. Habil. Kathrin Theissinger	RPTU Kaiserslautern-Landau, Germany

RAPPORTEURS:

Prof. Tony Heitkam	RWTH Aachen University, Germany
Prof. Adam Petrusek	Charles University, Czech Republic

OTHER MEMBERS OF THE JURY:

Prof. Klaus Schwenk	RPTU University Kaiserslautern-Landau, Germany
Prof. Joseph Schacherer	University of Strasbourg, France

Lena Bonassin

Genome characterisation of European freshwater crayfish

Summary

Freshwater crayfish are an ecologically important group of decapod crustaceans with a vital role in freshwater ecosystems. Many of these species are, however, among the most threatened species worldwide, with declines and local extinction of many populations. Their conservation is therefore critical to mitigate the biodiversity loss and preserve the health of the entire ecosystem. Despite their important role, genomic resources are lacking due to the large and repeat-rich genomes. In this study I address this genomic gap by characterising the genome among freshwater crayfish families to inform management of threatened populations. Across Decapoda and freshwater crayfish species, I identified a large proportion of transposable elements (TE) and satellite DNA (satDNA). Freshwater crayfish species with large genomes showed a particularly large content of satDNA. Based on the analysis of TE and satDNA, I identified species-specific patterns with a phylogenetic signal. To improve repeat detection, a new annotation pipeline was developed that detected 10% more repeats than the commonly used approaches. I also propose an optimised long-read sequencing workflow based on a salting-out DNA extraction protocol combined with PacBio sequencing. A combination of amplification-based, and amplification-free library preparation strategies was used to produce high-quality long-read sequencing data sufficient for *de novo* assembly. Both presented methodological advancements reduce the difficulties of working with challenging genomes. I further used genomic data to inform conservation actions in the endangered *Austropotamobius bihariensis*. A ddRADseq dataset revealed low heterozygosity (0.149 – 0.331), few private alleles (0 – 12), and small effective populations sizes (<100) with strong structuring. Together, my results highlight the need for conservation at a population scale to maintain the unique genetic makeup and mitigate the loss of variation through drift and inbreeding in *A. bihariensis*. Overall, this work provides a methodological framework and understanding of generating long-read sequencing data, TEs, satDNA and

genome-wide SNPs for genomic studies, comparative analyses and to translate genomic evidence into management actions for endangered species.

Résumé

Les écrevisses d'eau douce constituent un groupe de crustacés décapodes important sur le plan écologique, de par leur rôle essentiel dans les écosystèmes d'eau douce. Bon nombre de ces espèces figurent cependant parmi les espèces les plus menacées au monde, avec des déclins ou des extinctions locales de nombreuses populations. Leur conservation est donc essentielle pour limiter la perte de biodiversité et préserver la santé de l'ensemble de l'écosystème. Malgré leur rôle, les ressources génomiques font défaut pour les décapodes es décapodes en raison de la taille importante de leur génome et de sa richesse en éléments répétés. Dans cette étude, je réponds à ce manque de données génomique en caractérisant le génome des familles d'écrevisses d'eau douce afin d'éclairer la gestion des populations menacées. Chez les espèces de décapodes et d'écrevisses d'eau douce, j'ai identifié une grande proportion d'éléments transposables (TEs) et d'ADN satellite (satDNA). Les espèces d'écrevisses d'eau douce avec de grands génomes présentaient une teneur particulièrement élevée en satDNA. Sur la base de l'analyse des TEs et de l'ADN satellite, j'ai identifié des caractéristiques spécifiques à chaque espèce et mis un signal phylogénétique. Afin d'améliorer la détection des éléments répétés, un nouveau pipeline d'annotation a été développé, qui a permis de détecter 10 % d'éléments répétés en plus par rapport aux approches couramment utilisées. Je propose également une procédure optimisée pour le séquençage à lecture longue basée sur un protocole d'extraction d'ADN par "salting out" combiné au séquençage PacBio. Une combinaison de stratégies de préparation de librairies avec et sans amplification a été utilisée pour produire des données de séquençage à lecture longue de haute qualité, suffisantes pour un assemblage *de novo*. Ces deux avancées méthodologiques permettent de surmonter les difficultés liées à ces génomes particulièrement complexes. J'ai également généré et utilisé des données génomiques pour orienter les mesures de conservation de l'espèce menacée *Austropotamobius bihariensis*. Un ensemble de données ddRADseq a révélé une faible hétérozygotie (0,149 - 0,331), peu d'allèles privés (0 - 12) et de petites tailles de populations effectives (<100) avec une forte structuration. Dans l'ensemble, mes résultats soulignent la nécessité de mettre en œuvre des actions de conservation à l'échelle de la population afin de préserver la composition génétique unique et d'atténuer la perte de variation due à la dérive et à la consanguinité chez *A. bihariensis*. Globalement, ce travail fournit un cadre méthodologique et une meilleure compréhension de la génération de données de séquençage à lecture longue, de TEs, de satDNA et de SNPs à l'échelle du

génome, en vue d'études génomiques, d'analyses comparatives et pour traduire les données génomiques en actions de gestion pour les espèces menacées.

Acknowledgments

I would like to thank my supervisors Prof. Odile Lecompte, Dr.habil. Kathrin Theissinger and Prof. Klaus Schwenk for their support and opportunity to carry out this thesis. Thank you, Odile, for your guidance, for sharing your knowledge, your kind words and reassurance. Thank you for the always warm welcome in Strasbourg and the warm *pain au chocolat*. Kathrin, thank you for your advice and support throughout the thesis. Thank you for giving me the opportunity to work on different projects, your trust and encouragement helped me grow as a researcher.

I would like to extend my gratitude to Prof. Miklós Bálint for welcoming me into his group. I am grateful for the research stay before and during my PhD and the opportunity to carry out the majority of my experimental work in his lab.

Thank you to the ICUBE-CSTB members for their support and assistance. Thank you to the AG Stoeck for welcoming me into the group during the end of my thesis.

I would like to thank Prof. Lucian Pârvulescu for his dedication to crayfish research and for the opportunity to work on the idle crayfish for a large part of my thesis.

I am deeply grateful to Dr. Carola Greve and her team for the expert advice and continuous support during the genome sequencing project.

To Prof. Ivana Maguire and Assoc. Prof. Sandra Hudina, thank you for your unwavering support and collaboration throughout all these years.

I would like to thank all members and students of the FUG group for their support and for always having a free desk. Thank you for the shared time, and the fun memories and activities. Thank you to the TBG members for all the terrific TBG events.

I would like to extend a special thanks to all the colleagues who became friends during this journey.

Juliane, thank you for sharing this PhD journey with me, for being the first office-mate. Thank you for your patience with all the bioinformatic questions, thank you for your science advice, and sharing your R knowledge with me. The figures in this thesis would not be as they are without you. Thank you for sharing all the long hours and weekends during the last stretch of this thesis. Thank you for the friendship, for the shared experiences, the movie nights, board games and daily trips. Thank you for always taking the funniest pictures with the best memories.

Leonie, thank you for being the best lab manager, for all the advice and for always finding the solution. Thank you for having a plan B to Z. Thank you for always having time for truly listening and caring. Thank you for believing in me.

Johannes, thank you for introducing me to the must-know German phrases. Thank you for sharing this rollercoaster journey. Thank you for always finding a positive outlook and having coffee, milk, and healthy chocolate in your drawer.

Christelle, thank you for your support even if far away, thank you for all the advice and for patiently waiting through all the sequencing trials.

Caterina, thank you for always finding time to review my texts and thoughts, your comments and advice have always helped. Thank you for the hours-long talks, shared gossip and laughter.

Ljudevit, thank you for being by my side. Thank you for pushing me through the ups and downs of the research and for putting up with my millions of questions.

Thank you to Daniel and Ana for all the support, sharing their PhD experiences and advice, and the reminders to find time for activities outside the thesis.

Thank you, Yuma and Dea-Hyun, for welcoming me into your inner circle and making Frankfurt feel like home.

Thank you to all my friends GPS, Mašnica and Algebros: Goga, Pia, Donata, Martina, Valerija, Lea, Petar, Valerija, for keeping my spirits up and always asking for updates. Thank you for all the airport pick-ups, finding time to meet in my busy schedule and the online game rounds.

Hvala mojoj obitelji. Hvala Sara, mama, papà. Hvala na svakom ohrabrenju, hvala što me podržavate na svakom koraku.

Declaration

I here declare that I independently conducted the work presented in this thesis entitled "Genome characterisation of European freshwater crayfish". All used assistances are mentioned and involved contributors are either co-authors of or are acknowledged in the respective publication. This thesis has not been submitted elsewhere for an examination, as a thesis or for evaluation in a similar context to any other university or scientific institution. I am aware of and understand that a violation of the aforementioned conditions can have legal consequences. The Artificial Intelligence (AI) tool, DeepL (www.deepl.com) was used for the translation of the parts of the authors original text from English to French. No other uses of AI are reported for this thesis.

_____ _____

Place, date Signature

Table of contents

Summary I
Résumé III
Acknowledgments V
Declaration VII
Table of contents VIII
Figure list XII
Table list XVI
List of abbreviations XVII
I. General introduction 1
1.1. Decapod crustaceans and freshwater crayfish 1
Decapoda: a diverse and ecologically significant order 1
Global distribution and phylogeny of freshwater crayfish 3
Freshwater crayfish: keystone species in aquatic ecosystems 4
Decline of native freshwater crayfish populations in Europe 5
1.2. Genetic landscapes of endangered species 7
Genetic data in freshwater ecosystems 7
Genetic research on Decapoda and freshwater crayfish 8
1.3. Advancements in sequencing technologies and bioinformatic methodologies 9
Genome sequencing 10
Genome assembly 12
Genome annotation 13
1.4. Gaps in Decapoda and freshwater crayfish genomics 14
The unique challenges of genomes 15
Thesis objectives 23
Overview of the chapters 23
II. Chapter II 25
Abstract 26
Keywords 26
2.1. Introduction 26
2.2. Materials and Methods 29
2.2.1. Genomic Datasets 29
2.2.2. Identification and Annotation of Repetitive Elements 32

2.3. Results and Discussion 33
2.3.1. Construction of Repetitive Elements Reference 33
2.3.2. Annotation of Repetitive Elements in Decapoda Genomes 37
2.3.3. Proportion of Repetitive Elements in Decapoda Genomes 39
2.3.4. Correlation between Genome Size and Repetitive Elements 43
2.3.5. Frequency of satDNA Families Occurrence 45
2.3.6. Diversity of Repetitive Elements 46
2.3.7. Sequence Divergence Distribution of Transposable Elements 49
2.4. Conclusions 53
Funding 54
Acknowledgments 54
References 54
III. Chapter III 63
Abstract 64
Keywords 64
3.1. Background 64
3.2. Methods 67
3.2.1. Sampling and genomic DNA extraction 67
3.2.2. Flow cytometry analysis 68
3.2.3. Next-Generation Sequencing 69
3.2.4. Identification and annotation of repetitive DNA 69
3.2.5. Repeat classification and sequence analysis 70
3.2.6. Phylogenetic reconstruction and divergence analysis 70
3.2.7. Detailed characterisation of PlSAT3-411 and PlSAT57-664 71
3.2.8. Preparation of chromosome spreads and fluorescence *in situ* hybridisation (FISH) 71
3.3. Results 72
3.3.1. Repeat classification and sequence analysis 72
3.3.2. Phylogenetic reconstruction and divergence analysis 75
3.3.3. Detailed characterisation of PlSAT3-411 and PlSAT57-664 77
3.4. Discussion 81
3.4.1. High number of satellite DNA families in all freshwater crayfish species 81
3.4.2. SatDNAs among freshwater crayfish show shared characteristics 82
3.4.3. Phylogenetic signal of satDNA 83
3.4.4. Conserved satDNAs at the core of the freshwater crayfish satellitome 84
3.4.5. The challenges of studying satellitomes of non-model organisms 85
3.5. Conclusion 86

Funding 86
Acknowledgments 87
References 87
IV. Chapter IV 95
Abstract 96
Keywords 97
4.1. Introduction 97
4.2. Methods 100
4.2.1. Sampling of individuals and tissue 100
4.2.2. Genomic DNA extraction 100
4.2.3. Evaluation of DNA quantity and quality 102
4.2.4. Pre-extraction sorbitol washing complex homogenate protocol 102
4.2.5. Comparison of size selection protocols 102
4.2.6. Oxford Nanopore Technology library preparation 103
4.2.7. PacBio library preparation 103
4.2.8. Statistical analyses 105
4.3. Results 105
4.3.1. Total DNA concentration, purity and fragment length 106
4.3.2. Size selection 109
4.3.3. DNA sequencing 110
4.4. Discussion 112
4.5. Conclusion 117
Funding 118
Acknowledgment 118
References 118
V. Chapter V 123
Abstract 124
Keywords 124
5.1. Background 124
5.2. Methods 126
5.2.1. Sample collection and DNA extraction 126
5.2.2. ddRAD sequencing 127
5.2.3. ddRADseq data processing 128
5.2.4. Population genetic diversity 128
5.2.5. Population genetic structure 128
5.2.6. Ethical statement 129
5.3. Results 129

5.3.1. ddRAD data assembly 129
5.3.2. Population genetic diversity 130
5.3.3. Population genetic structure 132
5.4. Discussion 137
5.4.1. Population genetic diversity and structure 137
5.4.2. Conservation 139
5.5. Conclusion 141
Funding 141
Acknowledgement 142
References 142
VI. General discussion 149
6.1. Decapoda genome complexity is driven by repetitive elements 150
6.2. Optimised methodologies are crucial for studying freshwater crayfish genomes 156
6.3. Genomic approaches for informing conservation strategies of endangered freshwater crayfish 162
General conclusions 165
Conclusion générale 167
References 169
Appendix 189
Status and author contributions of publications included in the thesis 189
Supplementary material 192
Supplementary material - Chapter II. 192
Supplementary material - Chapter III 195
Supplementary material - Chapter IV 200
Supplementary material - Chapter V. 209
Curriculum vitae 210

Figure list

Figure I-1. Dorsolateral view of a generalised crayfish. Adapted from Cumberlidge et al., 2015..... 1

Figure I-2. Decapoda cladogram. Adapted from Wolfe et al., 2019. Organisms silhouettes from PhyloPic (phylopic.org) 2

Figure I-3. Global distribution and number of freshwater crayfish species. Adapted from Richman et al., 2015 3

Figure I-4. Double digest RAD sequencing (ddRADseq). Adapted from Peterson et al., 2012 9

Figure I-5. PacBio sequencing method (A) and HiFi read generation (B). Adapted from PacBio, 2021 11

Figure I-6. Nanopore sequencing technology. Adapted from Y. Wang et al., 2021 11

Figure I-7. Structure of satellite DNA. (A) SatDNA array organised in head-to-tail monomer units. (B) Higher order repeats (HOR). 18

Figure I-8. Satellite DNA evolution concepts: (A) SatDNA library concept and (B) concerted evolution. Adapted from Plohl et al., 2012..... 19

Figure I-9. Class, subclass, and superfamily classification of transposable elements. Adapted from Bourque et al., 2018 20

Figure II-1. Standardized annotation protocol for repetitive elements developed in this study. 33

Figure II-2. Proportion and content of repetitive elements in genomes. Percentage of repetitive ele-ments in the genome by class of repetitive elements. De, Dendrobranchiata; Ca, Caridea; Ac, Achelata; As, Astacidea; An, Anomura; Br, Brachyura; Oc, other Crustacea.41

Figure II-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 = 1.864×10^{-6}. (**B**). Correlation between assembly size and the percentage of TEs. Spearman rank correlation test: $\rho = 0.6$, p-value = 1.48×10^{-3}. (**C**). Correlation between estimated genome size and the load of TEs. Spearman rank correlation test: $\rho = 0.62$, p-value = 7.114×10^{-4}. (**D**). Correlation between

estimated genome size and the percentage of TEs. Spearman rank correlation test: ρ = 0.47, p-value =1.421 x 10^{-2} 45

Figure II-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. 46

Figure II-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. 48

Figure II-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. 50

Figure III-1. Proportions of repetitive elements in genomic reads of the 19 freshwater crayfish species used in this study. Each bar represents a species, with colours indicating different repetitive element categories. Repetitive elements are annotated as satellite DNA (satDNA), ribosomal DNA (rDNA), TEs belonging to Class I and Class II, and sequences without detailed annotation indicated as Unclassified repeats. Proportions are on a scale from 0 to 65 %. 73

Figure III-2. Frequency of satDNA sequences length across studied species. (A) Minisatellite DNA sequences ≤100 bp. (B) Macrosatellite DNA sequences >100 bp. Each colour represents a species. Vertical dashed lines represent the mean length value for each family. 74

Figure III-3. Density plot representing the GC content distribution for minisatellite and macrosatellite sequences by crayfish genus. Colours represent sequence type minisatellite (< 100 bp) and macrosatellite (> 100 bp) 75

Figure III-4. A) Total number of RE clusters per species identified in comparative analysis by Repeat Explorer2. B) Number of share clusters between species. C) Number of unique clusters per species. 76

Figure III-5. SatDNA sequence divergence landscapes for each species. Colours indicate different freshwater crayfish families. y-axes are scaled for each species. 77

Figure III-6. Variant repeat profiles of PlSAT3-411 satellite DNA family across the studied species. Different colours indicate A, T, C and G bases. The height of each bar indicates the coverage of each variant in the reads. 78
Figure III-7. Sequence divergence landscapes for each species for the PlSAT3-411 (left) and PlSAT57-664 (right) sequences. Colours indicate different species. y-axis is scaled for each species. 80
Figure III-8. Localisation of PlSAT3-411 satellite repeat family (in red) on metaphase chromosomes of (A) *A. torrentium* and (B) *P. leniusculus*. Red signals represent the Cy3 probe localisation, chromosomes are counterstained with DAPI. Scale bar = 10μm. 81
Figure IV-1. Workflow of sample processing from tissue to sequencing. The Sankey diagram illustrates the flow of samples through various stages of processing for different tissue types. Ribbons represent the sample pathways, originating from three tissue types (Claw, Leg muscle, Abdomen muscle). Samples then proceed through different DNA extraction methods (Sorbitol wash & Salting-out, Salting-out, MagAttract HMW, Phenol-chloroform, Nanobind Big DNA kit, Qiagen DNeasy). Following extraction, samples proceed to library preparation methods (PacBio amplification-based kit, PacBio amplification free kit, Nanopore transposase-based kit and Nanopore ligation kit), leading to sequencing. The width of the ribbons is proportional to the number of samples following that specific path. The numbers in parenthesis indicate the number of samples. 106
Figure IV-2. Violin dot plot of total DNA yield (ng) from three extraction methods. Colour indicates tissue types used for DNA extraction. 107
Figure IV-3. Absorbance ratio A260/280 (left) and A260/230 (right) for three extraction methods. Colour indicates tissue types used for DNA extraction. Dashed lines indicate optimal absorbance ratio values. 108
Figure IV-4. DNA fragment length (bp) at which the highest concentration of DNA extract was observed for three extraction method. A) Zoomed in y-axis 0-25000 bp and B) y axis 0-200000 bp. Colour indicates tissue types used for DNA extraction. 109
Figure IV-5. DNA fragment distribution electropherograms from a Femto pulse run using the genomic DNA 165 kb kit (lower marker at 1 bp, ladder range from 1 bp to 165 kb). Colours indicate original DNA sample (pink), AMPure PB bead (purple) and BluePippin (yellow) size selected sample. 110
Figure IV-6. HiFi read length (A) and HiFi yield (B) separated by library preparation method. Colours indicate samples used for library preparation, while shape indicates the polymerase used in the ultra-low input library preparation. 111

Figure IV-7. Sequencing metrics P0, P1 and P2 and correlation of P1 % with HiFi yield (Gb) and HiFi read length (bp) per sample for Low input and Ultra low input libraries. Panels A and C display heatmaps of sequencing metrics P0, P1, and P2 (%) for low input and ultra-low input libraries, respectively. Panels B and D show correlation of P1 % with HiFi sequencing yield (in Gb) and HiFi read length (bp), respectively. 112

Figure V-1. Distribution map of the sampled locations. Colours denote different river basins and arrows the direction of river flow. Population acronyms: DUD – Duduşoaia, COR – Corbului, TAL – Tâlniciorii, BOG – Boga, RAC – Racu, CUT – Cuţilor, IAD – Iadei, PRE – Preluca, ANI – Anişelului, MAR – Mare, STE - Starpă, BIS – Bistrii. The base map layout is provided by Earthstar Geographics (https://www.terracolor.net) and river basin boundaries were delineated using the HydroBASINS database (https://www.hydrosheds.org)..... 130

Figure V-2. Fixation coefficient (F_{ST}) between each population pair. Darker blue indicates higher F_{ST} values. Population acronyms: DUD – Duduşoaia, COR – Corbului, TAL – Tâlniciorii, BOG – Boga, RAC – Racu, CUT – Cuţilor, IAD – Iadei, PRE – Preluca, ANI – Anişelului, MAR – Mare, STE – Starpă, BIS – Bistrii. 133

Figure V-3. Population structure based on fastStructure analysis for K=5, 8, and 12. Different colours represent different genetic clusters. Each column represents one individual. Population acronyms: DUD – Duduşoaia, COR – Corbului, TAL – Tâlniciorii, BOG – Boga, RAC – Racu, CUT – Cuţilor, IAD – Iadei, PRE – Preluca, ANI – Anişelului, MAR – Mare, STE – Starpă, BIS – Bistrii. Columns with different colours indicate admixture of populations. 134

Figure V-4. PCA. Different colours denote different populations, and different symbols the river basins to which populations belong. A – PC1 and PC2, B – PC1 and PC3. Population acronyms: DUD – Duduşoaia, COR – Corbului, TAL – Tâlniciorii, BOG – Boga, RAC – Racu, CUT – Cuţilor, IAD – Iadei, PRE – Preluca, ANI – Anişelului, MAR – Mare, STE – Starpă, BIS – Bistrii. 135

Figure V-5. Co-ancestry matrix of pairwise genetic similarity between the individuals. Darker (blue and black) colours represent high level of genetic similarity and co-ancestry (relatedness), and light colours lower level of co-ancestry. The clustering of individuals is shown in a dendrogram on top of the matrix. Posterior probabilities values are 1 unless indicated on branches. Colour bars indicate river basins and populations. Population acronyms: DUD – Duduşoaia, COR – Corbului, TAL – Tâlniciorii, BOG – Boga, RAC – Racu, CUT – Cuţilor, IAD – Iadei, PRE – Preluca, ANI – Anişelului, MAR – Mare, STE - Starpă, BIS – Bistrii. 136

Table list

Table II-1. Genomic dataset used in this study..... 30

Table II-2. Number of RE libraries identified and annotated using species-specific libraries or a merged library from all species. RMo—RepeatModeler2, Tp—TAREAN pipeline.. 35

Table V-1. River basins and populations used in genetic analyses, number of individuals per population, percentage of polymorphic loci, number of private alleles, observed (H_O) and expected (H_E) heterozygosity and inbreeding coefficient (F_{IS}) 131

Table V-2. Number of individuals per river basin, percentage of polymorphic loci, number of private alleles, observed (H_O) and expected (H_E) heterozygosity and inbreeding coefficient (F_{IS}) of individuals grouped by river basins. 131

Table V-3. Effective population size estimation based on linkage diseqilibrium (N_eLD) and heterozygote excess (N_eb) for 0.02 and 0.01 minor allele frequency (MAF) and 95% CI based on jackknifing method – N_e estimator. Population acronyms: DUD – Duduşoaia, COR – Corbului, TAL – Tâlniciorii, BOG – Boga, RAC – Racu, CUT – Cuților, IAD – Iadei, PRE – Preluca, ANI – Anişelului, MAR – Mare, STE – Starpă, BIS – Bistrii..... 132

List of abbreviations

3C - chromosome conformation capture

ART ANOVA - aligned rank transform ANOVA

CCS - circular consensus sequencing

CENP-B - centromere protein-B

Ci - confidence interval

CLR - continuous long reads

CR1 - chicken repeat 1

CRE - cis-regulatory elements

CTAB - cetyltrimethylammonium-bromide

ddRADseq - double digest restriction site associated DNA sequencing

Df - degrees of freedom

Df res - residual degrees of freedom

DIRS - *Dictyostelium* intermediate repeat sequence 1

DToL - Darwin Tree of Life

eccDNA - extrachromosomal circular DNA

ENA - European Nucleotide Archive

ERGA - European Reference Genome Atlas

Fis - inbreeding coefficient

FISH - fluorescence *in situ* hybridisation

Fst - fixation index

GBS - genotyping-by-sequencing

GWAS - genome-wide association studies

H_E - expected heterozygosity

Hi-C - high-throughput chromosome conformation capture

HiFi - high fidelity reads

HMW - high molecular weight

H_O - observed heterozygosity

HOR - higher order repeat

HT - horizontal transfer

HTT - horizontal transfer of transposable elements

ICS - indigenous crayfish species

LD - linkage disequilibrium

LINE - long interspersed nuclear element

lncRNA - long non-coding RNA

LTR - long terminal repeats

MAF - minor allele frequency

MIR - mammalian-wide interspersed repeat

MITE - miniature inverted-repeat transposable elements

Mya - Million years ago

N_e - effective population size

N_{eb} - effective population size based on heterozygote excess

N_{eLD} - effective population size based on linkage disequilibrium

NGS - next generation sequencing

NICS - non-indigenous crayfish species

ns - not significant

ONT - Oxford Nanopore Technologies

PacBio - Pacific Biosciences

PCA - principal component analysis

PI - propidium iodide

PLE - Penelope-like elements

PVP - polyvinylpyrrolidone

RADseq - Restriction-site Associated DNA sequencing

rDNA - ribosomal DNA

RE - repeat element

RMo - RepeatModeler2

RNAseq - RNA sequencing

ROH - runs of homozygosity

RRS - reduced representation sequencing

satDNA - satellite DNA

SINE - short interspersed nuclear element

siRNA - small interfering RNA

SMRT - Single-molecule real-time

SNP - single nucleotide polymorphism

TE - transposable element

TIR - terminal inverted repeat

Tp - TAREAN pipeline

VGP - Vertebrate Genome project

WGS - whole genome sequencing

WSSV - white spot syndrome virus

ZMW - zero-mode waveguide

General introduction

1.1. Decapod crustaceans and freshwater crayfish

Decapoda: a diverse and ecologically significant order

Decapoda is a large order within the subphylum Crustacea that encompasses shrimps, crayfish, lobsters, and crabs. The order includes over 17000 species in marine, freshwater and semiterrestrial habitats across all continents, except Antarctica (Cumberlidge et al., 2015). The number of species is in constant change as new species are being discovered and taxonomically revised (De Grave et al., 2023). Decapoda have the greatest morphological diversity of all orders of crustaceans, which allowed their spread in diverse habitats. Nevertheless, the common characteristics among all Decapoda are extensive body segmentation, five pairs of jointed appendages and an exoskeleton composed of polysaccharides and inorganic salts (Cumberlidge et al., 2015). The general decapod body consists of a head, thorax, and abdomen. In shrimps, crayfish and crabs, the head and the thorax are fused into a cephalothorax (Figure I-1). The appendages are modified for different functions: sensory, feeding, locomotion, copulation, and swimming (Cumberlidge et al., 2015).

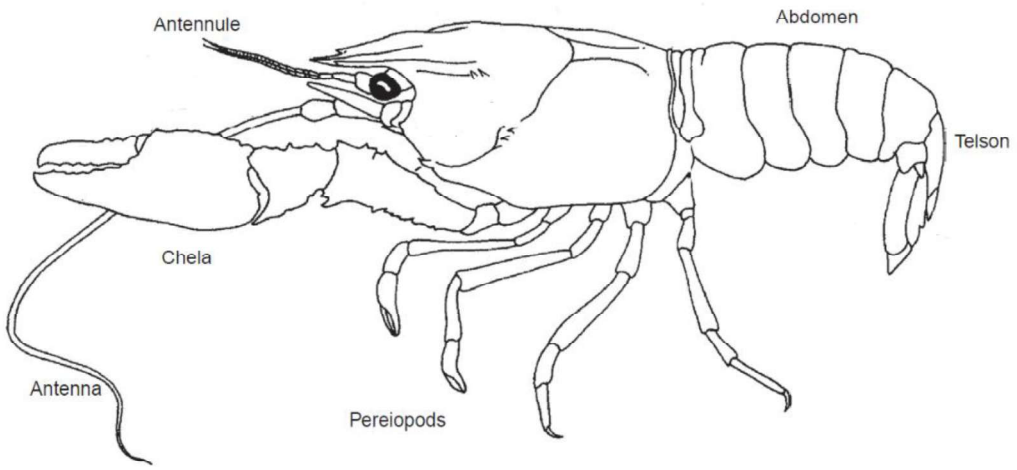


Figure I-1. Dorsolateral view of a generalised crayfish. Adapted from Cumberlidge et al., 2015

Decapoda play crucial roles in the ecosystem through their activities of burrowing, feeding and movement (Reynolds et al., 2013). They are primarily nocturnal, hiding during the day and foraging at night. Key contributions to the ecosystem include nutrient cycling and detritus processing, sediment bioturbation through burrowing and the structuring of benthic

communities. As ecosystem engineers, they can alter substrate composition, water flow and nutrient availability in aquatic and terrestrial environments (Albertson & Daniels, 2018). Their ecological roles are varied, as they enter the food webs as both primary consumers and prey for a wide range of predators. Predators of Decapoda include fish, amphibians, snakes, birds, and mammals, while their diet consists of a variety of plant and animal material, microorganisms in the substrate, and insects (Lavalli & Spanier, 2016).

The order Decapoda contains two suborders: Dendrobranchiata (shrimps and prawns) and Pleocyemata (true crabs, lobsters, and crayfish). The suborder Pleocyemata includes several infraorders. Stenopodidea (boxer shrimps), Procarididea and Caridea (true shrimps) form one group within the suborder. The second group, Reptantia, comprises the infraorders Achelata (spiny lobsters), Polychelida (benthic crustaceans), Astacidea (true lobsters and crayfish), Axiidea (mud shrimps), Gebiidea (mud lobsters and mud shrimps), Anomura (hermit crabs, squat lobsters, and king crabs), and Brachyura (true crabs). The different infraorders can be found in different marine, freshwater, and estuarine environments (Figure I-2) (Wolfe et al., 2019).

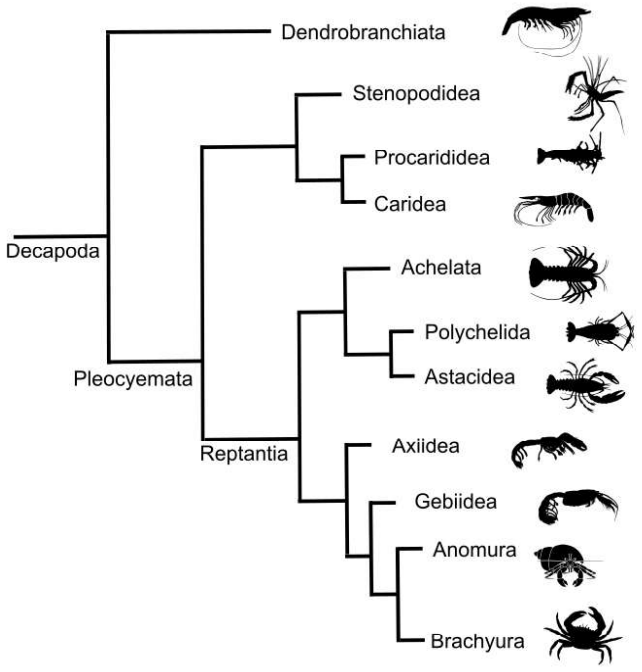


Figure I-2. Decapoda cladogram. Adapted from Wolfe et al., 2019. Organisms silhouettes from PhyloPic (phylopic.org)

Global distribution and phylogeny of freshwater crayfish

The infraorder Astacidea (freshwater crayfish) comprises two superfamilies: Astacoidea in the northern hemisphere and Parastacoidea in the southern hemisphere. The two superfamilies diverged from marine species between 320 and 290 Mya (Wolfe et al., 2019) and the split between the superfamilies occurred around 185 Mya with the break-up of Pangea (Crandall & Buhay, 2007). The superfamily Astacoidea includes three families, Astacidae, Cambaridae and Cambaroididae, while the family Parastacidae is the only family within the Parastacoidea superfamily. Freshwater crayfish families are distributed across the hemispheres: the Astacidae in Europe and western part of North America, Cambaridae in the eastern part of North America, Cambaroididae in eastern Asia and Parastacidae in South America, Madagascar, and Australia (Figure I-2) (Kawai & Crandall, 2016). The radiation within Parastacoidea occurred following the separation of the South American continent around 165-140 Mya, followed by Madagascar around 160 -120 Mya and finally the split of New Zealand from Australia around 80 Mya (Toon et al., 2010). Within the superfamily Astacoidea, Cambaroididae is the oldest family with a separation around 110 Mya, while the Cambaridae family is considered the youngest freshwater crayfish family with a radiation around 57 Mya (Wolfe et al., 2019).

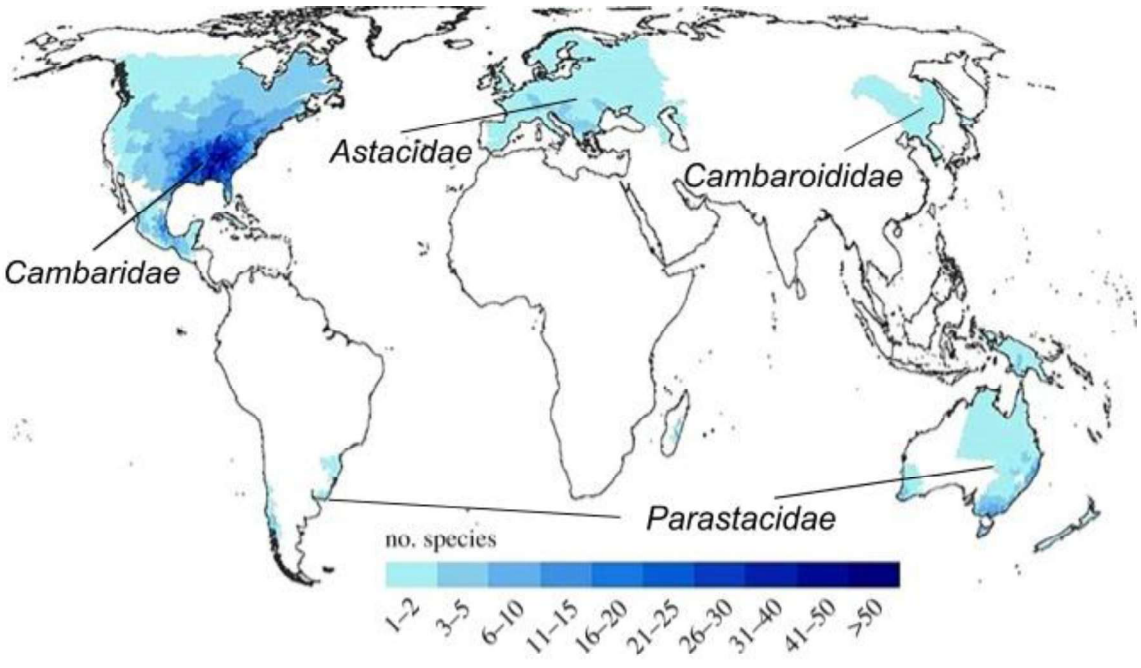


Figure I-3. Global distribution and number of freshwater crayfish species. Adapted from Richman et al., 2015

Globally there are two centres of diversity: one in the southeastern United States and one in southeastern Australia (Figure I-3). The majority of the ~650 freshwater crayfish species belong to the family Cambaridae with more than 400 species, with the most species-rich and widespread genera being *Procambarus* (167 species) and *Cambarus* (118 species). It has been hypothesised that the speciation of the family Cambaridae is driven by changes in river drainage patterns and glacial periods during the Pleistocene (Crandall et al., 1999; Cumberlidge et al., 2015). The second highest number of species is in the family Parastacidae. The diversity within this family is highest in southeastern Australia, with the *Euastacus* (53 species) and *Cherax* (52 species) genera being the most species-rich, while the most species-rich genus in South America is *Parastacus* (11) (Crandall & De Grave, 2017). Their radiation is a reflection of ancient continental drift (165 – 121 Mya) and more recently (6 Mya) of intense glaciation periods (Toon et al., 2010; Victoriano & D'Elía, 2021). The family Astacidae is mostly distributed across the Eurasian continent, while only the genus *Pacifastacus* is native to North America. This genus is basal to all other genera of the family Astacidae, however its phylogeographical placement has not been elucidated (Bracken-Grissom et al., 2014). The spread of genera present in Europe was highly influenced by glaciation during the last glacial maximum 29000 to 19000 years ago. Glacial refugia were found in southern Europe which became the centre of diversity from which the spread towards Northern Europe began. (Gross et al., 2021; Lovrenčić et al., 2020). The family Cambaroididae is the smallest family, comprising only one genus *Cambaroides* and six species. Its distribution is limited to eastern Russia, China, Korea, and Japan (Crandall & De Grave, 2017).

Freshwater crayfish: keystone species in aquatic ecosystems

Freshwater crayfish inhabit streams, rivers, lakes, and marshes with only a few species found in brackish waters. They are living within the same aquatic environment in all life stages and are dependent on permanent freshwater (Kawai & Crandall, 2016; Reynolds et al., 2013). When fluctuations of water levels occur, they can construct burrows with water at the bottom and humid air above it. They have an important role and impact within the ecosystem. They are defined as keystone species, surrogate species for water quality and flagship species for management decisions. Their key roles in environments include keystone trophic regulators and ecological engineers (Reynolds et al., 2013). Furthermore, they are considered water quality indicators and biodiversity indicators. They are dependent on habitat heterogeneity and water quality, and often adapted to specific habitats in terms of temperature, oxygen and

salinity (Füreder & Reynolds, 2003). Most freshwater crayfish require good quality of water where they search for refuges or create burrows. Astacidae and most Cambaridae prefer streams and lakes, while the minority of Cambaridae and Parastacidae species live in warm water ponds, ditches, and swamps (Crandall & Buhay, 2007).

Like many freshwater Decapoda, crayfish are mostly detritivores or omnivores, with some species being predominantly carnivorous or vegetation feeding (Reynolds et al., 2013). Crayfish have a clear impact on their habitat and on the structure of food webs by feeding on vegetation and invertebrates. Due to selective predation, they can alter the invertebrate composition and can eliminate macrophyte species due to intensive grazing (Olsson, 2008). Crayfish are also an important prey resource for larger predators such as fish, amphibians, reptiles, birds and mammals (Reynolds et al., 2013). The impact of crayfish on their ecosystem is clearly seen when a change in density or range occurs. Translocated crayfish, outside their native range can reach higher densities, altering the food-web interaction by predation and competition (Herrmann et al., 2022). As ecosystem engineers, the feeding activities and movement of crayfish affect the sediment in streams and the presence of resources for other organisms (Gherardi et al., 2010).

Decline of native freshwater crayfish populations in Europe

In Europe there are six indigenous crayfish species (ICS): the noble crayfish *Astacus astacus* (Linnaeus, 1758), the thick-clawed crayfish *Pontastacus pachypus* (Rathke, 1837), the narrow-clawed crayfish *Pontastacus leptodactylus* (Eschscholtz, 1823), the idle crayfish *Austropotamobius bihariensis* Pârvulescu, 2019, the white-clawed crayfish *Austropotamobius pallipes* (Lereboullet, 1858) and the stone crayfish *Austropotamobius torrentium* (Schrank, 1803). Because of the decreasing population numbers, the IUCN red list of threatened species lists *A. astacus* as vulnerable, and *A. bihariensis* and *A. pallipes* as endangered, while *A. torrentium* is listed as a priority species under the EU Habitats Directive on the conservation of natural habitats and of wild fauna and flora (Council Directive 92/43/EEC, 2007) . These species are impacted by the introduction of alien species and the spread of the crayfish plague (Richman et al., 2015). Specifically, in the case of the endemic *A. bihariensis*, the main threats are its small geographic range, restricted to the Apuseni mountains in Romania, and anthropogenic influence (urbanisation, forestry activities, pollution, reduction of habitat) (Ion et al., 2024). The general decline of European freshwater crayfish species is caused by climate change, extreme droughts, habitat alterations, introduction of diseases and water pollution (Tarandek et al., 2023). Current models estimate 87% of total freshwater crayfish

species being sensitive, and 80% of European species vulnerable to climate change mainly because of their habitat specialisation and range limitation (Hossain et al., 2018).

Beyond their ecological role, crayfish have a great cultural and commercial significance (Jussila, Edsman, et al., 2021). Because they are favoured in aquaculture and ornamental trade, they are among the most widely translocated aquatic invertebrates (Bláha et al., 2022; Jussila, Edsman, et al., 2021; Kouba et al., 2014). The long-distance (often cross continental) translocation facilitated the spread of non-indigenous crayfish species (NICS) and their diseases, most notably the crayfish plague and its causative pathogen *Aphanomyces astaci* Schikora, 1906 from North America (Jussila, Edsman, et al., 2021). Because of the role of crayfish as keystone species in their native habitats, their translocation can have disrupting consequences in the non-native habitat.

In European freshwaters, NICS are usually outcompeting and displacing ICS. This is facilitated through higher fecundity and faster population growth, higher dispersal rate and higher robustness than ICS (Holdich et al., 2009; Hudina et al., 2014). Unlike ICS, when in a non-native habitat, NICS can tolerate changes in the habitat like lower water quality, increased salinity, and decreased oxygen levels, allowing their successful spread in freshwater systems (Carvalho et al., 2022; Holdich et al., 2009; Soto et al., 2023). The adaptability and greater physiological plasticity are a result of NICS evolution in more dynamic and fluctuating environments compared to the pristine habitats of ICS (Marn et al., 2022). These characteristics raise the invasion potential and success of NICS. Moreover, NICS are often carriers of the crayfish plague pathogen, the oomycete *Aphanomyces astaci*, particularly deadly when infecting European crayfish. Since its first introduction in Europe in the 19th century, the pathogen has spread quickly causing mass mortalities and collapses of crayfish population throughout the continent (Alderman, 1996). The combination of competitive exclusion, environmental tolerance and disease transmission overwhelms native crayfish populations leading to their decline and local extinction. Consequently, there is the disappearance of other freshwater species through the impact on food webs and the collapse of the entire ecosystem (Lee et al., 2023). Several NICS have been introduced into Europe, mainly from North America and Australia (Holdich et al., 2009; Kouba et al., 2014; Laffitte et al., 2023). The NICS originating from North America with established populations in Europe are the signal crayfish *Pacifastacus leniusculus* (Dana, 1852), the red swamp crayfish *Procambarus clarkii* (Girard, 1852), the marbled crayfish *Procambarus virginalis* Lyko, 2017, the white river crayfish *Procambarus acutus* (Girard, 1852), the spiny-cheek crayfish

Faxonius limosus (Rafinesque, 1817), the calico crayfish *Faxonius immunis* (Hagen, 1870), the Kentucky River crayfish *Faxonius juvenilis* (Hagen, 1870), the virile crayfish *Faxonius virilis* (Hagen, 1870), and the rusty crayfish, *Faxonius rusticus* (Girard, 1852). Australian NICS established in European freshwaters are the Australian red claw crayfish *Cherax quadricarinatus* (Marten, 1868) and the common yabby *Cherax destructor* (Clark, 1936). Adding to the already established populations of NICS, there is a continuous escape of new species through aquarium and pet trade, which is challenging the biodiversity in European freshwaters.

1.2. Genetic landscapes of endangered species

Genetic data in freshwater ecosystems

Freshwater ecosystems are among the most diverse ecosystems on the planet, supporting 10% of all known species (Williams-Subiza & Epele, 2021). However, in the last decades, 35% of freshwater areas worldwide were lost, thereby accelerating the global biodiversity decline (Sayer et al., 2025). Yet, freshwater ecosystems have not been prioritised as marine and terrestrial habitats in global management actions. Based on the IUCN Red List of Threatened Species one quarter of freshwater species are threatened with extinction and have thus received IUCN classification as extinct in the wild, critically endangered, endangered, or vulnerable. The group with one of the highest extinction threats are Decapoda, with 30% of species being at risk. Additionally, 39% of Decapoda species are classified as data deficient, potentially increasing the number of species at risk (Sayer et al., 2025). Species' assessment is based on population size and decline, geographic range, and probability of extinction within the next century (IUCN, 2001). However, there are growing recommendations for including genetic and genomic data in the assessment of species' threat status (McLaughlin et al., 2025).

The preservation of a species' adaptive potential relies on the conservation of its genetic diversity. Key metrics used in the assessment of a species' genetic health status are heterozygosity (H), nucleotide diversity (π), haplotype diversity (h), runs of heterozygosity (ROH), allelic richness, effective population size (N_e) and fixation index (F_{ST}) (McLaughlin et al., 2025). These metrics can be studied at the individual or population level and can be measured from single or multiple genetic markers, or whole genomes (Kardos et al., 2021).

Single genetic markers often include mitochondrial DNA and ribosomal DNA genes and are used in species identification (barcoding), understanding geographic patterns of genetic

diversity and evolutionary relationships (Raza et al., 2016). As multiple marker approach, microsatellites are among the most widely used genetic markers for population level genetic characterisation (Abdul-Muneer, 2014). Microsatellites are short interspersed repetitive DNA sequences found throughout the genome, especially in non-coding regions. They are useful for studying population structure, genetic mapping and evolutionary processes because of their high mutation rate and analyses reproducibility among related species (Mason, 2015; Vieira et al., 2016). Genome-wide markers are single nucleotide polymorphisms (SNP). They are markers of choice for genetic diversity studies, genome wide association studies, genomic selection, and evolutionary history studies (Yirgu et al., 2023). SNP loci can be found in both coding and non-coding genomic regions which, alongside their abundance, allows their application in a variety of studies at a relatively low cost (Wenne, 2023). The most comprehensive view on genomic diversity is provided through the analysis of whole reference genomes, which allow the understanding of both coding and non-coding regions, including repetitive regions (Formenti et al., 2022; Theissinger et al., 2023).

Genetic research on Decapoda and freshwater crayfish

The use of whole genome sequencing (WGS) for conservation management practices in non-model organisms has become common practice only in the last decade. With the increase of SNPs and WGS studies, the number of studies using microsatellites decreased (Hogg et al., 2022). However, when a reference genome is not available, the commonly used approach remains microsatellites, or SNPs obtained from reduced representation sequencing approaches (RRS). RRS approaches include genotyping by sequencing (GBS), restriction-site associated DNA sequencing (RADseq) and double digest DNA restriction site-associated DNA sequencing (ddRADseq). RRS approaches use restriction enzymes to digest the genome into fragments, thus reducing the complexity of the genome and making it more cost-effective to generate a large number of SNPs across many individuals (Baird et al., 2008). GBS uses a single enzyme to digest DNA which is then sequenced. This method is very high-throughput but prone to high rates of missing data due to uneven sequencing coverage (Scheben et al., 2017). RADseq and ddRADseq approaches use an additional step of DNA fragment size selection which ensure consistent marker distribution, however the two methods are more complex and expensive than GBS. ddRADseq includes the use of two different restriction enzymes, a rare and common cutter, generating DNA fragments of various size. Following the enzymatic digestion, a size selection is performed to isolate fragments in a specific range and PCR amplified (Puritz et al., 2014) (Figure I-4).

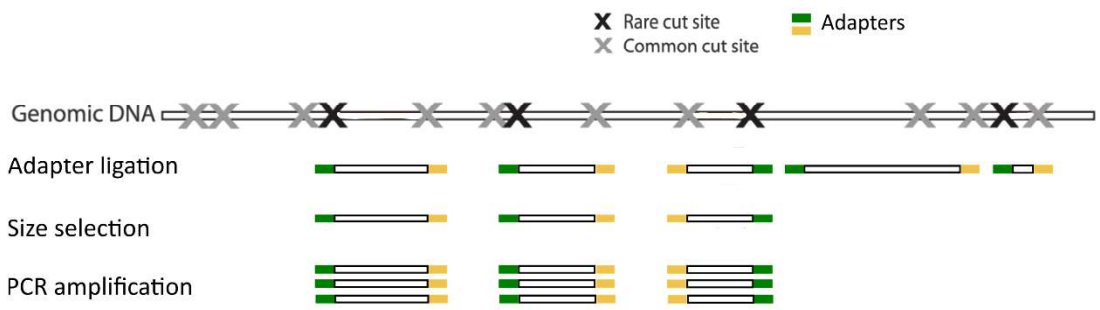


Figure I-4. Double digest RAD sequencing (ddRADseq). Adapted from Peterson et al., 2012

Microsatellites have been widely used to investigate genetic variation between and within freshwater crayfish populations providing insights into genetic diversity and structure, admixture, and gene flow thereby contributing critical data for conservation initiatives. These studies show high genetic diversity in south-eastern Europe, compared to the lower genetic diversity in central and Northern Europe for *A. astacus*, *A. torrentium* and *A. pallipes* (Berger et al., 2018; Laggis et al., 2017; Lovrenčić et al., 2022; Schrimpf et al., 2014). Microsatellites have also been used in studies aiming to identify populations suitable for selection in breeding programs and aquaculture (Liu et al., 2023). Studies using whole genome SNPs are rare within freshwater crayfish, mainly due to their large genomes and restricted resources for their sequencing.

1.3. Advancements in sequencing technologies and bioinformatic methodologies

In the last years, with improvement of sequencing technologies and with the increased number of global genome consortia, the number of sequenced genomes is increasing (Hogg et al., 2022). These genomes and their downstream applications can improve the knowledge about the species biology, evolutionary processes and functional variation (Hogg et al., 2022; Theissinger et al., 2023). For species of conservation concern, the use of genomic tools has helped decreasing biodiversity loss (Paez et al., 2022). Such tools include the reference genome, a highly contiguous, accurate and annotated genome assembly (Formenti et al., 2022). In the case of the European ash *Fraxinus excelsior*, genome sequencing of the ash tree and its fungal pathogen revealed traits associated with reduced susceptibility of the trees to the pathogen, which is crucial for breeding programmes to restock forests (Sollars et al., 2017; Theissinger et al., 2023). Genomic studies using a reference genome of the Florida

panther *Puma concolor* revealed reduced genetic variation and high inbreeding levels (Johnson et al., 2010). Translocation programs effectively decreased phenotypic defects and the population size increased (Supple & Shapiro, 2018). Still, globally, less than 3% of species listed as threatened by the IUCN Red List have a reference genome, which means that genomic resources for most critically endangered species are still lacking (Hogg et al., 2022).

Genome sequencing

Many of the sequencing efforts in producing reference genomes have been conducted using first- and second-generation genome sequencing technologies (Paez et al., 2022). First generation sequencing technologies includes the Sanger chain-termination dideoxy technique (Sanger et al., 1977). This low-throughput sequencing technology produces reads less than one kilobase (kb) in length (Heather & Chain, 2016). The development of second-generation sequencing technologies allowed to increase the throughput and reduce the costs of sequencing. The most successful among second-generation sequencing technologies is the Solexa/Illumina sequencing (Greenleaf & Sidow, 2014). It uses a sequencing by synthesis methodology using reversible terminator nucleotides, and a bridge amplification method which allows paired-end sequencing to occur simultaneously for thousands of molecules (Heather & Chain, 2016). Illumina reads are typically 50 bp to 300 bp long and have a 99,9 % accuracy, meaning there is a one incorrect base in 1000 (G. Tan et al., 2019). Third-generation DNA sequencing allows real-time sequencing of single molecules without the requirement of DNA amplification. The most widely used third-generation technologies are Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). These technologies allow to sequence long DNA molecules up to hundreds of kb with high accuracy (Espinosa et al., 2024).

PacBio sequencing utilises single molecule real time (SMRT) sequencing where a DNA polymerase is immobilised at the bottom of a zero-mode waveguide (ZMW) well (Figure I-5). Sequencing proceeds with the detection of fluorescently labelled nucleotides incorporated by the DNA polymerase in each ZMW (Espinosa et al., 2024). The first sequencing mode of PacBio is the Continuous Long Read (CLR) sequencing which generates extremely long reads, with tens of kilobases in length. The maximum length is limited by the length of the input DNA molecule and the polymerase activity. A major drawback of the CLR sequencing mode is the 5 – 1 % error rates (Rhoads & Au, 2015). A much higher accuracy is obtained with the Circular Consensus Sequencing (CCS) strategy where a consensus sequence is derived from multiple passes of a circular template molecule from a

single ZMW (Schell et al., 2025). Such circular templates, called SMRTbell libraries, are created by ligating hairpin adapter molecules to the DNA fragments ends. The adapters contain sequences that serve as starting points for the polymerase during sequencing. Using CCS strategy to generate high-fidelity (HiFi) reads, results in consensus sequences with 99,9 % accuracy, making them comparable to the accuracy of short read Illumina sequencing (Espinosa et al., 2024; Wenger et al., 2019).

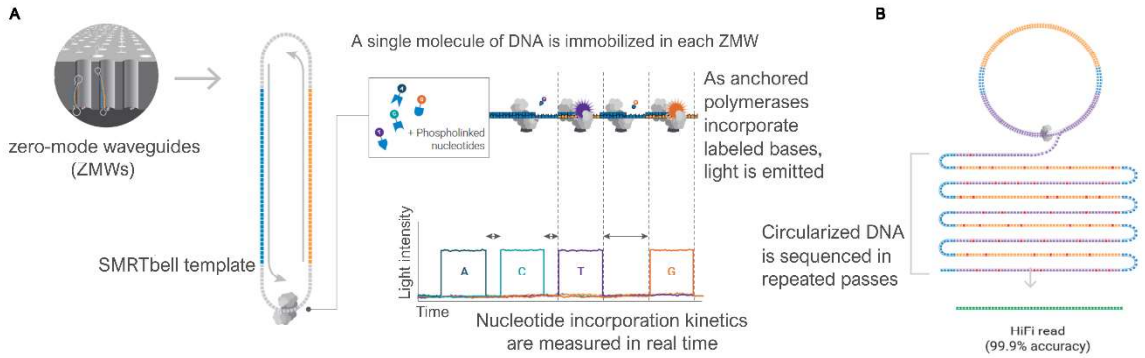


Figure I-5. PacBio sequencing method (A) and HiFi read generation (B). Adapted from PacBio, 2021

ONT Nanopore sequencing technology relies on a DNA molecule passing through a nanopore, tiny protein channel, which causes changes in the electrical current (Figure I-6). These changes are then decoded computationally during sequencing into nucleotides and DNA sequences (Jain et al., 2016). Nanopore sequencing is currently producing the longest sequence reads among all technologies with 99,75 % accuracy and is therefore suitable to tackle complex genomic regions, repetitive sequences and structural variations which could span large proportions of the genome (Espinosa et al., 2024; Schell et al., 2025).

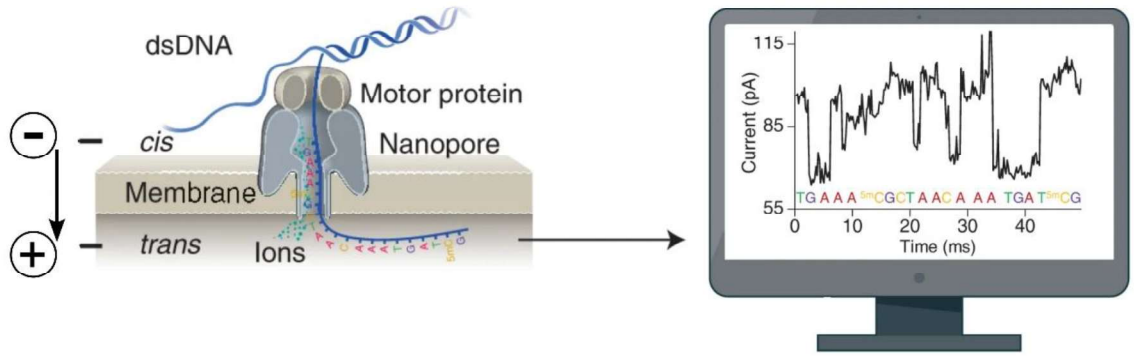


Figure I-6. Nanopore sequencing technology. Adapted from Y. Wang et al., 2021

DNA sequencing technologies have varying requirements for DNA input. Differences in DNA purity and fragment length are key considerations when choosing a sequencing methodology (Bista & Lino, 2025; Trigodet et al., 2022). All sequencing technologies require high-purity DNA, free from RNA, protein, and chemical contaminants (Dahn et al., 2022). Short-read Illumina sequencing is tolerant of minor impurities, however inhibitors can still reduce the sequencing quality (Hess et al., 2020). Long-read sequencing technologies, PacBio and ONT, are much more sensitive to contaminants. Both the polymerase (in PacBio sequencing) and the nanopore (in ONT sequencing) are affected by contaminants, leading to low sequencing quality and failed sequencing. The required DNA fragment length is the major factor distinguishing between the two sequencing technologies. In short read sequencing, genomic DNA is fragmented during library preparation to 300 – 500 bp. Long read sequencing technologies, on the other hand, require high molecular weight (HMW) DNA, generally over 50 kb in length. To obtain high-purity DNA, there are several extraction methods with different approaches to isolate DNA (Dahn et al., 2022). Column-based extraction uses a silica membrane which binds the DNA, while contaminants are washed away. This approach is commonly used for short-read sequencing, but it is not perfectly suitable for long read sequencing, as the numerous centrifugation steps can reduce the DNA fragment length (Kalendar et al., 2023). Organic solvent-based extraction, such as the phenol-chloroform method, uses a liquid-liquid separation to purify DNA. This method effectively removes contaminants, however it involves more hazardous chemicals than other extraction methods (Kalendar et al., 2023; Sambrook & Russell, 2006). Salt precipitation methods use high concentrations of salts like sodium acetate or ammonium acetate to precipitate the DNA out of the solution. This method is often used in combination with other methods because of the high purity of DNA that is obtained. Finally, magnetic bead-based extractions utilise beads that selectively bind DNA, allowing the easy separation of DNA from the rest of the sample. Since magnetic bead extraction does not utilise centrifugation steps it is the method of choice in many long read sequencing projects (Howard et al., 2025).

Genome assembly

After data generation by sequencing, the subsequent processes of genome assembly and annotation can be performed by a variety of available tools. Since in non-model organisms, genomic resources are scarce, genome assembly is usually done *de novo*. Assemblers for short read data are efficient, but often cannot resolve repeat sequences longer than the read length. In contrast, long-read assemblers can improve the contiguity of the assembly by

handling longer reads that can resolve complex regions of the genomes, still the read length remains the limiting factor in resolving repetitive regions (Treangen & Salzberg, 2012). Nevertheless, *de novo* assembly processes demand significant time and computational resources and are still challenging for genomes with high heterozygosity, high number of repeat elements and polyploid genomes (Espinosa et al., 2024). These factors create highly similar or ambiguous regions that generate conflicting information for the assembler, leading to fragmented or incorrect sequences that do not accurately represent the full genome (Treangen & Salzberg, 2012). The genome assembly process is complemented with information obtained from chromosome conformation capture (3C) methods used to provide chromosome-scale structural information (McCord et al., 2020). Data obtained from 3C methods is used to help scaffold and orient contigs in genome assemblies and provides insight in the three-dimensional structure of the genome (Belton et al., 2012). By capturing physical interaction between distant DNA segments, 3C methods provide information that cannot be obtained from short or long read sequencing alone, therefore the integration of 3C data enhances generating complete reference genomes.

Genome assemblies quality is primarily measured by the assembly contiguity and completeness metrics. Contiguity indicates the number of contigs (i.e., a continuous piece of genomic sequence) and their length distribution, measuring the success of the assembler to produce a genome without gaps. A more contiguous genome has fewer, longer sequences, ideally of chromosome length (Schell et al., 2025). Completeness refers to the amount of the original genome present in the final assembly. This is typically evaluated by comparing the total assembly size to the estimated genome size and by assessing the presence of a core set of conserved genes (Simão et al., 2015).

Genome annotation

After the genome assembly, protein-coding genes and repeat sequences are annotated to provide a comprehensive understanding of the genomic features and allow functional characterisation. For non-model organisms, these steps are often hindered by the underrepresentation or lack of data in databases (Schell et al., 2025). The annotation of protein coding genes follows the identification of genic regions: introns, exons, splice sites and start and stop codons, and is supported by known protein sequences in databases and/or transcriptome RNA sequencing data (Harrow et al., 2009). Repeat elements are usually annotated using tools that rely on public databases, masking repeat sequences based on

significant similarity. These databases contain information on transposable elements and can be taxon-specific, but often lack information on satellite DNA.

When reference sequences are missing from databases, repeat identification can be performed *de novo*. *De novo* repeat identification software utilise the repetitive features to identify repeats. They are based on alignment among sequences to identify homologies (self-comparison) or they search for repeated occurrences of short motifs that can be extended to larger sequences (k-mer approaches) (Jiang, 2013). Self-comparison approaches define repeats if sequences share a certain threshold of sequence similarity over a certain length. The results are then further clustered into repeat families. The k-mer approach identifies repeated DNA sequences by counting short, fixed-length substrings (k-mers) which act as starting points (seeds) to find and extend into larger repeating patterns (Jiang, 2013). Most of the *de novo* identification tools are challenged with the fragmentation of TE and the low sensitivity for highly divergent sequences which results in RE being split as multiple sequences (Storer et al., 2022). This highlights the need for use of multiple tools for comprehensive RE annotation.

De novo repeat identification tools identify and classify repetitive DNA from whole-genome or genome skimming data. When applied to genome skimming data, meaning low coverage ($0.01 - 0.05$X) genome sequencing data, these tools primarily detect high copy regions (Novák et al., 2020). Among these high copy regions are mitochondrial DNA, chloroplast DNA, nuclear ribosomal DNA and repeat elements. Still, building repeat libraries remains a manual and time-consuming process (Mann et al., 2024), and the development of standardised pipelines is needed.

1.4. Gaps in Decapoda and freshwater crayfish genomics

Within the order Decapoda, there are to date 58 species with published genomes of which 21 at chromosome level (NCBI Genomes, August 2025.). The sequencing technologies used for the sequencing of Decapoda species are mostly PacBio in combination with Hi-C sequencing, while only four genomes were sequenced using Nanopore technology. Many of these genomes are still highly fragmented and with low contiguity, with half of the genomes having N50 lower than 1 Mb, resulting in poor annotation for protein coding genes, non-coding and repetitive sequences (Yuan et al., 2021, 2023). The genome sizes of the sequenced species range from 1.4 to 7.1 Gb (Gregory, 2025), however, the assembly size is lower than the expected genome size for all species.

Among freshwater crayfish there are only four available genomes: *Procambarus virginalis*, *Procambarus clarkii*, *Cherax destructor* and *Cherax quadricarinatus*, belonging to the family Cambaridae and Parastacidae, respectively (Austin et al., 2022; Gutekunst et al., 2018; Liao et al., 2024; M. H. Tan et al., 2020; Z. Xu et al., 2021). The sequenced species include economically relevant organisms, important in aquaculture, while the genomes of many ecologically relevant species, including the European crayfish species, have not been studied (Yuan et al., 2023). Sequencing Decapoda genomes has provided valuable insights into sex determination mechanisms, environmental adaption, stress tolerance and disease resistance. The genomes have also facilitated the design of high-throughput SNP chips, which are essential for genetic selection in aquaculture and breeding programs. Beyond aquaculture, this information can also be applied to conservation management programs.

The unique challenges of genomes

The low number of Decapoda genomes is due to the limitations in sequencing and assembly caused by large genome sizes, large number of chromosomes, high polymorphism, and high number of repeat elements (Abdelrahman et al., 2017; Yuan et al., 2023). These factors complicate the processes of generating a reference genome. For instance, organisms with larger genome size require more sequencing data to achieve the necessary coverage for complete and accurate assembly. Beyond the amount of data, larger genomes tend to contain more repetitive elements. Their amount, length, localisation and sequence identity are contributing to fragmented genomes and gaps in the assembly (Tørresen et al., 2019). Furthermore, repetitive regions can often result in higher levels of polymorphism. Polymorphisms can cause sequences from the same locus to be mistakenly identified as sequences from different loci, or it can lead to contigs breaking at polymorphic regions during assembly (Claros et al., 2012; Huang et al., 2013; Kyriakidou et al., 2018).

Genome size

Genome size shows extreme variability among and within various taxonomic levels. Genome size variation reveals positive correlation with cell size, where larger genomes are associated with larger cells for structural reasons (Hessen & Persson, 2009). As genome size increases, the volume of the nucleus increases proportionally, and cells maintain a constant nuclear-to-cytoplasmic ratio (Veitia & Bottani, 2009). It has been hypothesised that this ratio is linked to the transcriptional capacity and protein synthesis rate (Wu et al., 2022). In organisms with larger genomes, the amount of DNA, including large amounts of repetitive DNA, contributes to larger nuclear volume. The increased cellular machinery, ribosomes and synthesised

proteins, required for the larger amount of coding genes and transcribed repetitive DNA, can lead to larger cell volumes (Balachandra et al., 2022). Furthermore, the expansion of the genome, driven by DNA duplication and repetitive element insertion/proliferation leads to larger nuclear volume and consequently to larger cytoplasmic volume to maintain a constant ratio (Veitia & Bottani, 2009). With increased cell and genome size, the duration of the cell cycle increases, therefore large cells develop more slowly. This suggests that the developmental rate of the organism decreases, and organisms with large genome sizes develop more slowly than related species with smaller organisms (Dufresne & Jeffery, 2011).

In particular, Decapoda organisms exhibit a genome size ranging from 1 Gb to 40 Gb (Gregory, 2025). Such variability is characteristic in all Crustacean orders (Hessen & Persson, 2009). In different Crustacean orders, genome size is correlated to body size, latitude, temperature and water depth, however, such correlation pattern has not been found within Decapoda (Iannucci et al., 2022). Instead, genome size in Decapoda is correlated with the developmental mode, with genome size being larger in species with direct development (less larval stages). Such direct development is characteristics of freshwater crayfish (infraorder Astacidea). Organisms with many larval stages have time-limited developmental windows which require rapid cellular division. In contrast, for species with direct development the development rate is less stringent, allowing for genome expansion (Gregory, 2002; Iannucci et al., 2022). Moreover, in the infraorders Anomura, Brachyura and Astacidea, freshwater species have larger genome sizes than species in terrestrial and intertidal habitats (Iannucci et al., 2022). Generally, freshwater species tend to have larger genomes than marine species possibly because of the harsher and more fluctuating freshwater environment (Dufresne & Jeffery, 2011).

The variation in genome sizes is greatly impacted by the repetitive sequences in the genomes: transposable elements (TE) and satellite DNA (satDNA). TEs have been shown to increase genome size by replicating in response to environmental changes and satDNA under replication shows strong correlation with genome size in some *Drosophila* species (Dufresne & Jeffery, 2011; Flynn & Yamashita, 2024). Increases in genome size can also result from whole-genome duplication and polyploidisation events, which lead to an increase in the number of chromosomes within the genome (Mayrose & Lysak, 2021). In Crustacea and Decapoda, chromosomes have been studied in a small number of species. The chromosomes are numerous, very small and punctiform making them difficult to analyse (Lécher et al., 1995). In the infraorder Astacidea, the chromosome numbers range from 2n=102 to 2n=276

(Boštjančić et al., 2021). Moreover, the presence of supernumerary B chromosomes has been demonstrated in various Decapoda species (Coluccia et al., 2004; Lécher et al., 1995). These extra chromosomes are not essential for the species' survival, but they often accumulate repetitive DNA and can influence host fitness. B chromosomes are highly variable in number, even among individuals of the same species. (Houben et al., 2014). The large variability in genome size and chromosome number makes genomic research in these species both challenging and important, as it provides a framework for understanding their evolution shaping their genomes.

Satellite DNA

Satellite DNA are tandemly repeated non-coding DNA sequences typically located in heterochromatic regions of the genomes, particularly in centromeres and telomeres (Garrido-Ramos, 2017). The typical structure of a satDNA consists of a repetition of monomeric units repeated head-to-tail (Figure I-7). Based on the length of the monomer, satDNA is categorised in three groups: microsatellites (2 – 10 bp), minisatellites (10 – 100 bp) and macrosatellites (>100 bp) (Garrido-Ramos, 2017). Microsatellites form arrays of repeated monomers up to 100 bp and are the most present class of satDNA in Eukaryota genomes, while minisatellites are forming arrays of total length up to 1000 bp. Macrosatellites arrays can span up to megabases of DNA in length and are present in lower abundance in the genome than the other two classes (Richard et al., 2008). SatDNA was first discovered with density-gradient ultracentrifugation of DNA using caesium chloride, in which the repeated fraction of the DNA separates from the rest of the DNA. Later, satDNA was studied using hybridisation techniques, PCR and/or cloning methods. However, it was not until the advent of next-generation sequencing (NGS) and high-throughput genome sequencing, that satDNA could be studied more effectively. This has led to a better understanding of satDNA composition and function (Garrido-Ramos, 2017).

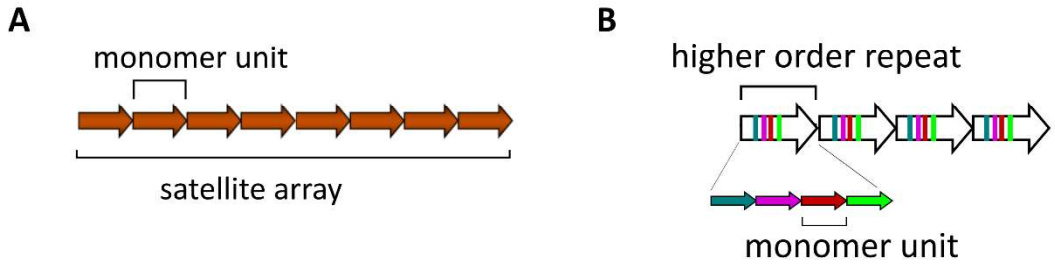


Figure I-7. Structure of satellite DNA. (A) SatDNA array organised in head-to-tail monomer units. (B) Higher order repeats (HOR).

Within a genome, satDNA is characterised by its nucleotide sequence, monomer length and copy number (Plohl et al., 2012). Furthermore, monomers of some satDNA can create higher-order repeats (HOR), in which two or more monomers create a new, longer monomer unit that becomes tandemly repeated. HOR structures are found across plant and animal species, and are common in centromeric repeats (Melters et al., 2013). SatDNA forms with a *de novo* duplication of a sequence of two or more base pairs in length (Ruiz-Ruano et al., 2016). This duplication occurs through DNA slippage during DNA replication or through reinsertion of extrachromosomal circular DNA (eccDNA) intermediaries (Garrido-Ramos, 2017). The spread of the satDNA through the genome happens through unequal crossover, rolling circle replication or transposition.

The collection of satDNA families in a genome is termed satellitome (Ruiz-Ruano et al., 2016). There are different satDNA families in one species, with typically one or a few predominant satDNA families in each species. SatDNA families can be species specific, shared among related species within a genus, shared by several genera within a family, by a whole family, several families or by a whole order (Garrido-Ramos, 2017; Martinsen et al., 2009; Petraccioli et al., 2015; Robles et al., 2004). Moreover, according to the library hypothesis, closely related species share a set of satDNA which differ by copy number in each species (Fry & Salser, 1977; Garrido-Ramos, 2017) (Figure I-8). Within a species, satDNAs undergo concerted evolution where a new variant of a satDNA sequence is homogenised within the satDNA family and is fixed between individuals of a same population within a species. The process of concerted evolution leads to greater similarity of satDNA within an individual and between individual of the same species than between further related individuals (Plohl et al., 2012).

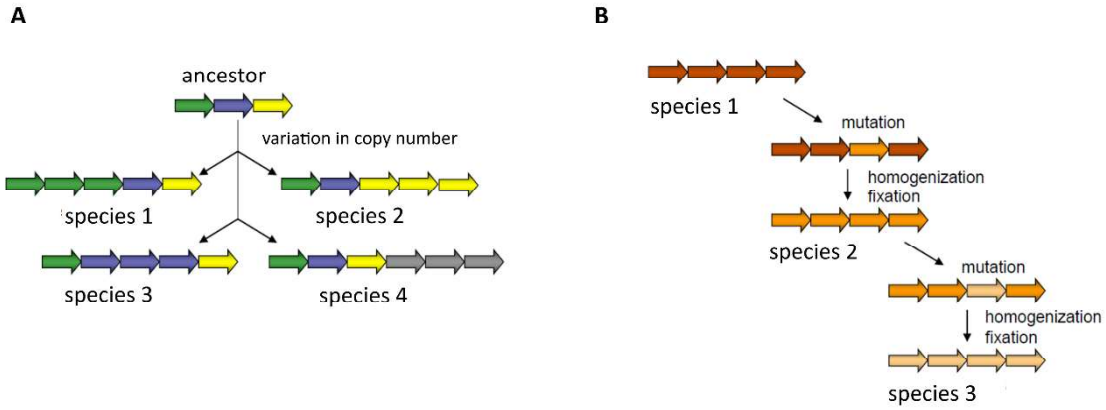


Figure I-8. Satellite DNA evolution concepts: (A) SatDNA library concept and (B) concerted evolution. Adapted from Plohl et al., 2012

satDNA is among the most dynamic component in the genomes, with changes in length and sequence happening within short evolutionary periods. Rapid amplification and sequence changes lead to reproductive isolation speciation (Šatović-Vukšić & Plohl, 2023). However, certain satDNA sequences can be conserved for long evolutionary periods because of functional constraints. Specific sequences can act as binding sites for proteins, such as the centromere protein B (CENP-B) box or are transcribed into small interfering RNAs (siRNA) involved in heterochromatin formation (Garrido-Ramos, 2017). satDNAs are involved in the assembly of centromeric chromatin, segregation of chromosomes and in preserving genome integrity (Šatović-Vukšić & Plohl, 2023). satDNA can also be transcribed into long noncoding RNAs (lncRNA) and siRNA. These transcripts are involved in heterochromatin regulation, centromere formation and gene expression regulation (Pezer et al., 2012).

In the 1970s, the first studies on satDNA in Crustacea species were conducted using traditional methods such as density gradient centrifugation, restriction enzyme digestion, DNA reassociation kinetics, hybridisation, and Sanger sequencing. These studies identified multiple satDNA conserved across various crustacean species. For instance, a GC-rich satellite discovered in the crab *Gecarcinus lateralis* (order Brachyura) was also found in lobster and shrimp species (order Decapoda), although the homology was lower, this finding suggested a phylogenetic signal (Skinner & Beattie, 1974). Later studies on Crustacea and Decapoda focused on the use of microsatellites with application in population genetic studies, and aquaculture (Feng & Li, 2008; Gross et al., 2021; Heras et al., 2016; Liu et al., 2023; Lovrenčić et al., 2022).

Whole genome sequencing studies on Crustacea and Decapoda rarely report the number of identified satDNA sequences. Often, assembled Decapoda genomes are highly fragmented, therefore difficult to annotate (Austin et al., 2022). Consequently, satDNA is frequently missing or is underrepresented in these genome assemblies. In the crayfish *P. clarkii*, tandem repeats constitute 5.21% of the genome. In contrast, satDNA is making up 27.5% of the genome in *P. leptodactylus* (Boštjančić et al., 2021). The difference in satDNA amount could be an underestimation due to the assembly not representing correctly the complement of tandem repeats or a true biological signal. This makes crayfish an interesting group to study on satDNA.

Transposable elements

Transposable elements (TE) are repetitive DNA sequences spread throughout the genome, with the ability to move the location within a genome (Bourque et al., 2018). Based on the transposition mechanism, TEs are divided into Class I Retrotransposons and Class II DNA transposons. Within each class, the different subclasses are divided based on the mechanism of chromosomal integration (Bourque et al., 2018) (Figure I-9). Class I retrotransposons utilise a copy-and-paste transposition mechanism in which an RNA intermediate is transcribed into a cDNA copy and integrated in the genome (Boeke et al., 1985). For long terminal repeats (LTR) integrases catalyse the cleavage and strand-transfer reaction. Non-LTR retrotransposons are mobilised through target-primed reverse transcription. The *Dictyostelium* intermediate repeat sequence (DIRS) elements are, unlike other Class I TE, using single stranded cDNA intermediates for transposition (Malicki et al., 2020). Class II DNA transposons are mobilised through a DNA intermediate in a cut-and-paste mechanism and include Helitron, Crypton and Maverick/Politron subclasses. In the case of Helitron elements, the transposition mechanism includes a circular DNA intermediate (Bourque et al., 2018; Munoz-Lopez & Garcia-Perez, 2010).

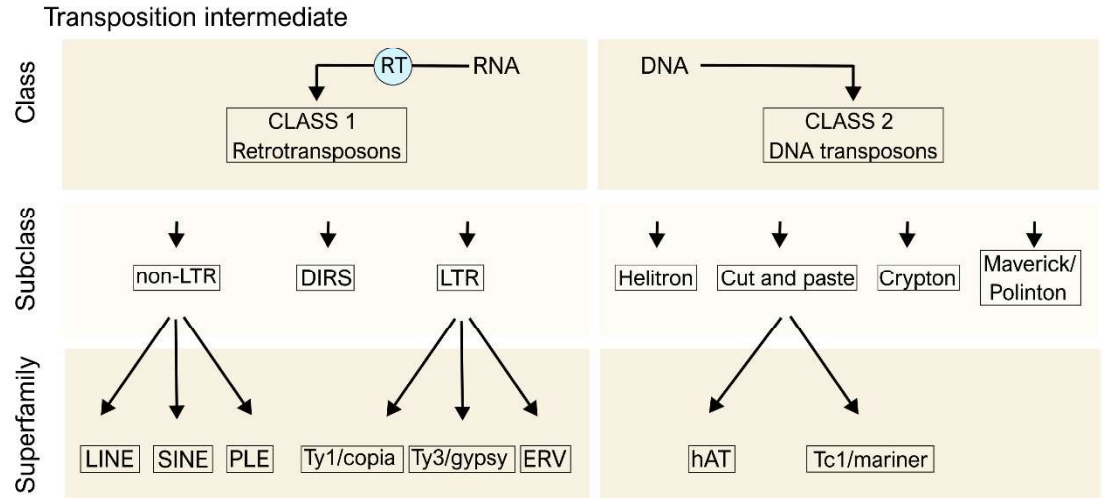


Figure I-9. Class, subclass, and superfamily classification of transposable elements. Adapted from Bourque et al., 2018

Within a genome, TEs are not distributed randomly, but exhibit preference for insertion sites. Insertions which cause deleterious effects on the genome, i.e. disrupt gene exons, are rapidly removed from the population, and rarely fixed (Bourque et al., 2018). The general transposition mechanism of TE leads to genome expansion and is counteracted by DNA deletion. These two processes are a key driver of genome size evolution. Moreover, the mobilisation of TE leads to structural variation in the genome, such as deletions, duplications, and inversions. The high homology between TE copies facilitates recombination effects in distant genomic regions (Aasegg Araya et al., 2025). The insertion of TEs can influence the regulation and expression of nearby genes. Furthermore, the transcription of TEs can regulate gene expression, chromatin accessibility or RNA interference (Lanciano & Cristofari, 2020). The presence of TEs within genomes is primarily attributed to vertical transfer, from parents to offspring by reproduction. However, TEs can be transmitted through the non-reproductive mechanism of horizontal transfer (HT), with viruses being major vectors of HT. Horizontally transferred TE into new genomes can lead to increased genomic variation, chromosome rearrangements and potential functional innovation (Gilbert & Feschotte, 2018).

TEs constitute a major but variable part of the genome in different species, such as 2.7% in the fish *Takifugu rubripes* and 85% in maize *Zea mays* (J. Wang et al., 2021). In Crustacea genomes, the amount of TEs varies from 6.74% in the crab *Eriocheir sinensis* (Y. Xu et al., 2023) to 78.22% in the krill *Euphausia superba* (Shao et al., 2023). Moreover, some TEs exhibited lineage specificity among Decapoda, with non-LTR/CR1 being abundant in Brachyura species, while non-LTR/CRE were found abundant in Astacidea species (Y. Xu et al., 2023). Within freshwater crayfish, TEs constitute from 27% of the genome in *P.*

virginalis, to 65% in *P. clarkii* (Liao et al., 2024; M. H. Tan et al., 2020). The most abundant TE in *Cherax* and *Procambarus* species are LTR/LINE elements which make approximately half of the whole TE content (Austin et al., 2022; M. H. Tan et al., 2020). In addition to the composition of TEs within a genome, temporal dynamics of TE activity have a significant evolutionary impact (Zeng et al., 2025). Certain TE classes, such as the LTR/Ty3 and Bel-Pao families, show low divergence across multiple Crustacea species, which suggest active proliferation that is driving genome expansion, while other classes, such as LTR/Copia show high divergence only in a few species, preserving genomic stability over time (Petersen et al., 2019; Zeng et al., 2025). Within Decapoda, many of the TEs cannot be classified, yet they are often key in genome evolution, as their activity coincides with shifts in genomic architecture or selective pressures (Colonna Romano & Fanti, 2022; Zeng et al., 2025). Furthermore, specific patterns of TE expression were revealed as critical in the process of moulting in the crab *E. sinensis* and shrimp *P. vannamei* (Zeng et al., 2025). Certain TEs were shown to be associated with responses to various environmental stressors or prolonged air exposure (Salem, 2024; Y. Xu et al., 2023). These studies highlight the diverse roles of TE in Crustacea, as drivers of genomic change and crucial regulators of physiological and adaptive responses.

Thesis objectives

In this work, I aim to characterise the genomic variability in Decapoda and particularly in freshwater crayfish, to address the gaps in understanding the evolution and adaptation of these species. I characterised the genomic variability of Decapoda on different levels beginning with a comparative study of the repeatome in 20 Decapoda species that demonstrated the role of transposable elements in genome size variation observed in these species. Furthermore, I characterised the diversity of satellite DNA in 19 freshwater crayfish species, highlighting the importance of the satellitome in their genomic diversity. In an effort to facilitate genome sequencing in European freshwater crayfish species, I tested different DNA extraction protocols on different crayfish tissues and evaluated various library preparation strategies for long-read genome sequencing. Finally, using population genomic approaches, I studied the genetic diversity of the endemic idle crayfish *Austropotamobius bihariensis*, indicating the need for conservation actions.

Overview of the chapters

The scientific contributions are presented in four chapters (II -V)

Chapter II: Abundance and diversification of repetitive elements in Decapoda genomes; published in Genes, doi: 10.3390/genes14081627

Chapter III: The extraordinary satellitome diversity of freshwater crayfish: a driver of genome evolution; under review in BMC Mobile DNA, doi: 10.21203/rs.3.rs-7499918/v1

Chapter IV: From DNA extraction to long read sequencing: assessment of workflow challenges on giant genomes of two non-model Astacidae (Crustacea: Decapoda) species; intended for submission in BMC Genomics

Chapter V: Genomic insights into the conservation status of the Idle Crayfish *Austropotamobius bihariensis* Pârvulescu, 2019: low genetic diversity in the endemic crayfish species of the Apuseni Mountains; published in BMC Ecology and Evolution, doi: 10.1186/s12862-024-02268-5

Chapter II and Chapter III explore the diversity of the repeatome in freshwater crayfish and Decapoda species. Chapter II aims to characterise the repetitive elements (REs) in Decapoda genomes. The repetitive fraction of the genome is often not fully characterised in genome

assemblies, therefore a new standardised bioinformatic pipeline was developed and applied to Decapoda genome assemblies. The study revealed a generally higher amount of REs in Decapoda than in other Crustacea species. Furthermore, a strong correlation was observed between assembly size and the load of TEs. Chapter III addresses the satellitome of the species based on low coverage genome sequencing and *de novo* repeat identification. The comprehensive analysis of satDNA diversity across freshwater crayfish species and families, phylogenetic reconstruction and investigation of chromosomal distribution aims to elucidate the role of satDNA in genome evolution. The analyses revealed a high proportion of repetitive DNA and an extraordinary diversity of satDNA families. I further characterised a conserved satDNA family identified in all species with chromosomal localisation using fluorescence *in situ* hybridisation (FISH), revealing its role as pericentromeric DNA. Both Chapter II and Chapter III showed that repetitive elements largely reflect the phylogenetic relationships within Astacidea and Decapoda, respectively. These results collectively underscore the significant impact of REs on Decapoda genome evolution providing a baseline for genomic studies.

In Chapter IV, I aimed to identify the most suitable workflow for long read genome sequencing, as basis for genomic studies in European freshwater crayfish species. I tested six DNA extraction protocols on three different tissue types to obtain high-quality high molecular weight (HMW) DNA. Based on DNA yield, purity, and fragment length, I identified extraction protocols suitable for long read sequencing. I tested two long read sequencing approaches and evaluated different library preparation methods. I identified the salting-out DNA extraction protocol in combination with amplification based PacBio library preparation as the best workflow for obtaining high amounts of sequencing reads suitable for genome assembly of two freshwater crayfish species.

In Chapter V, I focused on the freshwater crayfish *A. bihariensis*, whose geographical range is restricted to the Apuseni Mountains in Romania. Considering the outlined difficulties encountered in long-read genome sequencing in Chapter IV, a reference genome could not be generated within the timeframe of this thesis. Therefore, I conducted a population genomic survey based on a SNP dataset obtained via ddRADseq to obtain the species' genetic status by estimating population genetic diversity metrics. The study revealed critically low numbers of private alleles and small effective population size. Due to the reduced genetic diversity, the populations are at risk which makes them vulnerable to environmental changes, underscoring the urgent need for appropriate conservation measures.

Chapter II

Abundance and diversification of repetitive elements in Decapoda genomes

Christelle Rutz, Lena Bonassin, Arnaud Kress, Caterina Francesconi, Ljudevit Luka Boštjančić, Dorine Merlat, Kathrin Theissinger, Odile Lecompte

Published in Genes (2023), 14, 1627, doi: 10.3390/genes14081627

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

2.1. Introduction

With over 15,000 living species, Decapoda represents a diverse order of Crustacea that includes lobsters, crayfish, crabs, prawns, and shrimps (De Grave et al., 2009). They are a crucial component of marine and freshwater ecosystems (Reynolds et al., 2013, Souty-Grosset et al., 2006). 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) (Wolfe et al., 2019).

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 (Boštjančić et al., 2021) , 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 González-Tizón et al., 2013; Lécher et al., 1995), while this number can explode in Pleocyemata species to a 2n of 376 for the Astacidea *Pacifastacus leniusculus* (Crandall & De Grave, 2017; Niiyama 1962).

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 (Gregory, 2005). A high proportion of REs can greatly complicate genome sequencing and can lead to fragmented and incomplete assemblies (Pop, 2009; Tørresen et al., 2019; Treangen et al., 2012). 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 Shapiro & von Sternberg, 2005). 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 (Jurka et al., 2007).

SatDNAs consist of tandemly repeated patterns of nucleotides, called repeat units (monomers) (Garrido-Ramos, 2017). Different satDNA families are present in the genome, with usually only one or a few predominant families (Macas et al., 2007; Miga, 2015; Mravinac et al., 2005; Ruiz-Ruano et al., 2016). SatDNAs can have specific roles in gene and

genome regulation, such as chromosome organization, pairing, and segregation formation of the centromere locus (Plohl et al., 2008; 2012), in epigenetic regulation of heterochromatin establishment, and modulation of gene expression in response to stress (Biscotti et al., 2015; Pezer et al., 2012). In Crustacea, some SatDNA transcripts can have an impact on the inter-molt stage (Wang et al., 1999). 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 (Bennetzen et al., 2014; Bourque et al., 2018; Craig, 2002; Deininger et al., 2003). 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 (Barrón et al., 2014; Burns & Boeke, 2012; Deininger et al., 2003; Kim et al., 2012). Organisms living in challenging environmental conditions can have more TEs in their genome, increasing genome plasticity to respond to stress factors (Lanciano & Mirouze, 2018). 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) (Di Stefano, 2022; Kojima, 2019; Slotkin & Martienssen, 2007; Wicker et al., 2007). 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 (Flynn et al., 2020) and RepeatMasker (Smit et al., 2013). However, a wide variety of additional tools have been developed, such as RECON (Bao & Eddy, 2002), RepeatScout (Price et al., 2005) and LtrHarvest/Ltr_retriever (Ou & Jiang,

2018), REPET (Flutre et al., 2011), RepeatExplorer (Novák et al., 2013) (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.

2.2. Materials and Methods

2.2.1. 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 (Supplementary table II-1). 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 II-1) (Holt et al., 2018). 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 II-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 II-1. Genomic dataset 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	Meng et al., 2021
	Penaeus indicus	GCA018983055.1	1936	88.5	SRR12969543	2810	Swathi et al., 2018
	Penaeus japonicus	GCF017312705.1	1705	96.6	DRR278744	2170	Swathi et al., 2018
	Penaeus monodon	GCF015228065.1	2394	83.9	SRR11278066	2200	Swathi et al., 2018
	Penaeus vannamei	GCF003789085.1	1664	84.8	SRR13661692	2270	Swathi et al., 2018
Caridea	*Caridina multidentata*	GCA002091895.1	1949	25.2	DRR054559	3230	Kawato et al., 2021
	Macrobrachium nipponense	GCA015104395.1	1985	41	SRR9026393	4600	Jin et al., 2021
Achelata	*Panulirus ornatus*	GCA018397875.1	1926	70	SSR13822589	3230	Veldsman et al., 2021
Astacidea	*Procambarus virginalis*	GCA020271785.1	3701	67	SRR12901906	3500	Gutekunst et al., 2018
	Procambarus clarkii	GCF020424385.1	2735	94.3	SRR14457195	8500	Shi et al., 2018
	Cherax destructor	GCA009830355.1	3337	81.7	SRR10467055	4500	Austin et al., 2022
	Cherax quadricarinatus	GCA009761615.1	3237	69.9	SRR10484712	5000	Tan et al., 2020
	Homarus americanus	GCF018991925.1	2292	93	SRR12699166	7700	Polinski et al., 2021
Anomura	*Paralithodes camtschaticus*	GCA018397895.1	3810	44.2	SRR13805857	7290	Veldsman et al., 2021
	Paralithodes platypus	GCA013283005.1	4805	71.7	SRR1145749	5490	Tang et al., 2021

	Birgus latro	GCA018397915.1	2959	57.7	SRR13816158	6220	Veldsman et al., 2021
Brachyura	*Chionoecetes opilio*	GCA016584305.1	2003	91	SRR11278230	1655	
	Eriocheir sinensis	GCA013436485.1	1272	92.6	SRR11971329	2230	Liu et al., 2016
	Portunus trituberculatus	GCF017591435.1	1005	93.5	SRR9964028	2250	Liu et al., 2016
	Callinectes sapidus	GCA020233015.1	998	90.4	SRR15834103	2290	Jimenez et al., 2010
Other Crustacea	*Amphibalanus amphitrite* (Cirripedia)	GCA019059575.1	808	93.9	SRR9595623	481	Kim et al., 2019
	Armadillidium vulgare (Isopoda)	GCA004104545.1	1725	84.5	SRR8156178	1660	Chebbi et al., 2019
	Daphnia magna (Phyllopoda)	GCA020631705.2	161	98.6	SRR15012074	238	Routtu et al., 2014
	Darwinula stevensoni (Podocopida)	GCA905338385.1	382	90.3	SRR8695251	437	Tran Van et al., 2021
	Eurytemora affinis (Copepoda)	GCA000591075.2	389	91	SRR2452640	616	Rasch et al., 2004
	Hyalella azteca (Amphipoda)	GCA000764305.4	551	93.8	SRR1556043	1050	Poynton et al., 2018

2.2.2. Identification and Annotation of Repetitive Elements

2.2.2.1. Identification of Satellite DNA Families

For each species, a set of Illumina paired-end reads was randomly chosen in the SRA database (Table II-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 II-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 (Chikhi & Medvedev, 2014). The sets of reads were then analysed using the TAREAN pipeline, Galaxy version 2.3.8.1 (Novák et al., 2017) (reads trimmed at 100 bp and default parameters) to compile each species-specific library of satellite elements.

2.2.2.2. 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 (Flynn et al., 2020) 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 (Bao et al., 2015) 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 II-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.

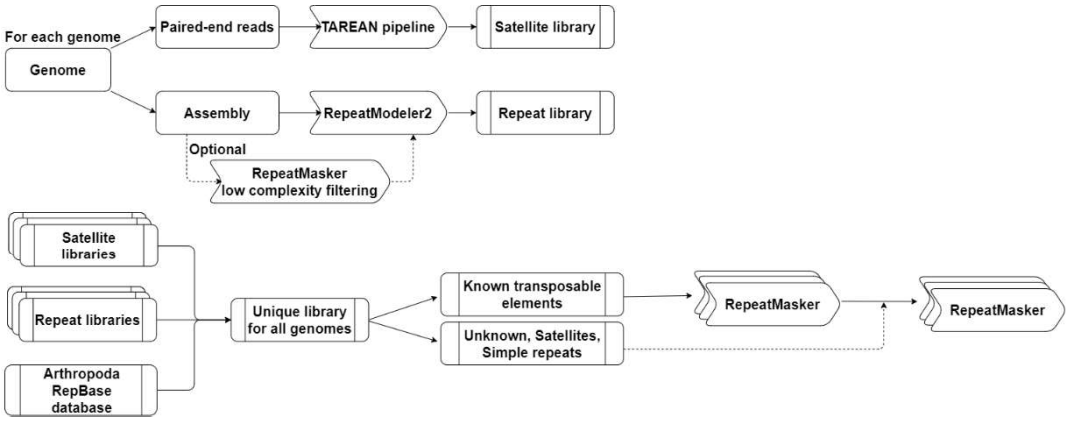


Figure II-1. Standardized annotation protocol for repetitive elements developed in this study.

2.2.2.3. 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 (Smit et al., 2013) following a two-step approach (Figure II-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.

2.2.2.4. 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 (Demšar et al., 2013). The sequence divergence distribution was calculated as Kimura distances (rates of transitions and transversions) using the RepeatMasker tools "calcDivergenceFromAlign.pl" and "createRepeatLandscape.pl".

2.3. Results and Discussion

2.3.1. 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 II-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 (Trangean & Salzberg, 2012). 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 II-2).

Table II-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
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 II-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) 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) improved the classification of the RepBase database (Bao & Eddy, 2002), 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 (Supplementary figure II-1). Well-categorized REs have also been submitted to RepBase.

2.3.2. 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 II-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 II-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* (Boštjančić et al., 2021). 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 (Miga, 2015; Silva et al., 2023). However, a large number of satDNA repeats has already been found in Arthropoda, such as *Triatoma infestans* (42 families, genome size 1.4 Gb) (Pita et al., 2017), *Locusta migratoria* (62 families, genome size 6 Gb) (Ruiz-Ruano et al., 2016), the morabine grasshoppers (129 families, genome size 5 Gb) (Palacios-Gimenez et al., 2020), and the fish *Megaleporinus microcephalus* (164 families, assembly size 1.2 Gb) (Utsunomia et al., 2019). It should be noted that our results may still underestimate the real number of satDNA families, due to the fragmentation of available assemblies (Supplementary table II-1). 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 II-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 (Supplementary figure II-1). 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 (sproul et al., 2022). This is because assemblies based on long reads are often more contiguous (Logsdon et al., 2020; Paajanen et al., 2019). 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 (Supplementary table II-1).

2.3.3. 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 II-1, Figure II-2, and Supplementary table II-1). The non-Decapoda *H. azteca* also presents fewer REs, with 26.12%, and is one of the genomes assembled with short reads only (Figure II-2 and Supplementary table II-1), 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

II-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 II-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.

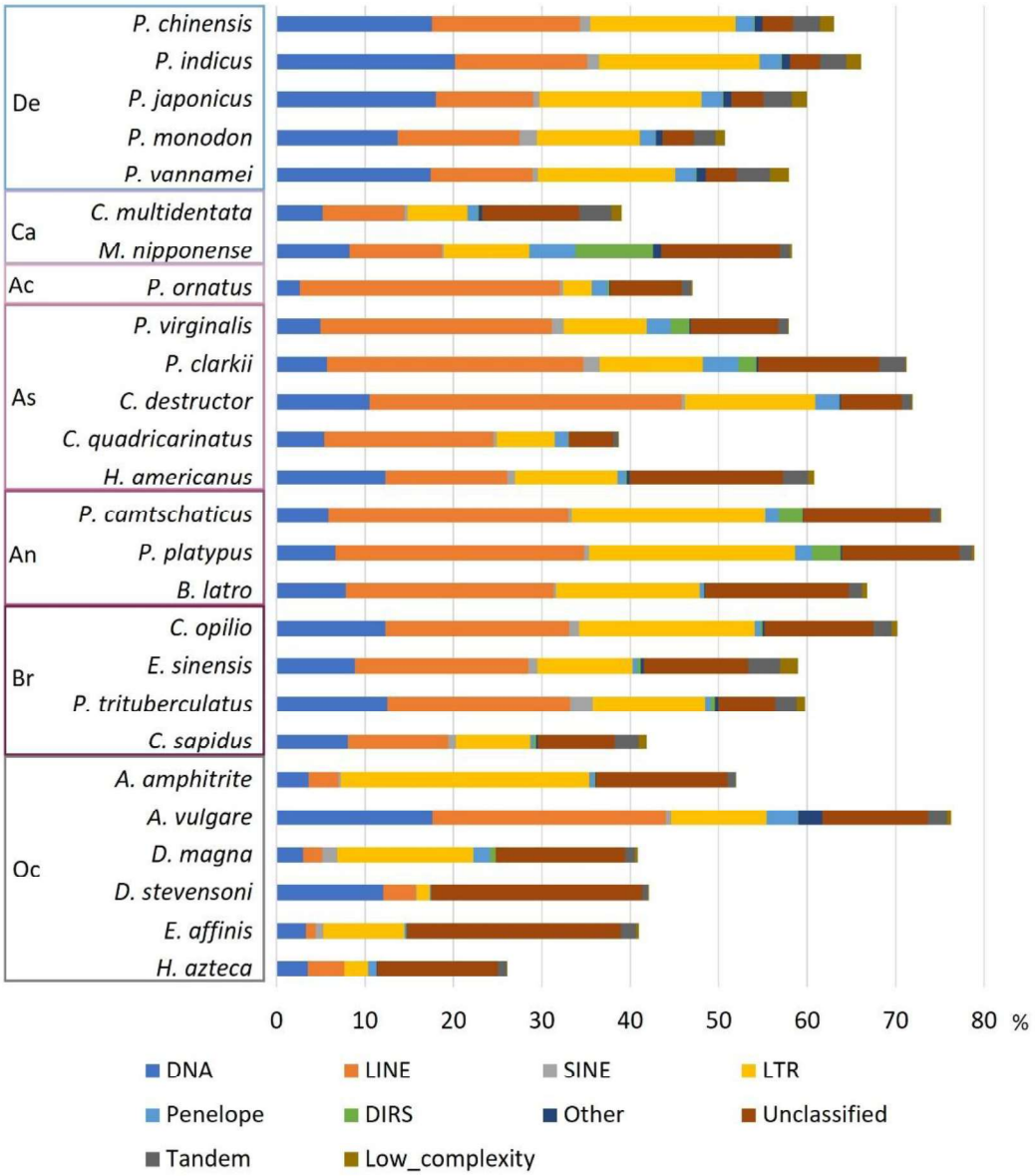


Figure II-2. Proportion and content of repetitive elements in genomes. Percentage of repetitive ele-ments in the genome by class of repetitive elements. De, Dendrobranchiata; Ca, Caridea; Ac, Achelata; As, Astacidea; An, Anomura; Br, Brachyura; Oc, other Crustacea.

According to RE studies of Decapoda species included in assembly publications, the proportion of REs varies from 8% to 82% (Austin et al., 2022; Bacvaroff 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., 2020; Uengwetwanit et al., 2021; Veldsman et al., 2021; Wang et al. 2022; Xu et al., 2021; Yuan et al., 2021; Zhang et al., 2019). 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 (Tan et al., 2020). 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). 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 (Gutekunst et al., 2018; Polinski et al., 2021). 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 II-2). For the *P. platypus* genome, we observed similar overall results to Tang et al. (2021) (Figure II-2). 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 (Tan et al., 2020; Tang et al., 2020), while here we determined that repetitive elements represent 58.93% of the genome (Figure II-2). Taken together, these results show that our method provides greater or equal proportion of REs but with a better characterization.

The Decapoda species studied here all presented high proportions of REs, ranging from 58% to 79% (Figure II-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% (Petersen et al., 2019; Sproul et al., 2022; Wu & Lu, 2019). 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 (Shao et al., 2023). In terms of TE landscape, Decapoda presents only a few SINE elements, as for all Arthropoda (Figure II-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 (Petersen et al., 2019; Sproul et al., 2022; Wu & Lu, 2019). Here, we showed that DNA transposons were the major subclass in all Dendrobranchiata species, followed by LINEs (Figure II-2). This is similar to what is observed in most insect species, where DNA transposons are generally the major TE group present in genomes (Petersen et al., 2019; Sproul et al., 2022; Wu & Lu, 2019). Interestingly,

our results revealed a different situation in the studied Pleocyemata species, where LINE and LTR elements are more abundant (Figure II-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 (Petersen et al., 2019; Sproul et al., 2022). 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. 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 (Lyon, 2006; Slotkin & Martienssen, 2007). 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.

2.3.4. 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 = 1.864×10^{-6}) and the percentage of TEs ($\rho = 0.6$, p-value = 1.48×10^{-3}) (Figure II-3A,B). The estimated genome size (Table II-2) was positively correlated with the load of TEs ($\rho = 0.62$, p-value = 7.114×10^{-4}), but there was no significant correlation with the percentage of TEs ($\rho = 0.47$, p-value = 1.421×10^{-2}) (Figure II-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 (Supplementary figure II-1). The correlations between the percentage of REs and both assembly and estimated genome size were not significant, with $\alpha = 0.005$ (Supplementary figure II-1).

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) 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) 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 (Petersen et al., 2019; Sproul et al., 2022). 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.

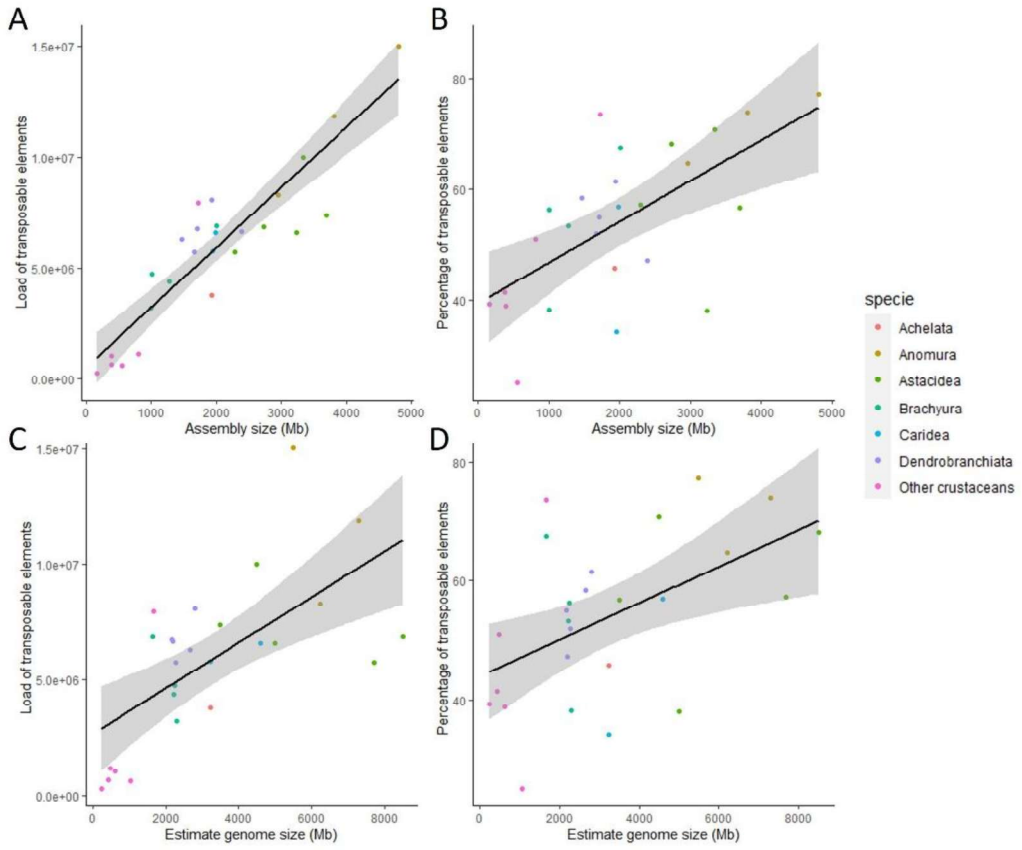


Figure II-3. Correlation between genome size and TEs. Correlation plots between assembly or estimated genome size and load (number of copies) or percentage of TEs. Orders and suborders are indicated by different colours. (**A**). Correlation between assembly size and the load of TEs. Spearman rank correlation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

2.3.5. Frequency of satDNA Families Occurrence

In Crustacea, and particularly in Decapoda, we annotated a large number of different satDNA families (Table II-2) and evaluated the occurrence of each family in each genome (Figure II-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 II-2), only a few are predominant in each genome (Figure II-4), as seen in several other studies (Mravinac et al., 2005; Miga, 2015; Ruiz-Ruano et al., 2016). 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 II-1) and presented the largest numbers of families that were highly repeated in their genomes (Figure II-4). They also tend to have the highest total number of families (Table II-2). This suggests that satDNA is a key factor in explaining the huge variations in genome size observed in Decapoda.

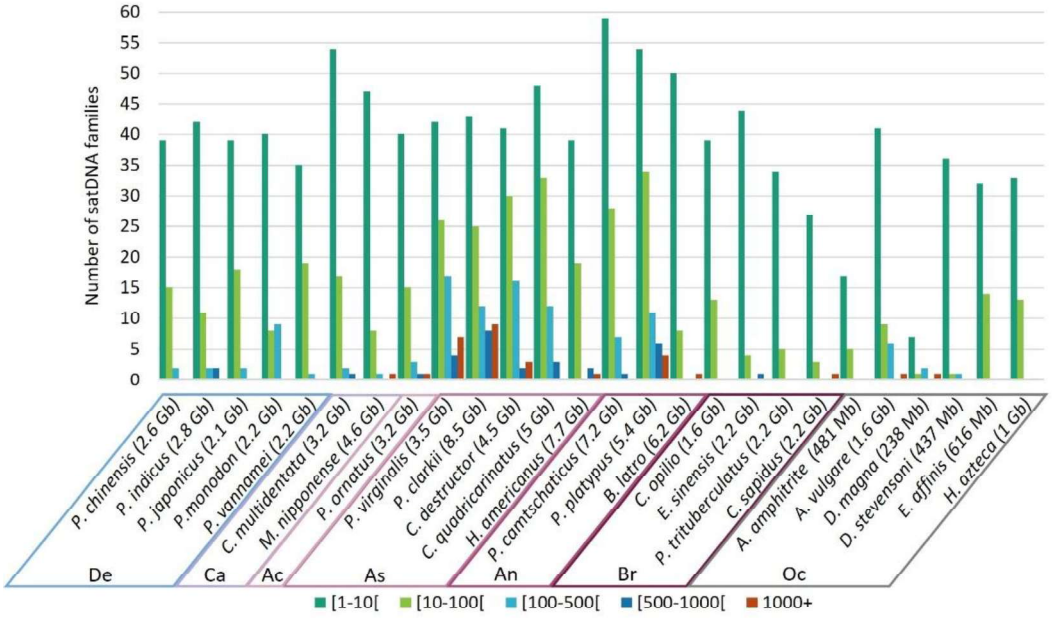


Figure II-4. Distribution of satDNA families according to the number of occurrences in each genome. Low-frequency families (less than 10 occurrences) are indicated in dark green, while highly abundant families with more than 1000 occurrences are indicated in red. Number indicated for each species is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

2.3.6. 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 II-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 (petersen et al., 2019), and appear highly conserved across all the genomes (Figure II-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 II-5). This dendrogram mainly followed the currently known species phylogeny (Wolfe et al., 2019) except for *A. vulgare*,

whose RE proportions and composition were more similar to Decapoda (Figure II-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 II-2). This may explain why *A. vulgare* is clustered with Decapoda species and not with other crustaceans (Figure II-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 (Wolfe et al., 2019). Many studies, including Petersen et al. (2019), Sproul et al. (2022), and Wu and Lu (2019), 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 II-5) (De Grave et al., 2009). 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 (Dodsworth et al., 2015; 2017). This can be explained by the capacity of some REs to have a high conservation and synteny within species (Silva et al., 2003; Vitales et al., 2020; Zhu et al., 2003). This approach could therefore be used in the future to determine the phylogeny of non-model species using low-coverage, low-cost sequencing.

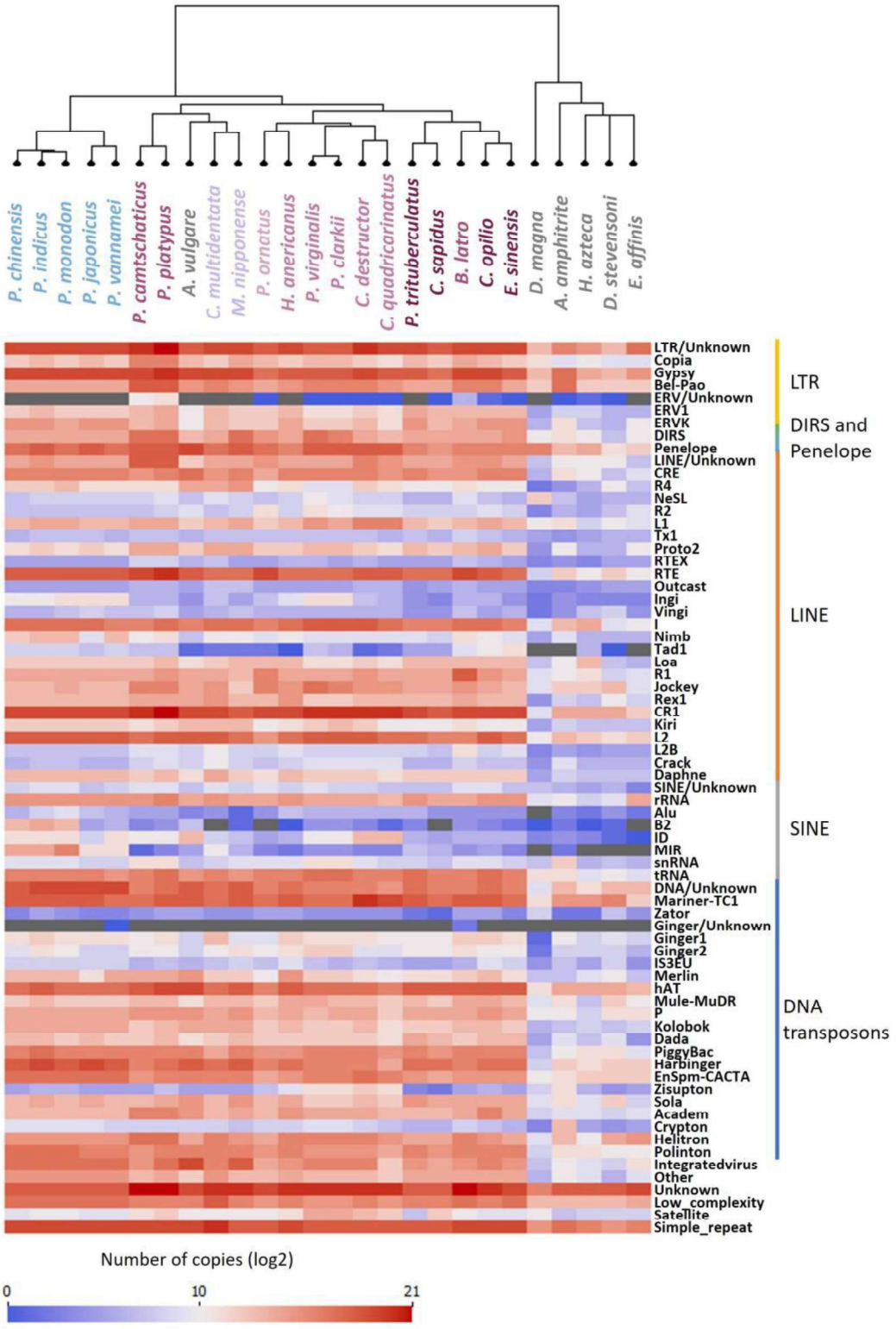


Figure II-5. Diversity of repetitive elements. Log2 of the load of each family of repetitive elements identified for each genome was graduated between 0 (blue) and 21 (red). Gray colour indicates raw values of 0, before log2 transformation. The dendrogram was produced according to repeat profile by clustering.

2.3.7. 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 II-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.

In Dendrobranchiata, sequence divergence landscapes were similar for the five species (Figure II-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 (Figure II-2 and Figure II-5). Interestingly, the content of repeats showed that DNA transposons are the most widespread among the suborder (Figure II-2). However, the expansion of DNA transposons was older and more spread out over time (Figure II-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 (Figure II-5 and Figure II-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 II-6).

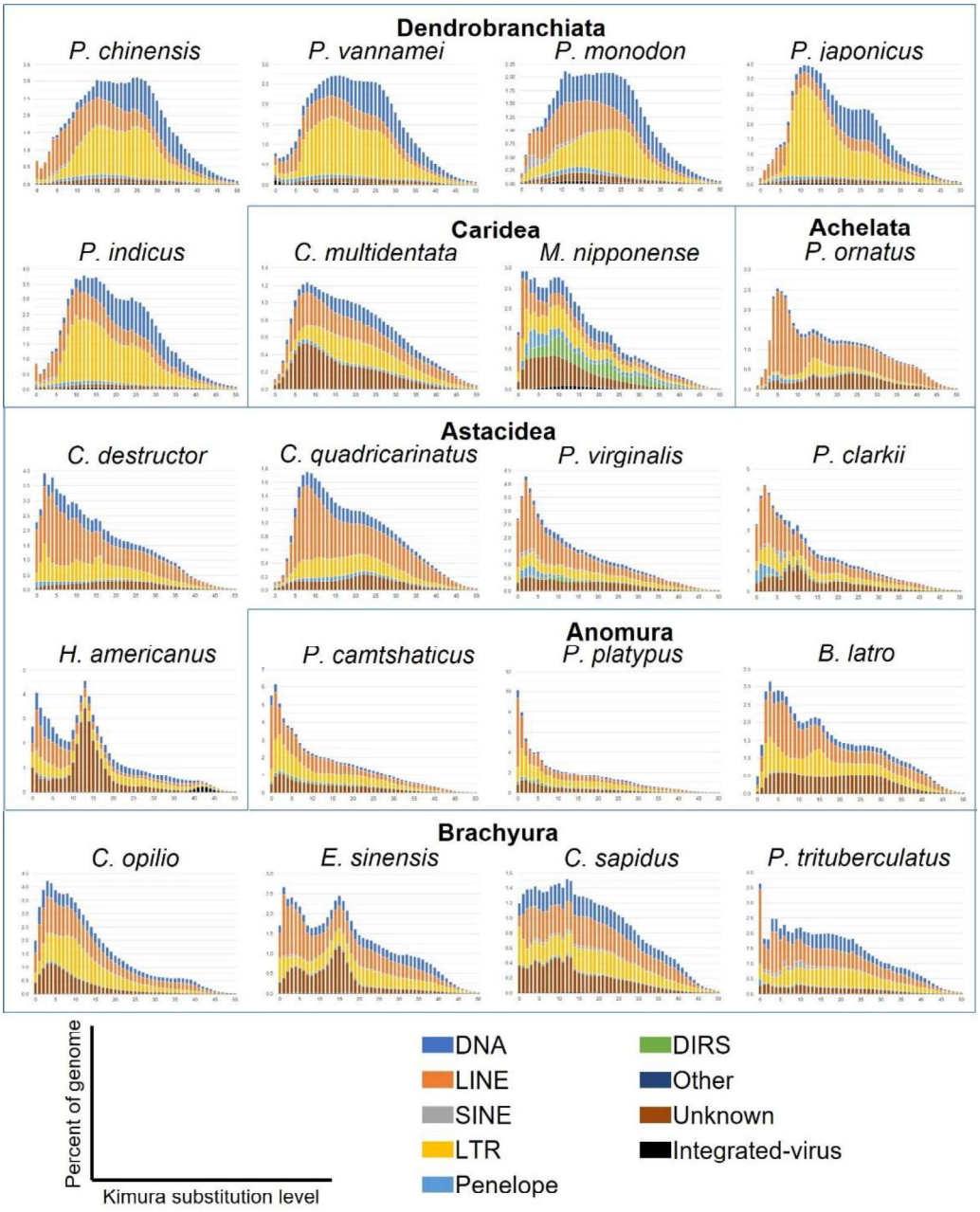


Figure II-6. Sequence divergence distribution of TEs representing TE accumulation history based on Kimura 2P distance. Percentage of sequence divergence, or Kimura substitution level, is indicated on the x-axis. On the y-axis is the percentage of the genome occupied by each TE type; the scale is different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

The two Caridea species presented a different sequence divergence landscape (Figure II-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 II-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 II-6).

Within Astacidea, *H. americanus* has a different TE landscape compared to the other four species belonging to the infraorder (Figure II-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 II-2). Integrated virus in *H. americanus* sequences corresponds to the white spot syndrome virus (WSSV) (Bao et al., 2020), suggesting that *H. americanus* faced this virus a long time ago and these sequences were then propagated (Figure II-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 (Cawthorn, 2011; Clark et al., 2013). 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 II-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 II-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 II-2). In Anomura, the intragroup with the highest percentage of LTRs within Decapoda (Figure II-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 II-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 (Supplementary figure II-2).

A clear differentiation in sequence divergence distribution between Dendrobranchiata and Pleocyemata species was observed, as seen with the proportion and diversity of repeats (Figure II-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 (Lanciano & Mirouze et al., 2018; Shapiro & von Sternberg et al., 2005). 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 (Sotero-Caio et al., 2017). 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.

2.4. 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.

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.

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.

References

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. *Frontiers in Genetics*, *12*(January), 1–8. https://doi.org/10.3389/fgene.2021.695763

Bachvaroff, T. R., McDonald, R. C., Plough, L. V., & Sook Chung, J. (2021). Chromosome-level genome assembly of the blue crab, Callinectes sapidus. *G3 Genes|Genomes|Genetics*, *11*(9). https://doi.org/10.1093/g3journal/jkab212

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

Bao, W., Tang, K. F. J., & Alcivar-Warren, A. (2020). The Complete Genome of an Endogenous Nimavirus (Nimav-1_LVa) From the Pacific Whiteleg Shrimp Penaeus (Litopenaeus) Vannamei. *Genes (Basel)*, *11*(1). https://doi.org/10.3390/genes11010094

Bao, Z., & Eddy, S. R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research*, *12*(8). https://doi.org/10.1101/gr.88502

Barrón, M. G., Fiston-Lavier, A.-S., Petrov, D. A., & González, J. (2014). Population genomics of transposable elements in Drosophila. *Annual Review of Genetics*, *48*(561). 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. *Annual Review of Plant Biology*, *65*(1), 505–530. https://doi.org/10.1146/annurev-arplant-050213-035811

Biscotti, M. A., Canapa, A., Forconi, M., Olmo, E., & Barucca, M. (2015). Transcription of tandemly repetitive DNA: functional roles. *Chromosome Research*, *23*(3), 463–477. https://doi.org/10.1007/s10577-015-9494-4

Boštjančić, L. L., Bonassin, L., Anušić, L., Lovrenčić, L., Besendorfer, V., Maguire, I., Grandjean, F., Austin, C. M., Greve, C., Hamadou, A. Ben, & Mlinarec, J. (2021). The Pontastacus leptodactylus (Astacidae) Repeatome Provides Insight Into Genome Evolution and Reveals Remarkable Diversity of Satellite DNA. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.611745

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 Biology*, *19*(1), 199. https://doi.org/10.1186/s13059-018-1577-z

Burns, K. H., & Boeke, J. D. (2012). Human Transposon Tectonics. *Cell*, *149*(4), 740–752. https://doi.org/10.1016/j.cell.2012.04.019

Cawthorn, R. J. (2011). Diseases of American lobsters (Homarus americanus): A review. *Journal of Invertebrate Pathology*, *106*(1), 71–78. https://doi.org/10.1016/j.jip.2010.09.010

Chebbi, M. A., Becking, T., Moumen, B., Giraud, I., Gilbert, C., Peccoud, J., & Cordaux, R. (2019). The Genome of Armadillidium vulgare (Crustacea, Isopoda) Provides Insights into Sex Chromosome Evolution in the Context of Cytoplasmic Sex Determination. *Molecular Biology and Evolution*, *36*(4), 727–741. https://doi.org/10.1093/molbev/msz010

Chikhi, R., & Medvedev, P. (2014). Informed and automated k -mer size selection for genome assembly. *Bioinformatics*, *30*(1), 31–37. https://doi.org/10.1093/bioinformatics/btt310

Clark, K. F., Greenwood, S. J., Acorn, A. R., & Byrne, P. J. (2013). Molecular immune response of the American lobster (Homarus americanus) to the White Spot Syndrome Virus. *Journal of Invertebrate Pathology*, *114*(3), 298–308. https://doi.org/10.1016/j.jip.2013.09.003

Craig, N. L., Cragie, R., Gellert, M., & Lambowitz, A. M. (2002). *Mobile DNA II*. ASM Press.

Crandall, K. A., & De Grave, S. (2017). An updated classification of the freshwater crayfishes (Decapoda: Astacidea) of the world, with a complete species list. *Journal of Crustacean Biology*, *37*(5), 615–653. https://doi.org/10.1093/jcbiol/rux070

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. *RAFFLES BULLETIN OF ZOOLOGY*, *21*.

Deininger, P. L., Moran, J. V, Batzer, M. A., & Kazazian, H. H. (2003). Mobile elements and mammalian genome evolution. *Current Opinion in Genetics & Development*, *13*(6), 651–658. https://doi.org/10.1016/j.gde.2003.10.013

Demšar, J., Curk, T., Erjavec, A., Črt, G., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: Data Mining Toolbox in Python. *JMLR*, *14*(71).

Di Stefano, L. (2022). All Quiet on the TE Front? The Role of Chromatin in Transposable Element Silencing. *Cells*, *11*(16), 2501. https://doi.org/10.3390/cells11162501

Dodsworth, S., Chase, M. W., Kelly, L. J., Leitch, I. J., Macas, J., Novak, P., Piednoel, M., Weiss-Schneeweiss, H., & Leitch, A. R. (2015). Genomic Repeat Abundances Contain Phylogenetic Signal. *Systematic Biology*, *64*(1), 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 Systematics and Evolution*, *303*(8), 1013–1020. https://doi.org/10.1007/s00606-016-1356-9

Flutre, T., Duprat, E., Feuillet, C., & Quesneville, H. (2011). Considering Transposable Element Diversification in De Novo Annotation Approaches. *PLoS ONE*, *6*(1), e16526. https://doi.org/10.1371/journal.pone.0016526

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. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117

Garrido-Ramos, M. A. (2017). Satellite DNA: An evolving topic. *Genes*, *8*(9). https://doi.org/10.3390/genes8090230

González-Tizón, A. M., Rojo, V., Menini, E., Torrecilla, Z., & Martínez-Lage, A. (2013). Karyological analysis of the shrimp palaemon serratus (Decapoda: Palaemonidae). *Journal of Crustacean Biology*, *33*(6), 843–848. https://doi.org/10.1163/1937240X-00002185

Gregory, T. R. (2005). Chapter 1—Genome Size Evolution in Animals. In T. R. Gregory (Ed.), *The Evolution of the Genome* (pp. 3–87). Academic Press.

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. *Nature Ecology & Evolution*, *2*(3), 567–573. https://doi.org/10.1038/s41559-018-0467-9

Holt, C., Campbell, M., Keays, D. A., Edelman, N., Kapusta, A., Maclary, E., T. Domyan, E., Suh, A., Warren, W. C., Yandell, M., Gilbert, M. T. P., & Shapiro, M. D. (2018). Improved Genome Assembly and Annotation for the Rock Pigeon (Columba livia). *G3 Genes|Genomes|Genetics*, *8*(5), 1391–1398. https://doi.org/10.1534/g3.117.300443

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*(3), 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*(1), 1–9. https://doi.org/10.1093/gigascience/giaa160

Jurka, J., Kapitonov, V. V., Kohany, O., & Jurka, M. V. (2007). Repetitive sequences in complex genomes: Structure and evolution. *Annual Review of Genomics and Human Genetics*, *8*, 241–259. https://doi.org/10.1146/annurev.genom.8.080706.092416

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). *Frontiers in Marine Science*, *8*. https://doi.org/10.3389/fmars.2021.808354

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 Genes|Genomes|Genetics*, *11*(11). https://doi.org/10.1093/g3journal/jkab268

Kim, J.-H., Kim, H. K., Kim, H., Chan, B. K. K., Kang, S., & Kim, W. (2019). Draft Genome Assembly of a Fouling Barnacle, Amphibalanus amphitrite (Darwin, 1854): The First Reference Genome for Thecostraca. *Frontiers in Ecology and Evolution*, *7*. https://doi.org/10.3389/fevo.2019.00465

Kim, Y.-J., Lee, J., & Han, K. (2012). Transposable Elements: No More "Junk DNA." *Genomics & Informatics*, *10*(4), 226. https://doi.org/10.5808/GI.2012.10.4.226

Kojima, K. K. (2019). Structural and sequence diversity of eukaryotic transposable elements. *Genes and Genetic Systems*, *94*(6), 233–252. https://doi.org/10.1266/ggs.18-00024

Lanciano, S., & Mirouze, M. (2018). Transposable elements: all mobile, all different, some stress responsive, some adaptive? *Current Opinion in Genetics & Development*, *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. *Invertebrate Reproduction and Development*, *27*(2), 85–114. https://doi.org/10.1080/07924259.1995.9672440

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 Oceanologica Sinica*, *35*(6), 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. *Frontiers in Marine Science*, *9*. https://doi.org/10.3389/fmars.2022.895119

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, *21*(10), 597–614. https://doi.org/10.1038/s41576-020-0236-x

Lyon, M. F. (2006). Do LINEs Have a Role in X-Chromosome Inactivation? *BioMed Research International*, *2006*(1). https://doi.org/10.1155/JBB/2006/59746

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*(1), 427. https://doi.org/10.1186/1471-2164-8-427

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. *Scientific Reports*, *11*(1), 7533. https://doi.org/10.1038/s41598-021-87237-4

Miga, K. H. (2015). Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosome Research*, *23*(3), 421–426. https://doi.org/10.1007/s10577-015-9488-2

Mravinac, B., Plohl, M., & Ugarković, Đ. (2005). Preservation and High Sequence Conservation of Satellite DNAs Suggest Functional Constraints. *Journal of Molecular Evolution*, *61*(4), 542–550. https://doi.org/10.1007/s00239-004-0342-y

Niiyama, H. (1962). On the Unprecedentedly Large Number of Chromosomes of the Crayfish, Astacus Trowbridgii Stimpson. *Annotationes Zoologicae Japonenses*, *35*, 229–233.

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 Research*, *45*(12), e111–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*(6), 792–793. https://doi.org/10.1093/bioinformatics/btt054

Ou, S., & Jiang, N. (2018). LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiology*, *176*(2), 1410–1422. https://doi.org/10.1104/pp.17.01310

Paajanen, P., Kettleborough, G., López-Girona, E., Giolai, M., Heavens, D., Baker, D., Lister, A., Cugliandolo, F., Wilde, G., Hein, I., Macaulay, I., Bryan, G. J., & Clark, M. D. (2019). A critical comparison of technologies for a plant genome sequencing project. *GigaScience*, *8*(3). https://doi.org/10.1093/gigascience/giy163

Palacios-Gimenez, O. M., Koelman, J., Palmada-Flores, M., Bradford, T. M., Jones, K. K., Cooper, S. J. B., Kawakami, T., & Suh, A. (2020). Comparative analysis of morabine grasshopper genomes reveals highly abundant transposable elements and rapidly proliferating satellite DNA repeats. *BMC Biology*, *18*(1), 199. https://doi.org/10.1186/s12915-020-00925-x

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 Evolutionary Biology*, *19*(1), 1–15. 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 M. A. Garrido-Ramos (Ed.), *Genome Dynamics* (Volume 7, pp. 153–169). Karger.

Pita, S., Panzera, F., Mora, P., Vela, J., Cuadrado, Á., Sánchez, A., Palomeque, T., & Lorite, P. (2017). Comparative repeatome analysis on Triatoma infestans Andean and Non-Andean lineages, main vector of Chagas disease. *PLOS ONE*, *12*(7), e0181635. https://doi.org/10.1371/journal.pone.0181635

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*(1–2), 72–82. https://doi.org/10.1016/j.gene.2007.11.013

Plohl, M., Meštrović, N., & Mravinac, B. (2012). Satellite DNA evolution. *Genome Dynamics*, *7*(July), 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. *Science Advances*, *7*(26). https://doi.org/10.1126/sciadv.abe8290

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics*, *10*(4), 354–366. https://doi.org/10.1093/bib/bbp026

Poynton, H. C., Hasenbein, S., Benoit, J. B., Sepulveda, M. S., Poelchau, M. F., Hughes, D. S. T., Murali, S. C., Chen, S., Glastad, K. M., Goodisman, M. A. D., Werren, J. H., Vineis, J. H., Bowen, J. L., Friedrich, M., Jones, J., Robertson, H. M., Feyereisen, R., Mechler-Hickson, A., Mathers, N., … Richards, S. (2018). The Toxicogenome of Hyalella azteca : A Model for Sediment Ecotoxicology and Evolutionary Toxicology. *Environmental Science & Technology*, *52*(10), 6009–6022. https://doi.org/10.1021/acs.est.8b00837

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics*, *21*(Suppl 1), i351–i358. https://doi.org/10.1093/bioinformatics/bti1018

Rasch, E. M., Lee, C. E., & Wyngaard, G. A. (2004). DNA–Feulgen cytophotometric determination of genome size for the freshwater-invading copepod Eurytemora affinis. *Genome*, *47*(3), 559–564. https://doi.org/10.1139/g04-014

Reynolds, J., Souty-Grosset, C., & Richardson, A. (2013). Ecological roles of crayfish in freshwater and terrestrial habitats. *Freshwater Crayfish*, *19*(2), 197–218. https://doi.org/10.5869/fc.2013.v19-2.197

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., Stelzer, C.-P., Solorzano, E., Thomas, W. K., Pfrender, M. E., & Ebert, D. (2014). An SNP-based second-generation genetic map of Daphnia magna and its application to QTL analysis of phenotypic traits. *BMC Genomics*, *15*(1), 1033. https://doi.org/10.1186/1471-2164-15-1033

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. *Scientific Reports*, *6*(June), 1–14. https://doi.org/10.1038/srep28333

Shao, C., Sun, S., Liu, K., Wang, J., Li, S., Liu, Q., Deagle, B. E., Seim, I., Biscontin, A., Wang, Q., Liu, X., Kawaguchi, S., Liu, Y., Jarman, S., Wang, Y., Wang, H.-Y., Huang, G., Hu, J., Feng, B., … Fan, G. (2023). The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights. *Cell*, 1–16. https://doi.org/10.1016/j.cell.2023.02.005

Shapiro, J. A., & Von Sternberg, R. (2005). Why repetitive DNA is essential to genome function. *Biological Reviews of the Cambridge Philosophical Society*, *80*(2), 227–250. https://doi.org/10.1017/S1464793104006657

Shi, L., Yi, S., & Li, Y. (2018). Genome survey sequencing of red swamp crayfish Procambarus clarkii. *Molecular Biology Reports*, *45*(5), 799–806. https://doi.org/10.1007/s11033-018-4219-3

Silva, B. S. M. L., Picorelli, A. C. R., & Kuhn, G. C. S. (2023). In Silico Identification and Characterization of Satellite DNAs in 23 Drosophila Species from the Montium Group. *Genes*, *14*(2). https://doi.org/10.3390/genes14020300

SILVA, J. C., SHABALINA, S. A., HARRIS, D. G., SPOUGE, J. L., & KONDRASHOV, A. S. (2003). Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genetical Research*, *82*(1), S0016672303006268. https://doi.org/10.1017/S0016672303006268

Slotkin, R. K., & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, *8*(4), 272–285. https://doi.org/10.1038/nrg2072

Smit, A., Hubley, R., & Green, P. (2013). *RepeatMasker Open-4.0*.

Sotero-Caio, C. G., Platt, R. N., Suh, A., & Ray, D. A. (2017). Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biology and Evolution*, *9*(1), 161–177. https://doi.org/10.1093/gbe/evw264

Souty- Grosset, C., Holdich, D. M., Noël, P. Y., Reynolds, J. D., & Haffner, P. (2006). Atlas of Crayfish in Europe. *Patrimoines Naturels Collection*, *64*(187).

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 Research*, *33*(10), 1708–1717. https://doi.org/10.1101/gr.277387.122

Swathi, A., Shekhar, M. S., Katneni, V. K., & Vijayan, K. K. (2018). Genome size estimation of brackishwater fishes and penaeid shrimps by flow cytometry. *Molecular Biology Reports*, *45*(5), 951–960. https://doi.org/10.1007/s11033-018-4243-3

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. *Frontiers in Genetics*, *11*(March). https://doi.org/10.3389/fgene.2020.00201

Tang, B., Wang, Z., Liu, Q., Wang, Z., Ren, Y., Guo, H., Qi, T., Li, Y., Zhang, H., Jiang, S., Ge, B., Xuan, F., Sun, Y., She, S., Yam Chan, T., Sha, Z., Jiang, H., Li, H., Jiang, W., … Li, Y. (2021). Chromosome-level genome assembly of Paralithodes platypus provides insights into evolution and adaptation of king crabs. *Molecular Ecology Resources*, *21*(2), 511–525. https://doi.org/10.1111/1755-0998.13266

Tang, B., Wang, Z., Liu, Q., Zhang, H., Jiang, S., Li, X., Wang, Z., Sun, Y., Sha, Z., Jiang, H., Wu, X., Ren, Y., Li, H., Xuan, F., Ge, B., Jiang, W., She, S., Sun, H., Qiu, Q., … Li, Y. (2020). High-Quality Genome Assembly of Eriocheir japonica sinensis Reveals Its Unique Genome Evolution. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.01340

Tang, B., Zhang, D., Li, H., Jiang, S., Zhang, H., Xuan, F., Ge, B., Wang, Z., Liu, Y., Sha, Z., Cheng, Y., Jiang, W., Jiang, H., Wang, Z., Wang, K., Li, C., Sun, Y., She, S., Qiu, Q., … Ren, Y. (2020). Chromosome-level genome assembly reveals the unique genome evolution of the swimming crab (Portunus trituberculatus). *GigaScience*, *9*(1). https://doi.org/10.1093/gigascience/giz161

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 Research*, *47*(21), 10994–11006. https://doi.org/10.1093/nar/gkz841

Tran Van, P., Anselmetti, Y., Bast, J., Dumas, Z., Galtier, N., Jaron, K. S., Martens, K., Parker, D. J., Robinson-Rechavi, M., Schwander, T., Simion, P., & Schön, I. (2021). First annotated draft genomes of nonmarine ostracods (Ostracoda, Crustacea) with different reproductive modes. *G3 (Bethesda, Md.)*, *11*(4). https://doi.org/10.1093/g3journal/jkab043

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, *13*(1), 36–46. https://doi.org/10.1038/nrg3117

Uengwetwanit, T., Pootakham, W., Nookaew, I., Sonthirod, C., Angthong, P., Sittikankaew, K., Rungrassamee, W., Wongsurawat, T., Jenjaroenpun, P., Sangsrakru, D., Khudet, J., Koehorst, J. J., Schaap, P. J., Karoonuthaisiri, N., Science, N., Agency, T. D., Thani, P., Science, N., Agency, T. D., … Unit, V. (2020). *A chromosome-level assembly of the black tiger shrimp (*.

Utsunomia, R., Silva, D. M. Z. de A., Ruiz-Ruano, F. J., Goes, C. A. G., Melo, S., Ramos, L. P., Oliveira, C., Porto-Foresti, F., Foresti, F., & Hashimoto, D. T. (2019). Satellitome landscape analysis of Megaleporinus macrocephalus (Teleostei, Anostomidae) reveals intense accumulation of satellite sequences on the heteromorphic sex chromosome. *Scientific Reports*, *9*(1), 5856. https://doi.org/10.1038/s41598-019-42383-8

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), 1–16. https://doi.org/10.1186/s12864-021-07636-9

Vitales, D., Garcia, S., & Dodsworth, S. (2020). Reconstructing phylogenetic relationships based on repeat sequence similarities. *Molecular Phylogenetics and Evolution*, *147*(February), 106766. https://doi.org/10.1016/j.ympev.2020.106766

Wang, Q., Ren, X., Liu, P., Li, J., Lv, J., Wang, J., Zhang, H., Wei, W., Zhou, Y., He, Y., & Li, J. (2022). Improved genome assembly of Chinese shrimp (Fenneropenaeus chinensis) suggests adaptation to the environment during evolution and domestication. *Molecular Ecology Resources*, *22*(1), 334–344. https://doi.org/10.1111/1755-0998.13463

Wang, S. Y., Biesiot, P. M., & Skinner, D. M. (1999). Toward an Understanding of Satellite DNA Function in Crustacea. *American Zoologist*, *39*(3), 471–486.

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. *Nature Reviews Genetics*, *8*(12), 973–982. https://doi.org/10.1038/nrg2165

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. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1901), 20190079. https://doi.org/10.1098/rspb.2019.0079

Wu, C., & Lu, J. (2019). Diversification of Transposable Elements in Arthropods and Its Impact on Genome Evolution. *Genes*, *10*(5), 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*(5), 3274–3284. https://doi.org/10.1016/j.ygeno.2021.07.017

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

Zhang, X., Yuan, J., Sun, Y., Li, S., Gao, Y., Yu, Y., Liu, C., Wang, Q., Lv, X., Zhang, X., Ma, K. Y., Wang, X., Lin, W., Wang, L., Zhu, X., Zhang, C., Zhang, J., Jin, S., Yu, K., … Xiang, J. (2019). Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-018-08197-4

Zhu, L., Swergold, G. D., & Seldin, M. F. (2003). Examination of sequence homology between human chromosome 20 and the mouse genome: intense conservation of many genomic elements. *Human Genetics*, *113*(1), 60–70. https://doi.org/10.1007/s00439-003-0920-x

Chapter III

The extraordinary satellitome diversity of freshwater crayfish: a driver of genome evolution

Lena Bonassin, Ljudevit Luka Boštjančić, Christelle Rutz, Caterina Francesconi, Leonie Schardt, Damian Baranski, Carola Greve, Lucian Pârvulescu, Jelena Mlinarec, Višnja Besendorfer, Ivana Maguire, Kathrin Theissinger, Odile Lecompte

Under review in BMC Mobile DNA doi:10.21203/rs.3.rs-7499918/v1

Abstract

Background: Repetitive elements, particularly satellite DNA (satDNA), play a significant role in genome evolution and organisation. However, their diversity and evolutionary dynamics remain poorly understood in non-model organisms. Freshwater crayfish (Decapoda, Astacidea) have large genomes with a high chromosome number and are rich in satDNAs. This makes them attractive for studying the impact of satDNA on genome evolution.

Results: In this study, we investigated the repetitive genomic landscape of 19 species representing four freshwater crayfish families. Our analysis revealed a high proportion of repetitive DNA in all studied species, with the total repeat content ranging from 30% to 66%. The number of satDNA families was remarkably high (54–622 families per species), with minisatellites (<100 bp) forming the largest component of the satellitome. Family-specific patterns emerged: Astacidae and Cambaroididae showed the highest satDNA proportions, while Cambaridae and Parastacidae were dominated by Class I transposable elements. Species of the family Parastacidae showed the largest number of unique satDNA clusters and were clearly separated from other families, reflecting their phylogenetic divergence and distinct biogeographic history. We identified specific satDNAs conserved across all species, among them the PlSAT3-411, pointing to their important functional roles as pericentromeric satDNA.

Conclusion: This study provides the first comprehensive comparative analysis of satDNA in freshwater crayfish. Our results highlight the dynamic nature of repetitive DNA and underscore its importance in genome organisation and evolutionary history.

Keywords

satellite DNA, repetitive elements, Decapoda, library hypothesis, concerted evolution

3.1. Background

Freshwater crayfish are a group of taxonomically diverse decapod crustaceans belonging to two superfamilies: Astacoidea (families Astacidae, Cambaridae, Cambaroididae) and Parastacoidea (family Parastacidae) (Crandall & De Grave, 2017). Their distribution range

spans all continents except Antarctica and Africa (with the exception of Madagascar), with particularly high diversity in southeastern North America and southeastern Australia (Crandall & Buhay, 2007; Ion, Bloomer, et al., 2024). Based on phylogenetic evidence, the two superfamilies Astacoidea (in the Northern hemisphere) and Parastacoidea (in the Southern hemisphere) split around 261/268 Mya – 241 Mya (Bracken-Grissom et al., 2014; Wolfe et al., 2019). Within the superfamily Astacoidea, the oldest family Cambaroididae is distributed in East Asia, while the family Astacidae is distributed in Europe, with the exception of the genus *Pacifastacus*, which is found in western North America. The family Cambaridae, considered the youngest freshwater crayfish family, is distributed in eastern North America (Crandall & Buhay, 2007). Freshwater crayfish play crucial roles in aquatic ecosystems worldwide. They are considered keystone species and ecosystem engineers, contributing to various ecological processes in their environment (Holdich, 2002). In recent decades, population numbers of many freshwater crayfish species have experienced severe declines due to diverse anthropogenic influence (Ion, Ács, et al., 2024; Jussila et al., 2021) and are the subject of many conservation efforts (Ács et al., 2025; Bonassin et al., 2024; Schrimpf et al., 2017; Theissinger et al., 2022).

Freshwater crayfish families exhibit high diversity in their genomic characteristics (Rutz et al., 2023; Tan et al., 2020). Chromosome numbers in the infraorder Astacidea range from 2n = 102 in *Procambarus digueti* (family Cambaridae) up to 2n = 276 in *Procambarus virginalis* (family Cambaridae) (Boštjančić et al., 2021). This chromosomal diversity is further highlighted by significant variations in genome size and repetitive DNA content across different species and habitats. Within the infraorder Astacidea, freshwater crayfish generally possess larger genomes than the species occurring in shallow and deep water marine environments (Iannucci et al., 2022). In particular, genome sizes in freshwater crayfish range from 3.16 Gb in *Cherax quadricarinatus* (family Parastacidae) to 19.2 Gb in *Astacus astacus* (family Astacidae), while in marine species, genome sizes range between 4.16 Gb in *Hommarus gammarus* and 4.79 Gb in *Nephrops norvegicus* (Iannucci et al., 2022). A large portion of these giant genomes consists of repetitive elements (REs), with satellite DNA (satDNA) families being particularly abundant in freshwater crayfish (Boštjančić et al., 2021; Rutz et al., 2023). Thus, freshwater crayfish represent an interesting model for investigating the processes that drive the evolution of repeat-rich genomes.

The repetitive fraction of any eukaryote genome plays a crucial role in genome organisation, functioning and evolution (Šatović-Vukšić & Plohl, 2023). Repetitive DNA can broadly be

divided into transposable elements (TEs) and satDNA (Garrido-Ramos, 2017). TEs, including retrotransposons and DNA transposons, are characterised by their ability to proliferate, and are therefore widely dispersed within the genome. They can play a role in genome reorganisation, in regulation of gene expression and epigenetic processes (Bourque et al., 2018). SatDNA plays a crucial role in the formation of chromosomal structures and in genome stability, forming long tandem arrays primarily in telomeric, centromeric and pericentromeric regions of the genome (Shatskikh et al., 2020). In most animal and plant species, satDNA contributes to the formation of centromeres (Talbert & Henikoff, 2022). These long arrays allow the attachment of histones essential in the formation of the kinetochore. The ability of satDNAs to undergo homogenisation and rapid evolution is essential for maintaining their structural and functional roles in centromeric regions (Plohl et al., 2008). Pericentromeric satDNA PlSAT3-411 has been previously identified and characterised in the crayfish *Pontastacus leptodactylus,* where it has been found to colocalise with the pericentromeric regions on all chromosomes (Boštjančić et al., 2021). In the same species, the PlSAT57-664 satDNA showed complex units that include the complete PlSAT3-411 unit and four direct sub-repeats, but with lower abundance than the PlSAT3-411 (Boštjančić et al., 2021). The most abundant satDNAs are often implicated in the crucial and conserved function of centromeres. Despite their importance, centromeric sequences often do not share conserved properties such as monomer length, GC content, or common sequence motifs (Melters et al., 2013). These findings underscore the significance of satDNA organisation in understanding the evolution of centromeric regions in crayfish and other taxa.

SatDNA sequences exhibit great variability in monomer size, nucleotide sequence, genomic distribution and abundance among closely related species (Plohl et al., 2012). This unique collection of satDNAs distributed in the genome is termed the species' satellitome. The diversity of satDNA emphasises their rapid evolution while maintaining structural importance. Divergence of satDNA sequences can be detected at different taxonomic levels, with certain sequences diverging from population to species or higher taxonomic levels (Ugarković & Plohl, 2002). The library hypothesis predicts that related species share a common library of satDNAs inherited from a common ancestor, with differences primarily being quantitative due to differential amplification of certain variants (J. P. M. Camacho et al., 2022; Shatskikh et al., 2020; Veseljak et al., 2024). This rapid divergence between species is often explained by concerted evolution, a pattern emerging from processes like satDNA amplification and homogenisation that maintains sequence similarity within a

species while allowing independent changes to accumulate in different lineages (Šatović-Vukšić & Plohl, 2023; Ugarković & Plohl, 2002). SatDNA has long been challenging to study due to the limitations of sequencing technologies and bioinformatic tools (Louzada et al., 2020). Nowadays, low coverage genome skimming data used in combination with assembly free software allows *de novo* repeat identification from short read sequencing data (Louzada et al., 2020; Theissinger et al., 2023). This approach enables comprehensive studies across taxa, including those with large genomes whose assemblies are often still missing (Lower et al., 2018). Despite their biological importance, the evolutionary dynamics of satDNA, and their potential impact on intra- and interspecific diversification are still poorly understood in many invertebrate groups.

This study aims to elucidate the characteristics and distribution of repetitive DNA, specifically satDNA, across 19 species belonging to four freshwater crayfish families: Astacidae, Cambaridae, Cambaroididae, and Parastacidae. Previous research has highlighted large genome sizes with a high proportion of REs and satDNA for several crayfish species (Boštjančić et al., 2021; Rutz et al., 2023). Given this previously observed abundance of satDNA in crayfish, and the known rapid evolution of these sequences in other taxa, we hypothesise that satDNA abundance and diversity vary among freshwater crayfish families, with diversification of satDNA reflecting family-specific evolutionary events. Through comparative analyses, we aim to identify conserved and lineage specific repeat sequences and investigate the relationship between the phylogenetic structure and satellitome composition of the species. Understanding the composition and evolution of repetitive DNA in freshwater crayfish is crucial for unravelling the mechanisms that drive genome expansion, structural diversity, and adaptation in this ecologically relevant group.

3.2. Methods

3.2.1. Sampling and genomic DNA extraction

The male individual of *Astacus astacus* was purchased from the crayfish farm Flusskrebszucht Frömel (Kavelstorf, Germany). The male individual of *Austropotamobius torrentium* was collected from the stream Dolje (Podsused, Croatia) with the permission of the of Croatian Ministry of Economy and Sustainable Development (517-10-1-2-22-4). *Pacifastacus leniusculus* was collected from the river Korana (Croatia) and *Faxonius immunis* from Neuburg am Rhein (Germany). One adult male individual of *A. bihariensis*

was collected from the Valea Iadului river in Romania (46,7447 N 22,5597 E) with the necessary authorisation from the Romanian Academy (1/CJ/13.01.2021), the Romanian Ministry of Water and Forests (DGB/2/R5787/16.08.2022), the Apuseni Nature Park Administration (199/09.09.2022), the National Agency for Protected Areas (882/15.09.2022), and the Environmental Protection Agencies in the geographical area where the specimen was sampled (76/20.09.2022).

Genomic DNA was extracted using a 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 proteins and cellular debris the samples were centrifuged at 5000 x g for 10 min, and to precipitate the DNA the samples were centrifuged at 5000 x g for 5 min. Finally, the DNA pellet was resuspended in 100 µL nuclease free water. DNA was quantified using the QuantiFluor® dsDNA System on the Quantus™ Fluorometer (Promega, USA).

3.2.2. Flow cytometry analysis

The genome size was estimated for *Astacus astacus* and *Austopotamobius bihariensis* following a flow cytometry protocol with propidium iodide-stained nuclei (Hare & Johnston, 2012). For each species, haemolymph of a −80 °C frozen adult sample and neural tissue of the internal reference standard *Acheta domesticus* (female, 1C = 2Gb) was mixed with 2 mL of chopping buffer. Three different buffers were used for different measurements: Galbraith's buffer (Galbraith et al., 1983), Phosphate buffer saline and Otto's buffer (Otto, 1992). The suspension was filtered through a 42-µm nylon mesh and stained with the intercalating fluorochrome propidium iodide (PI, Thermo Fisher Scientific) and treated with RNase II A (Sigma-Aldrich), each with a final concentration of 25 µg/mL. The mean red PI fluorescence of stained nuclei was quantified using a CytoFLEX flow cytometer (Beckman-Coulter, USA) with a solid-state laser emitting at 488 nm. Fluorescence intensities of 5000 nuclei per sample were recorded. We used the CytExpert 2.3 software for histogram analyses. The total quantity of DNA in the sample was calculated as the ratio of the mean fluorescence signal of the 2C peak of the stained nuclei of the crayfish sample divided by the mean fluorescence signal of the 2C peak of the stained nuclei of the reference standard times the 1C amount of DNA in the reference standard. The genome size is reported as 1C, the mean amount of DNA in Gbp in a haploid nucleus.

3.2.3. Next-Generation Sequencing

Extracted DNA was fragmented using the Bioruptor® Pico sonication device (Diagenode, Hologic Inc., Liege, Belgium) for 21 cycles of 30 s ON followed by 30 s OFF. Illumina libraries were prepared according to the BEST protocol (Carøe et al., 2018). Preparation of PCR reactions was automated using a Biomek i7 Hybrid workstation (Beckman Coulter, Brea, CA, USA). Library lengths were verified on a TapeStation system (Agilent Technologies, Santa Clara, CA, USA). Libraries of all species were barcoded, pooled, and sequenced on an Illumina NovaSeq 6000 at Novogene (Cambridge, UK) to obtain 2 x 150 bp paired-end reads. Sequence quality was assessed using FastQC v0.11.9 (Andrews, 2010) and quality trimming was performed using the Trimmomatic software (Bolger et al., 2014). The reads generated during the current study have been deposited in the NCBI SRA repository, BioProjectID PRJNA1293697, sample accession numbers SAMN50031846-SAMN50031850.

3.2.4. Identification and annotation of repetitive DNA

For the identification of repetitive DNA, the following species were chosen from all crayfish families: *Astacus astacus*, *Austropotamobius torrentium*, *Austropotamobius pallipes*, *Austropotamobius bihariensis*, *Pontastacus leptodactylus*, *Pacifastacus leniusculus*, *Faxonius immunis*, *Faxonius limosus*, *Procambarus clarkii*, *Procambarus acutus*, *Cambarus robustus*, *Cambaroides japonicus*, *Cambaroides dauricus*, *Cambaroides schrenckii*, *Cambaroides similis*, *Cherax destructor*, *Cherax robustus*, *Cherax quadricarinatus*, *Parastacus brasiliensis*. Accession numbers of Illumina paired-end reads obtained in this study and accession numbers of datasets from publicly available studies from the European Nucleotide Archive (ENA) are listed in Supplementary table III-1. Quality control, read pre-processing and RepeatExplorer2 (Galaxy Version 2.3.12.1) analysis were performed following the protocol described in (Novák et al., 2020). Reads of each analysed species were filtered against a customised database containing mitochondrial sequences of the species using the RepeatExplorer Utilities: Preprocessing of FASTQ paired-end reads (Galaxy Version 1.0.0.3). The GenBank accession numbers of the mitochondrial sequences used for filtering the reads are listed in Supplementary table III-1. Similarity-based clustering analysis using RepeatExplorer2 (Novák et al., 2013) was performed using 500 000 reads. The reconstruction of monomer sequences of individual satDNA families was performed using TAREAN

analysis (Novák et al., 2017). After individual clustering analysis, the reads were concatenated and subjected to comparative analysis using RepeatExplorer2.

3.2.5. Repeat classification and sequence analysis

Repeat classification was done following the procedure described in (Boštjančić et al., 2021). After *de novo* identification of repetitive elements in RepeatExplorer2 (Galaxy Version 2.3.12.1), contigs were further classified using Censor (Kohany et al., 2006) and with similarity searches using BLASTN 2.5.0. (C. Camacho et al., 2009) and BLASTX 2.5.0. (C. Camacho et al., 2009) against public databases and against the repetitive elements identified in Rutz et al. (2023). Similarity between the clusters obtained from comparative analysis and individual RepeatExplorer2 runs were analysed using BLASTN 2.5.0.

Statistical analyses were conducted in R version 4.2.1 (R Core Team, 2024). The normality of GC content and length of satDNA sequences was assessed using the Shapiro Wilk's test, and the correlation between the two variables was calculated using the Spearman rank correlation test. The distribution of GC content and length of satDNA sequences was assessed using the Wilcoxon test. Comparisons of repeat unit length and GC content among all freshwater crayfish genera and between families was assessed using a Kruskal-Wallis test. Post-hoc pairwise Wilcoxon rank-sum tests with Bonferroni correction were applied to identify specific group differences. For all analyses, the significance level used was $\alpha=0.05$.

3.2.6. Phylogenetic reconstruction and divergence analysis

Phylogenetic reconstruction was based on (Vitales et al., 2020) using a custom bash and R script (Supplementary file III-1 and Supplementary file III-2). A distance matrix was calculated based on the observed/expected number of edges in clusters between species obtained in comparative analysis using the distance.comb function from the R package sidier v4.1.0 (Pajares, 2013). A dendrogram was built using the R package pvclust v2.2.0 (Suzuki et al., 2019) with ward.D2 as the method for hierarchical clustering, Bray–Curtis dissimilarity distance method and 1 000 bootstrap replications. The heatmap was constructed using the R package ComplexHeatmap v2.20.0 (Gu, 2022; Gu et al., 2016).

The satDNA consensus sequences obtained from the comparative analysis were used to estimate divergence. First, for all sequences dimers or higher repeat numbers up to 200 nt were generated using the dimerator.py script.This collection was used as a reference for running the RepeatMasker software against reads from each species. RepeatMasker v4.1.2-p1

(Smit et al., 2013) was used with the parameters -a -nolow -no_is. The sequence divergence distribution was calculated as Kimura distances using the RepeatMasker tool calcDivergenceFromAlign.pl.

3.2.7. Detailed characterisation of PlSAT3-411 and PlSAT57-664

SatDNAs from each species with similarity to the PlSAT3-411 and PlSAT57-664 were extracted using BLASTN (C. Camacho et al., 2009) and a consensus sequence was produced using Bowtie2 (Langmead & Salzberg, 2012). For both sequences, dimers were produced using the dimerator.py script (https://github.com/fjruizruano/ngs-Protocols/blob/master/dimerator.py, 2025). Patterns in the PlSAT3-411 and PlSAT57-664 in all species were examined applying the RepeatProfiler workflow (Https://github.com/johnssproul/RepeatProfiler, 2025; Negm et al., 2021). For the correlation analysis, species were grouped by family. Both consensus sequences were used as reference for the RepeatMasker software against reads from each species.

3.2.8. Preparation of chromosome spreads and fluorescence *in situ* hybridisation (FISH)

Chromosome spreads were prepared according to the method described in (Mlinarec et al., 2011). Specific primers for the satDNA sequence PlSAT3-411 (Boštjančić et al., 2021) were used for amplification of probes for FISH. PCRs were performed using GoTaq Green Master Mix (Promega, USA) in 25 μL reactions: 12.5 μL GoTaq® Green Master Mix, 2.5 μL of each primer, 1 μL DNA. The PCR program consisted of 3 min denaturation at 95°C, 35 cycles each with 1 min denaturation at 95°C, 30 s annealing at 55°C, 1 min extension at 72°C, and a final extension of 20 min. Amplicons were visualised on a 2% agarose gel and purified from gel slices using the ReliaPrepTM DNA Clean-Up and Concentration System (Promega, USA). Cloning was performed using a pGEM-T Easy Vector System (Promega, USA) according to the manufacturer's protocol. Individual clones were purified using the PureYield™ Plasmid Miniprep System (Promega, USA) and sequenced by Macrogen (Amsterdam, The Netherlands).

Plasmid vectors containing the satDNA monomer sequence were labelled with Aminoallyl-dUTP-Cy3 (Jena Bioscience GmbH, Jena, Germany) using the Nick Translation Reagent Kit (Abbott Molecular Inc., USA) according to the manufacturer's protocol with the following modifications: plasmid DNA (700 ng) was labelled in a reaction volume of 25 μL using 2.5

µL of enzyme mixture for 6 h at 15°C. FISH was performed according to (Boštjančić et al., 2021). The preparations were mounted in Dako Fluorescence Mounting Medium (Dako North America Inc., USA) and stored at 4°C overnight. Signals were visualised using an Olympus BX51 microscope, equipped with a cooled CCD camera (Olympus DP70).

3.3. Results

3.3.1. Repeat classification and sequence analysis

The genome size of *A. astacus* and *A. bihariensis* was measured from haemolymph. Results showed that the average 1C DNA value was 16.89 Gbp for *A. astacus* and 11.58 Gbp for *A. bihariensis* (Supplementary table III-2). The number of obtained reads per species (*A. torrentium*, *A. bihariensis*, *F. immunis*, *P. leniusculus*, *A. astacus)* resulting from low coverage sequencing ranged from 13.5 M to 40.8 M reads corresponding to 0.15 – 0.53 x coverage. (Supplementary table III-1). Reads obtained in this study and from public databases were used in the *de novo* identification of repeats in 19 freshwater crayfish species (Supplementary table III-1) with the RepeatExplorer2 pipeline based on low-coverage Illumina reads. The clusters identified by RepeatExplorer2 are shown in Supplementary table III-3. The number of identified clusters ranged between 620 in *A. astacus* and 756 in *P. leptodactylus*. The read proportion of each element, grouped by repeat type, is shown in Figure III-1. All sequences were annotated as satDNA, ribosomal DNA (rDNA), TEs belonging to Class I and Class II, or repeats (sequences without detailed annotation). The total proportion of REs in each species ranged from 30% in *P. clarkii* to 66% in *C. similis*. Ribosomal DNA (rDNA) was identified with 0.08% to 1.5% abundance per species, except for *P. leniusculus*, *A. torrentium* and *P. brasiliensis* where no sequences were assigned to rDNA sequences (Figure III-1, Supplementary table III-3). In each genome, the most abundant REs were either annotated to satDNAs or Class I TEs, which were represented by LINE and LINE/DIRS elements. In the Astacidae family, the highest proportion of reads was represented by satDNA sequences ranging from 17% in *A. torrentium* to 41.7% in *P. leniusculus*. The second highest proportion of reads in the Astacidae family belonged to Class I TEs, while no Class II elements were annotated. In the Cambaridae family, the highest proportion of reads were assigned to Class I TEs (between 16% in *P. acutus* and 24% in both species of the genus *Faxonius*), while 8 to 14% of sequences were assigned to satDNA. Class II TE were assigned to 0.01% (Maverick element) and 0.05% (Helitron element) of the sequences in *F. immunis* and *P. acutus*, respectively. In the Cambaroididae family, most

sequences were assigned to Class I TEs in *C. japonicus* and C. *dauricus* with 19% and 27%, respectively, while in *C. similis* and *C. schrenckii* most sequences were assigned to satDNA with 57% and 27%, respectively. In *C. japonicus*, 0.01% of reads were assigned to Class II TE Maverick. In the Parastacidae family, most sequences were assigned to Class I TEs, ranging between 20% in *Ch. robustus* and 29% in *C. quadricarinatus*. Among the studied species, distinct satDNA sequences dominate their respective genomes (Supplementary table III-3, Supplementary table III-4). In genomes with the highest proportion of satDNA, the most abundant satDNA are CsiSAT1-17 in *C. similis* with 38% of reads assigned, PleSAT1-21 in *P. leniusculus* with 22% of reads assigned and ApaSAT1-21 in *A. pallipes*.

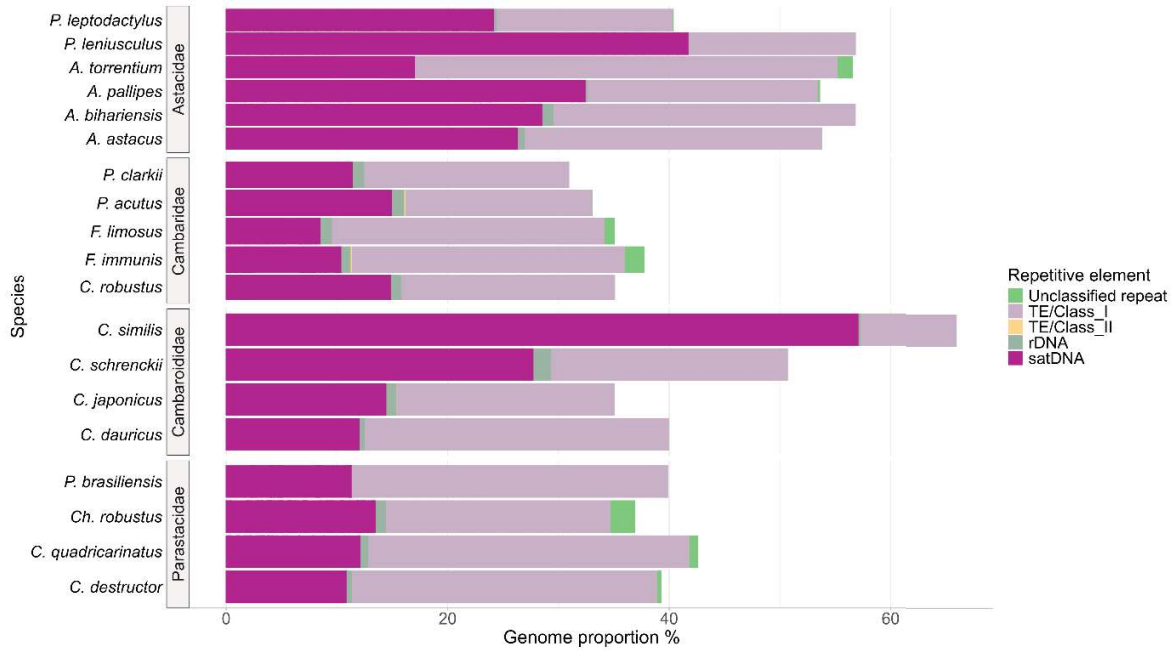


Figure III-1. Proportions of repetitive elements in genomic reads of the 19 freshwater crayfish species used in this study. Each bar represents a species, with colours indicating different repetitive element categories. Repetitive elements are annotated as satellite DNA (satDNA), ribosomal DNA (rDNA), TEs belonging to Class I and Class II, and sequences without detailed annotation indicated as Unclassified repeats. Proportions are on a scale from 0 to 65 %.

The range of the repeat unit length of satDNA sequences was from 12 to 2657 bp for the Astacidae family, from 14 to 4898 bp for the Cambaridae family, from 12 to 1920 bp for the Cambaroididae family and 13 to 2700 bp for the Parastacidae family (Supplementary table III-3). The distribution of the repeat unit length of satDNA sequences is concentrated below 100 bp (Figure III-2). The mean repeat unit length for minisatellite sequences (below 100 bp) were 46.9 bp for Astacidae, 35.2 bp for Cambaridae, 38.1 bp for Cambaroididae and 37.1 bp

for Parastacidae. The mean repeat unit length for macrosatellite sequences (above 100 bp) were 201.7 bp for Astacidae, 622.1 bp for Cambaridae, 368.3 bp for Cambaroididae and 608.5 bp for Parastacidae (Figure III-2). Overall, the length of repeat unit sequences of satellites was significantly different between genera (Kruskal-Wallis test, p = 5.157e-68) and between families (Kruskal-Wallis tests, p = 2.537e-70) (Supplementary table III-5). From now on, we will refer to satDNA sequences below 100 bp as minisatellites, satDNA sequences above 100 bp as macrosatellites, and all satDNA as simply satellites.

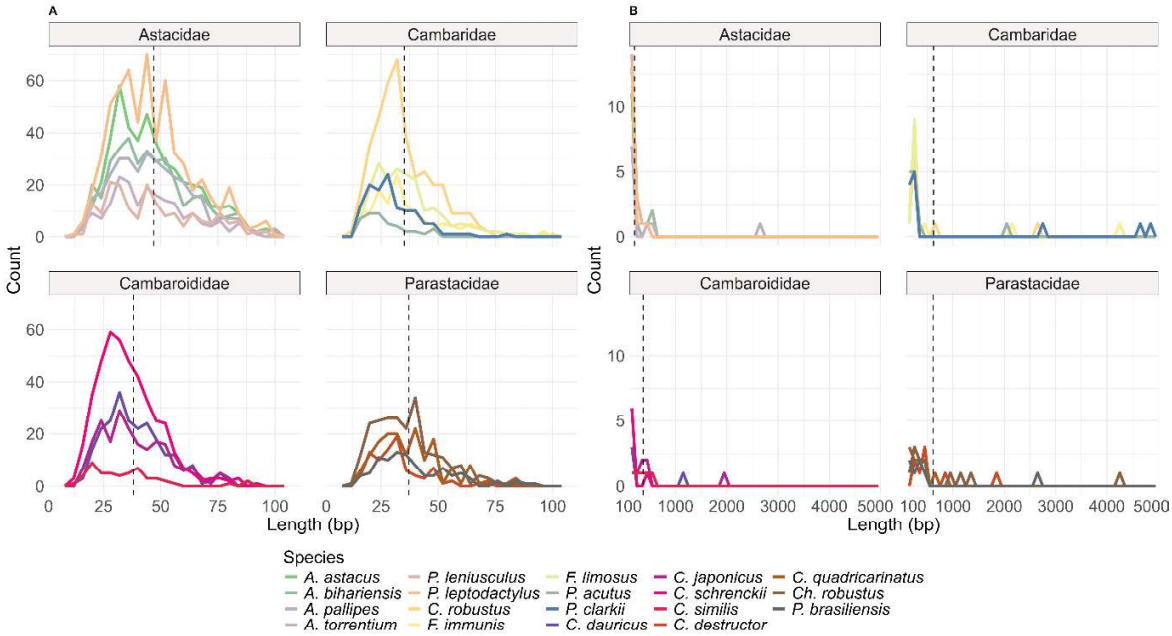


Figure III-2. Frequency of satDNA sequences length across studied species. (A) Minisatellite DNA sequences ≤100 bp. (B) Macrosatellite DNA sequences >100 bp. Each colour represents a species. Vertical dashed lines represent the mean length value for each family.

The GC content of the satDNA sequences across all studied species varied between 17.2% (PbrSAT14-29 in *P. brasiliensis*) and 80% (FimSAT163-15 in *F. immunis*). The median GC content value for minisatellites was 49.19% and 46.15% for macrosatellites (Supplementary table III-3). The GC content differed significantly between macrosatellites and minisatellites (Wilcoxon rank-sum test, p = 1.048e-05) (Figure III-3). We observed two different populations of macrosatellites in terms of GC content in the genera *Astacus*, *Austropotamobius*, *Pacifastacus*, *Pontastacus* and *Cambarus*: one peak of satDNA distributed at around 30% GC content and a second peak at 50% GC content (Figure III-3). GC content was different for all satDNA between genera (Kruskal-Wallis test, p= 5.252 e-67)

and between families (Kruskal-Wallis test, p = 1.0156 e-56) (Supplementary table III-5). A negative correlation was observed between GC content and repeat unit length overall in satellite DNA (Spearman's rho = -0.056, p = 1.581 e-04) and in minisatellites (Spearman's rho = -0.035, p = 0.0205). On family level, a negative correlation was observed only for the family Astacidae and Cambaridae for satDNA (Supplementary figure III-1, Supplementary table III-6).

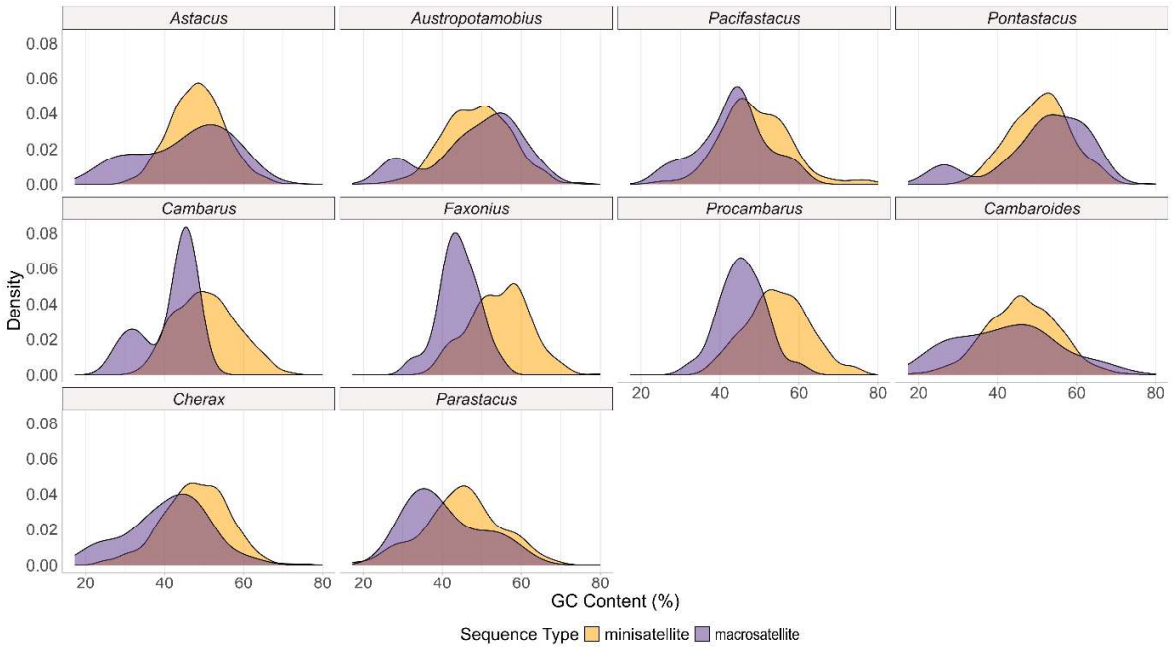


Figure III-3. Density plot representing the GC content distribution for minisatellite and macrosatellite sequences by crayfish genus. Colours represent sequence type minisatellite (< 100 bp) and macrosatellite (> 100 bp)

3.3.2. Phylogenetic reconstruction and divergence analysis

Hierarchical clustering was performed based on the clusters obtained from comparative analysis of all species with RepeatExplorer2. Based on the cluster similarities, two clusters are observed: one for species of the family Parastacidae, and another for all the other freshwater crayfish families. Species belonging to the same family clustered together except for *F. limosus* and *P. clarkii* (family Cambaridae) that grouped together with the species of the family Astacidae (Supplementary figure III-2).

Based on comparative analysis, we identified in total 395 RE clusters (Figure III-4). The highest number of RE clusters were present in the species *A. torrentium* and the lowest in *P. brasiliensis* (Figure III-4.A). Most of the identified clusters appeared in the species within the family Astacidae, while the lowest number of shared RE clusters was present between the species of the family Parastacidae and all other species (Figure III-4.B). The number of unique clusters per species was generally low (<5), except for *P. brasiliensis and P. leniusculus* with 18 and 11 unique clusters respectively (Figure III-4.C).

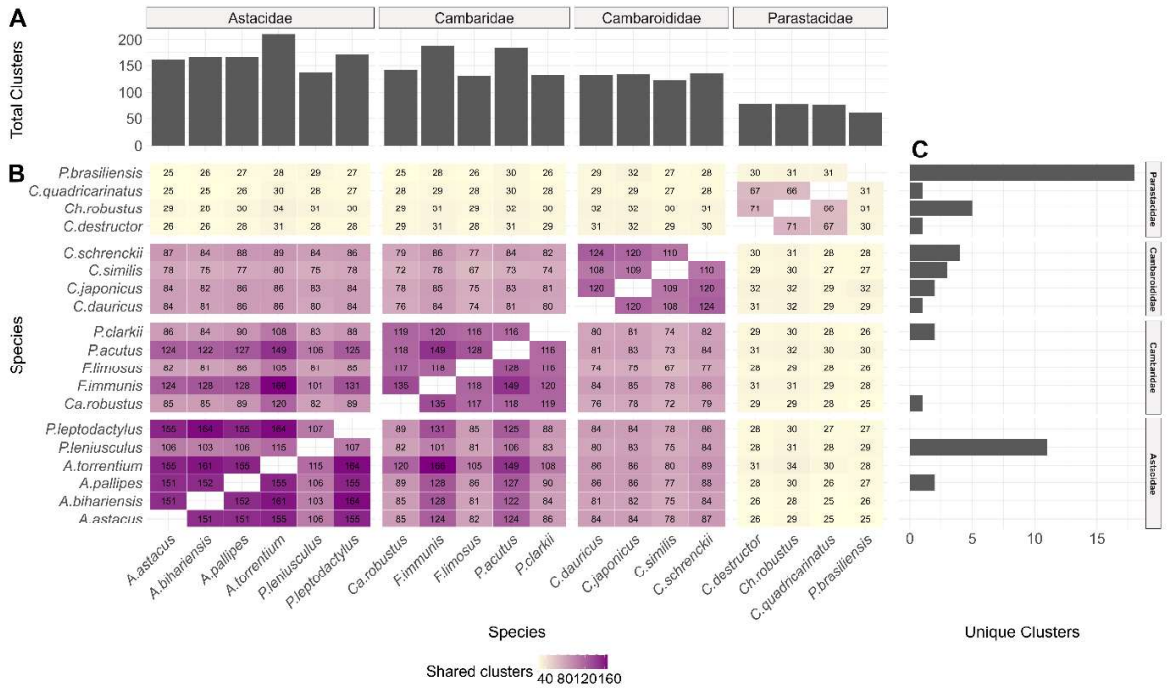


Figure III-4. A) Total number of RE clusters per species identified in comparative analysis by Repeat Explorer2. B) Number of share clusters between species. C) Number of unique clusters per species.

Sequence divergence landscapes were calculated for all satDNA sequences in each species. The analysis revealed distinct patterns across the crayfish species and families (Figure III-5). In the family Astacidae, a peak was observed at lower substitution levels (0 – 10%). In Cambaridae, distinct patterns were observed between species of the genus *Faxonius* (one peak at 0 – 10%), the genus *Procambarus* (multiple peaks between 0 - 20%), and *Ca. robustus* (peak between 5 – 20%). In the family Cambaroididae, distinct patterns were observed between *C. similis and C. schrenckii* (one peak at 0 – 10%), and between *C. japonicus* and *C. dauricus* (one peak at 0 – 10 % and another peak at 20 – 30%). In *Ch. robustus*, multiple peaks were present at 0 -30% divergence, while the other species in the

family Parastacidae show one peak at lower substitution levels (0 – 10%). The overall satDNA content varies between species as indicated by the different coverage (Figure III-5).

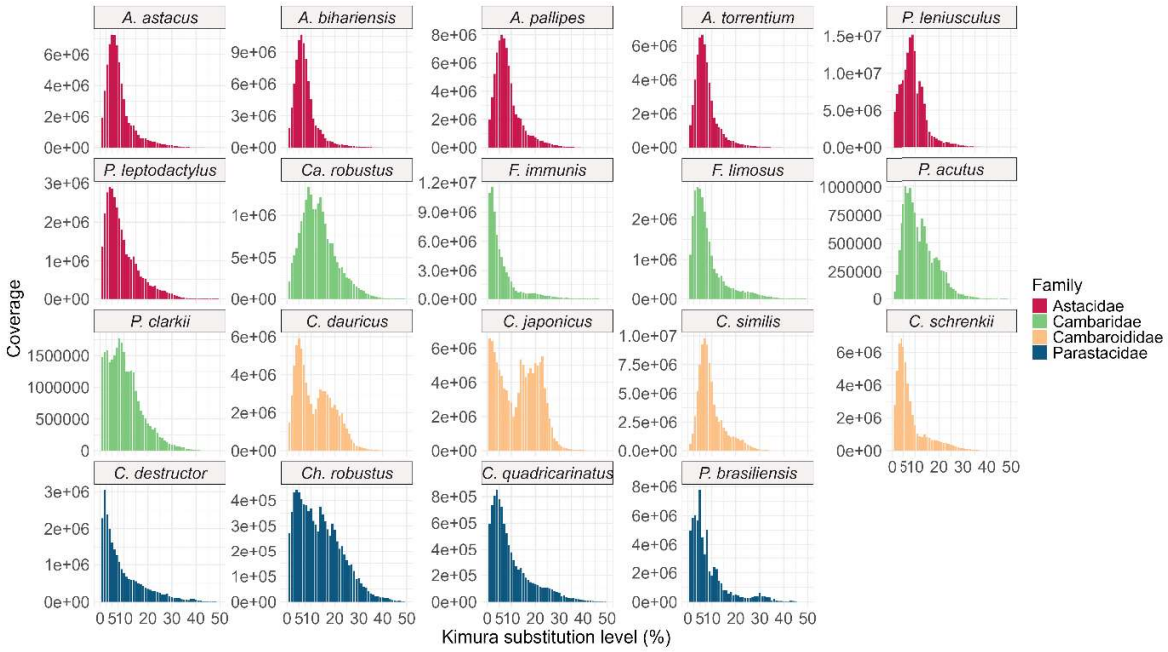


Figure III-5. SatDNA sequence divergence landscapes for each species. Colours indicate different freshwater crayfish families. y-axes are scaled for each species.

3.3.3. Detailed characterisation of PlSAT3-411 and PlSAT57-664

The PlSAT3-411 satDNA family was previously identified as pericentromeric satellite DNA in *P. leptodactylus* (Boštjančić et al., 2021) and was selected for further analysis. Dimer consensus sequences of the PlSAT3-411 sequence were used to obtain repeat profiles of the sequence in each species. The colour enhanced profiles showed the presence of the entire sequence in the species from the Astacidae and Cambaroididae family with similar depth (Supplementary figure III-3). In the Cambaridae family, only parts of the sequence were mapped, while the sequence is not present in the Parastacidae family. Variant profiles show family-specific signatures in the pattern of variants relative to the consensus PlSAT3-411 sequence (Figure III-6).

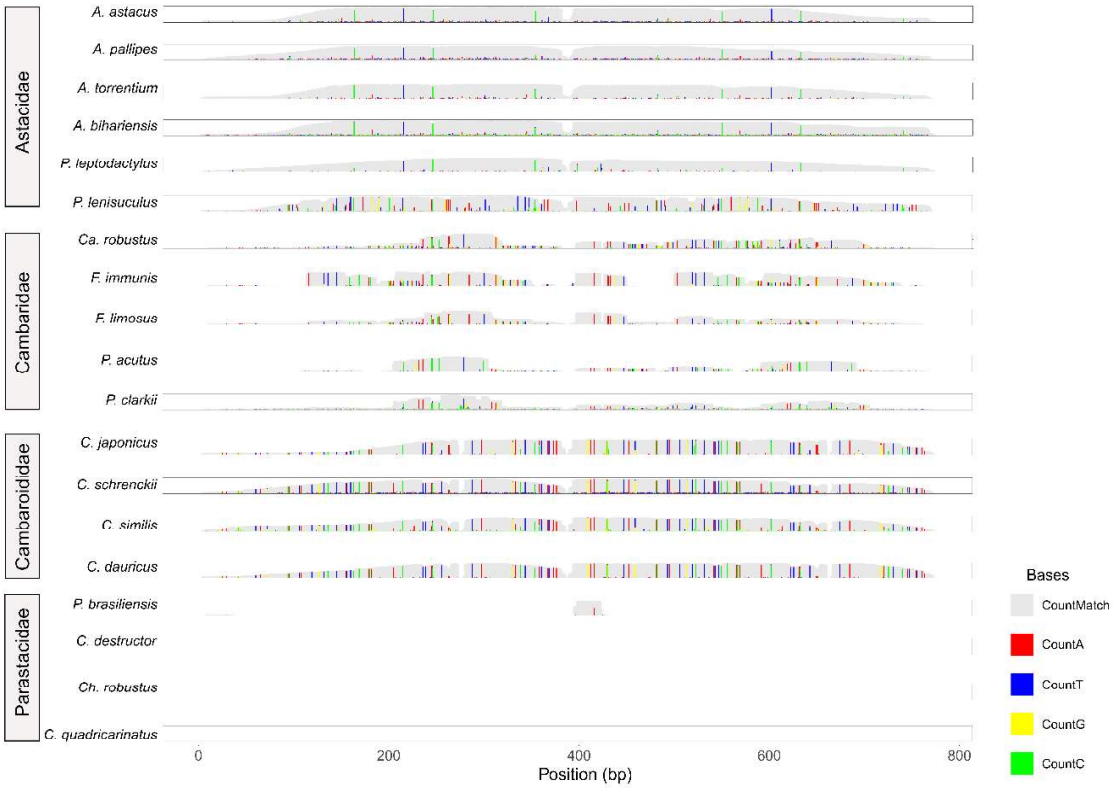


Figure III-6. Variant repeat profiles of PlSAT3-411 satellite DNA family across the studied species. Different colours indicate A, T, C and G bases. The height of each bar indicates the coverage of each variant in the reads.

The distribution of Kimura substitution levels was explored for the PlSAT3-411 and PlSAT57-664 sequences in each species. The landscapes varied among species and between the two satellite sequences. In the Astacidae and Cambaroididae family, the PlSAT3-411 sequence showed high coverage and a peak at lower substitution levels (0 - 5%). In contrast, the PlSAT57-664 sequence showed 10^5-fold lower coverage than the PlSAT3-411 sequence. The pattern was opposite for the species in the Cambaridae family, with the PlSAT57-664 sequence having a double coverage compared to the PlSAT3-411 in all genera, except in the genus *Faxonius* where the coverage of the PlSAT57-664 sequence was 10^5-fold higher than the PlSAT3-411 sequence. In the genera *Cambarus* and *Procambarus* (Cambaridae family) the divergence peak was present at 5 – 10% for the PlSAT3-411 sequence and 10 – 30% for the PlSAT57-664 sequence. In the genus *Faxonius*, the PlSAT57-664 showed a peak at 0 – 10 % divergence with higher coverage for *F. immunis* than *F. limosus* (Figure III-7). To check for the chromosomal localisation of the PlSAT3-411, FISH was performed on metaphase spreads of *A. torrentium* and *P. leniusculus*. The chromosome

mapping revealed hybridisation sites in the pericentromeric regions on all chromosomes in both species (Figure III-8).

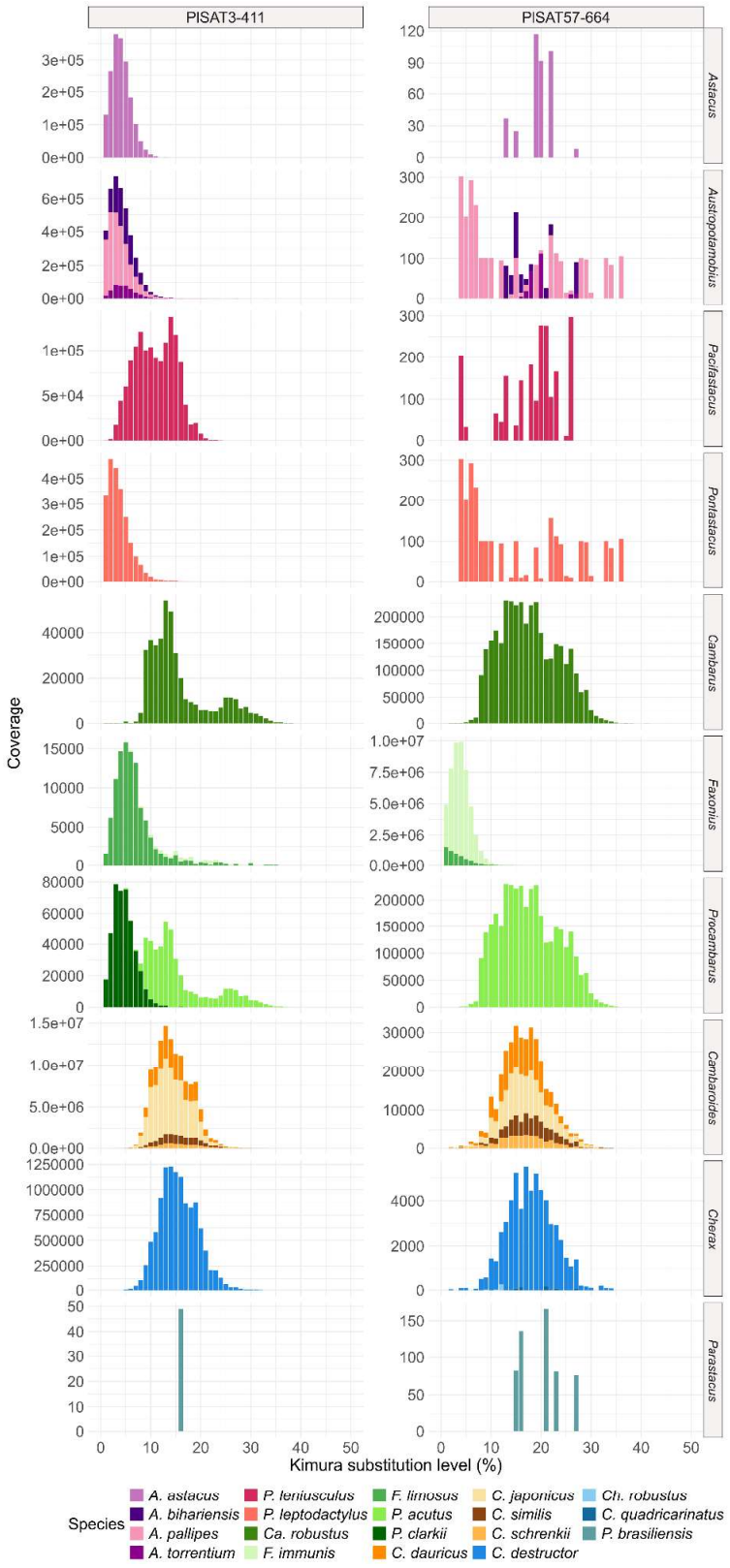


Figure III-7. Sequence divergence landscapes for each species for the PlSAT3-411 (left) and PlSAT57-664 (right) sequences. Colours indicate different species. y-axis is scaled for each species.

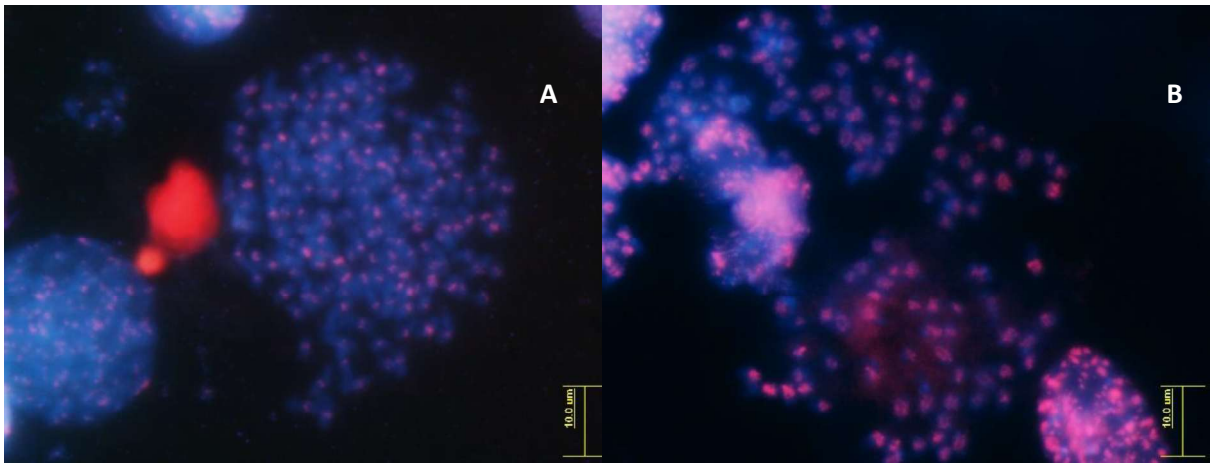


Figure III-8. Localisation of PlSAT3-411 satellite repeat family (in red) on metaphase chromosomes of (A) *A. torrentium* and (B) *P. leniusculus*. Red signals represent the Cy3 probe localisation, chromosomes are counterstained with DAPI. Scale bar = 10µm.

3.4. Discussion

We investigated the repetitive genomic landscape of freshwater crayfish to gain novel understanding of the genome evolution through satellite DNA in 19 species across four families. In line with our hypothesis, our findings highlight universal patterns present across-freshwater crayfish families, as well as lineage specific sequences, providing unique insights into repetitive element diversification in freshwater crayfish. The relevance of our results regarding our research hypothesis is discussed below.

3.4.1. High number of satellite DNA families in all freshwater crayfish species

The high number of REs identified across all studied species aligns with previous findings in Decapoda, where the proportion of REs constitutes 58% to 79% of genomes (Rutz et al., 2023). Our results confirm an overall high repetitive DNA content in all species, ranging from 30% in *P. clarkii* to 66% in *C. similis* (Figure III-1). We observed family-specific patterns in the distribution of transposable elements and satDNA, with Astacidae and Cambaroididae showing the highest proportion of satDNA, while Cambaridae and Parastacidae genomes were dominated by Class I transposable elements (Figure III-1). The lack or low number of identified Class II TEs observed here has already been reported in other Decapoda (Rutz et al., 2023; Xu et al., 2023). Class I TEs tend to accumulate in low-

recombination regions such as peri-centromeric heterochromatin, while Class II are more common in gene rich regions (Wells & Feschotte, 2020). The “copy and paste” proliferation mechanism of Class I TE can lead to TE copy accumulation, contributing to larger genome sizes (Wells & Feschotte, 2020). The prevalence of TEs has been identified as a key factor underlying genome size variation in Decapoda (Liu et al., 2024; Rutz et al., 2023).

Eukaryotes exhibit a wide array of satDNA families, ranging from 2 in the plant *Tanacetum cinerariifolium* (Mlinarec et al., 2019) to 226 in the frog *Proceratophrys boiei* (João Da Silva et al., 2023) and 258 in *P. leptodactylus* (Boštjančić et al., 2021). Even within genera, satDNA counts can vary greatly, as seen in the beetle *Tribolium freemani* with 135 satDNA families and *T. castaneum* with 57 satDNA (Gržan et al., 2023; Veseljak et al., 2024). In our study, the number of satDNA families varies greatly, but is overall high, between 54 and 622 (Supplementary table III-3). This illustrates the remarkable diversification of the satellitome in freshwater crayfish. A high proportion of satDNA in all species in this study is formed by minisatellites (<100 bp), especially sequences shorter than 50 bp (Figure III-2). A high abundance of minisatellites has also been found in other Decapoda genomes (Molina et al., 2020; Zhang et al., 2019). Minisatellites are often associated with euchromatic regions (Garrido-Ramos, 2017; Zhao et al., 2012), but they can also be associated with long arrays present in pericentromeric and subtelomeric regions (Garrido-Ramos, 2017). In *P. leptodactylus,* the minisatellites PlSAT6-70 and PlSAT14-79 are located interstitially along the chromosome arms, considered part of euchromatic regions (Boštjančić et al., 2021). The formation of new minisatellite sequences arises from DNA replication, DNA recombination or DNA repair (Richard et al., 2008). Some minisatellites found in euchromatic regions or within genes appear to be essential for protein function (Richard et al., 2008). Furthermore, it has been proposed that minisatellite sequences facilitate interchromosomal DNA exchanges (Pathak & Ali, 2012; Piazza et al., 2012). Based on this, the presence of a large number of minisatellites in freshwater crayfish could allow for rapid adaptation of the organism to the environment.

3.4.2. SatDNAs among freshwater crayfish show shared characteristics

Our study revealed distinct satDNA sequence characteristics across the 19 species belonging to the four extant crayfish families, including variations in repeat unit length, GC content (Supplementary table III-3), and substitution landscapes (Figure III-5). We observed a

negative correlation between GC content and satellite length overall for satellites and minisatellites (Figure III-3, Supplementary figure III-1). A negative correlation has been found in the grasshopper *Locusta migratoria* and in the fish *Astyanax paranae*, where short satDNA tend to arise from GC-rich regions, while longer monomers are more AT-rich (Ruiz-Ruano et al., 2016; Silva et al., 2017). Other studies in fish genera *Megaloporinus*, *Psalidodon* and *Astyanax* did not establish correlation between these two characteristics (Goes et al., 2022; Utsunomia et al., 2019). It is generally considered that AT-rich satDNA can induce DNA curvature, important for facilitating the tight packing of DNA and proteins in heterochromatin (Ugarković, 2005). Heterochromatin acts as a scaffold for nuclear architecture, enabling the efficient organisation and proper chromosome segregation during mitosis and meiosis (Jagannathan et al., 2018), therefore ensuring the proper organisation of the large number of chromosomes.

3.4.3. Phylogenetic signal of satDNA

Among the analysed freshwater crayfish families, species of the family Parastacidae shows the highest number of unique clusters (Figure III-4). Hierarchical clustering based on the repeat profiles separated the family Parastacidae from the other families (Supplementary figure III-2), consistent with their distinct biogeographic history and deep phylogenetic divergence from Northern Hemisphere crayfish (Crandall & De Grave, 2017). The clustering of the three other families into a single group reflects higher repeat similarity, despite their taxonomic diversity. While hierarchical clustering of satDNA profiles aligns with the divergence of Parastacidae, the clustering of the other crayfish families suggests that satDNA evolution is also influenced by lineage-specific homogenisation processes within the whole superfamily Astacoidea. Repeatome based analyses have revealed clear phylogenetic signals in different plant groups (Herklotz et al., 2021; Vitales et al., 2020), and several studies showed the use of single satDNA as phylogenetic markers (Dias et al., 2021; Utsunomia et al., 2017). The clustering of species in our study based on satDNA does not fully mirror their taxonomy. Stronger phylogenetic signals could arise by performing clustering analysis at the family or genus level or by focusing only on a smaller number of satDNA sequences. Clustering can be further influenced by high intraspecific diversity and the presence of multiple divergent satDNA subfamilies. Low sequence divergence peaks in the family Astacidae and in two Cambaroididae species, *C. similis* and *C. schrenckii* (0 – 10%, Figure III-5) suggest the rapid spread and amplification of similar repeat units within a species, resulting in a high abundance of recently homogenised sequences that have not yet

accumulated many mutations (Palomeque & Lorite, 2008; Rico-Porras et al., 2024). This could reflect rapid and localised speciation events or rapid population increases potentially driven by paleohydrogeological changes. Conversely multiple and broader peaks present in *Ca. robustus*, *P. acutus*, *Ch. robustus*, *C. dauricus* and *C. japonicus* (Figure III-5) indicate that different subfamilies within a satDNA family are diverging independently, thus creating more sequence variants. This pattern suggests older and more complex speciation histories which aligns with the wide distribution and large number of species present in the families Parastacidae and Cambaridae or reflects distinct diversification events within the *Cambaroides* genus.

3.4.4. Conserved satDNAs at the core of the freshwater crayfish satellitome

The presence of certain satDNA sequences with low sequence divergence (as seen in the PlSAT3-411, Figure III-6, Figure III-7) can be attributed to the concerted evolution and strong purifying selection in functionally constrained genomic regions (Thakur et al., 2021). SatDNA sequences found near important functional regions such as ribosomal loci or within euchromatic regions can exhibit minimal divergence due to functional constraints and potentially deleterious effects of mutations (Schaper, 2014; Thakur et al., 2021). The distribution of the PlSAT3-411 sequence throughout all the species aligns with patterns of concerted evolution, where the repetitive DNA sequences maintain a greater similarity among repeats within a species than between species (Garrido-Ramos, 2017; Plohl et al., 2012). This conservation may reflect functional roles in centromeric/pericentromeric regions, while its absence in Parastacidae suggests deep evolutionary divergence. Due to abundance of PlSAT57-664 in the family Parastacidae and its sequence similarity to PlSAT3-411, we can speculate that PlSAT57-664 family may have a similar role across these species, however, this hypothesis should be confirmed in future by chromosomal localisation studies. The functional roles of centromeric satDNA include contributing to structural integrity of the centromere (Bloom, 2014), maintenance of the heterochromatin, epigenetic regulation through non-coding RNAs (Ugarković et al., 2022), recruitment of centromere specific proteins during cell replications and maintenance of genome stability (Shatskikh et al., 2020). High AT content and canonical size of 170 or 340 bp have been proposed to support the packaging of DNA in the heterochromatin, providing crucial satDNA function typical for centromeric satellites (Ugarković, 2005). The functional role of the PlSAT3-411 family, as

pericentromeric satDNA, has already been hypothesised in (Boštjančić et al., 2021). The contrasting abundance of PlSAT3-411 and PlSAT57-664 among the freshwater crayfish families parallels the library hypothesis of satDNA evolution, which suggests that related species share a common library of satDNA sequences inherited from the last common ancestor but differentially amplified across closely related lineages. This process leads to species-specific satDNA profiles, with similar sequences in different proportions observed among the species with a common ancestor (J. P. M. Camacho et al., 2022; Šatović-Vukšić & Plohl, 2023). While satDNA is usually highly dynamic, several studies have reported conservation over periods longer than 50 Mya (dos Santos et al., 2021). In molluscs, two satDNAs, BIV160 and PjHaaI, have been discovered to be conserved for 540 Mya (Petraccioli et al., 2015; Plohl et al., 2010). The divergence between Northern and Southern Hemisphere crayfish families (Parastacidae vs. Astacidae, Cambaridae, Cambaroididae) occurred around 241 Mya (Wolfe et al., 2019), suggesting that the PlSAT3-411 is at least 241 Mya old. This makes the PlSAT3-411 one of the most ancient satDNAs discovered so far.

3.4.5. The challenges of studying satellitomes of non-model organisms

The study of REs, especially satDNA, across diverse organisms poses computational challenges (Šatović-Vukšić & Plohl, 2021). satDNA consists of long arrays of highly repetitive sequences which makes them challenging to sequence and assemble leading to gaps in genome assemblies (Šatović-Vukšić & Plohl, 2021). Bioinformatic tools used in genome assembly methodologies, often misclassify, or fail to identify satDNA sequencing, making them underrepresented in databases. Moreover, satellitome studies are still often performed on one or a few species, with limited insight in diversity of satDNA across different genera or families. Here we focused on species representative of all four freshwater crayfish families. We analysed species from all genera in the families Astacidae and Cambaroididae, 23% of genera from the family Cambaridae and 15% of genera from the family Parastacidae. This approach allowed the capturing of a wide diversity of repetitive DNA patterns and identify lineage-specific satDNA trends. The assembly of reference genomes for species with large, highly repetitive genomes, such as freshwater crayfish, remains a significant challenge for current sequencing technologies. To date, only four freshwater crayfish genomes have been published, none featuring satDNA as a driver of genome evolution. Low coverage genome sequencing circumvents the challenges associated with assembling highly repetitive regions in large, complex genomes. This approach has been demonstrated to provide reliable repeat profiles across a range of diverse taxa. Although bioinformatic annotation of repetitive DNA

is complicated by the dynamic organisation and homology of satDNA (Lower et al., 2018), our study addressed these challenges by integrating established tools such as RepeatExplorer and TAREAN, which perform *de novo* identification of REs and satDNA. We further characterised the satDNA employing metrics like monomer clustering and GC-content analysis. These findings not only enrich our knowledge of crayfish genomes but also provide a replicable framework for studying satDNA in other non-model organisms.

3.5. Conclusion

Here we provide novel insights into the repetitive DNA and highlight conserved and lineage-specific patterns of satDNA evolution across 19 freshwater crayfish species from four families. Our results reveal a high number of satDNA families and a high proportion of REs in all species, confirming the important role of repetitive DNA in shaping freshwater crayfish genomes. Our results show distinct family-level patterns, with Parastacidae exhibiting the highest number of unique repeat clusters and phylogenetic separation from Northern Hemisphere families, consistent with their evolutionary divergence. The observed variation in satDNA repeat unit length, GC content, and substitution landscapes reveals dynamic processes, including concerted evolution and differential amplification, as predicted by the library hypothesis. The conservation of certain satDNAs across all species highlights their likely functional roles, particularly in centromeric or pericentromeric regions. This study advances our understanding of genome evolution in decapods and emphasises the value of satDNA in unravelling both evolutionary history and functional genome organisation in crustaceans.

Funding

This work was funded by the Agence Nationale de la Recherche (GEODE: ANR-21-CE02-0028), the Deutsche Forschungsgemeinschaft (GEODE: DFG 490760095/TH 1807/7-1), by the institutional project financed by the University of Zagreb and by the University of Zagreb Student union grant for the project Development of new karyotypization techniques of freshwater crayfish (family: Astacidae). Kathrin Theissinger is supported through the DFG Heisenberg programme (534452071/TH 1807/10-1). Ljudevit Luka Boštjančić is supported through the Deutsche Bundesstiftung Umwelt (39954/01). Lucian Pârvulescu is supported

through the grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI–UEFISCDI, project number PN-III-P4-ID-PCE-2020-1187, within PNCDI III.

Acknowledgments

The authors kindly acknowledge Juliane Romahn for the support with server management and data visualisation, Sandra Hudina for the collection of *Pacifastacus lenisuculus* samples and Johannes Meka for the collection of *Faxonius immunis* sample.

References

Ács, A. R., Ion, M. C., Miok, K., Laza, A. V., Pitic, A., Robnik-Šikonja, M., & Pârvulescu, L. (2025). Threats Assessment of the Endemic Idle Crayfish (Austropotamobius bihariensis Pârvulescu, 2019): Lessons From Long-Term Monitoring. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *35*(1). https://doi.org/10.1002/aqc.70033

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*.

Bloom, K. S. (2014). Centromeric Heterochromatin: The Primordial Segregation Machine. *Annual Review of Genetics*, *48*(1), 457–484. https://doi.org/10.1146/annurev-genet-120213-092033

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

Bonassin, L., Pârvulescu, L., Boštjančić, L. L., Francesconi, C., Paetsch, J., Rutz, C., Lecompte, O., & Theissinger, K. (2024). Genomic insights into the conservation status of the Idle Crayfish Austropotamobius bihariensis Pârvulescu, 2019: low genetic diversity in the endemic crayfish species of the Apuseni Mountains. *BMC Ecology and Evolution*, *24*(1), 78. https://doi.org/10.1186/s12862-024-02268-5

Boštjančić, L. L., Bonassin, L., Anušić, L., Lovrenčić, L., Besendorfer, V., Maguire, I., Grandjean, F., Austin, C. M., Greve, C., Hamadou, A. Ben, & Mlinarec, J. (2021). The Pontastacus leptodactylus (Astacidae) Repeatome Provides Insight Into Genome Evolution and Reveals Remarkable Diversity of Satellite DNA. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.611745

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 06 Biological Sciences 0604 Genetics. *Genome Biology*, *19*(1), 1–12. https://doi.org/10.1186/s13059-018-1577-z

Bracken-Grissom, H. D., Ahyong, S. T., Wilkinson, R. D., Feldmann, R. M., Schweitzer, C. E., Breinholt, J. W., Bendall, M., Palero, F., Chan, T. Y., Felder, D. L., Robles, R., Chu, K. H., Tsang, L. M., Kim, D., Martin, J. W., & Crandall, K. A. (2014). The emergence of lobsters: Phylogenetic relationships, morphological evolution and divergence time

comparisons of an ancient group (Decapoda: Achelata, astacidea, glypheidea, polychelida). *Systematic Biology*, *63*(4), 457–479. https://doi.org/10.1093/sysbio/syu008

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 1–9. https://doi.org/10.1186/1471-2105-10-421/FIGURES/4

Camacho, J. P. M., Cabrero, J., López-León, M. D., Martín-Peciña, M., Perfectti, F., Garrido-Ramos, M. A., & Ruiz-Ruano, F. J. (2022). Satellitome comparison of two oedipodine grasshoppers highlights the contingent nature of satellite DNA evolution. *BMC Biology*, *20*(1), 1–24. https://doi.org/10.1186/s12915-021-01216-9

Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S. S. T., Sinding, M. H. S., Samaniego, J. A., Wales, N., Sicheritz-Pontén, T., & Gilbert, M. T. P. (2018). Single-tube library preparation for degraded DNA. *Methods in Ecology and Evolution*, *9*(2), 410–419. https://doi.org/10.1111/2041-210X.12871

Crandall, K. A., & Buhay, J. E. (2007). Global diversity of crayfish (Astacidae, Cambaridae, and Parastacidae—Decapoda) in freshwater. In *Freshwater Animal Diversity Assessment* (pp. 295–301). Springer Netherlands. 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. *Journal of Crustacean Biology*, *37*(5), 615–653. https://doi.org/10.1093/jcbiol/rux070

Dias, C. A. R., Kuhn, G. C. S., Svartman, M., Santos Júnior, J. E. dos, Santos, F. R., Pinto, C. M., & Perini, F. A. (2021). Identification and characterization of repetitive DNA in the genus Didelphis Linnaeus, 1758 (Didelphimorphia, Didelphidae) and the use of satellite DNAs as phylogenetic markers. *Genetics and Molecular Biology*, *44*(2). https://doi.org/10.1590/1678-4685-gmb-2020-0384

dos Santos, R. Z., Calegari, R. M., de Andrade Silva, D. M. Z., Ruiz-Ruano, F. J., Melo, S., Oliveira, C., Foresti, F., Uliano-Silva, M., Porto-Foresti, F., & Utsunomia, R. (2021). A Long-Term Conserved Satellite DNA That Remains Unexpanded in Several Genomes of Characiformes Fish Is Actively Transcribed. *Genome Biology and Evolution*, *13*(2), 1–16. https://doi.org/10.1093/gbe/evab002

Galbraith, D. W., Harkins, K. R., Maddox, J. M., Ayres, N. M., Sharma, D. P., & Firoozabady, E. (1983). Rapid Flow Cytometric Analysis of the Cell Cycle in Intact Plant Tissues. *Science*, *220*(4601), 1049–1051. https://doi.org/10.1126/science.220.4601.1049

Garrido-Ramos, M. A. (2017). Satellite DNA: An evolving topic. *Genes*, *8*(9). https://doi.org/10.3390/genes8090230

Goes, C. A. G., dos Santos, R. Z., Aguiar, W. R. C., Alves, D. C. V., Silva, D. M. Z. de A., Foresti, F., Oliveira, C., Utsunomia, R., & Porto-Foresti, F. (2022). Revealing the Satellite DNA History in Psalidodon and Astyanax Characid Fish by Comparative Satellitomics. *Frontiers in Genetics*, *13*(June), 1–12. https://doi.org/10.3389/fgene.2022.884072

Gržan, T., Dombi, M., Despot-Slade, E., Veseljak, D., Volarić, M., Meštrović, N., Plohl, M., & Mravinac, B. (2023). The Low-Copy-Number Satellite DNAs of the Model Beetle Tribolium castaneum. *Genes*, *14*(5). https://doi.org/10.3390/genes14050999

Gu, Z. (2022). Complex heatmap visualization. *IMeta*, *1*(3). https://doi.org/10.1002/imt2.43

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

Hare, E. E., & Johnston, J. S. (2012). Genome Size Determination Using Flow Cytometry of Propidium Iodide-Stained Nuclei. In V. Orgogozo & M. Rockman (Eds.), *Molecular Methods for Evolutionary Genetics. Methods in Molecular Biology*. Humana Press. https://doi.org/10.1007/978-1-61779-228-1_1

Herklotz, V., Kovařík, A., Wissemann, V., Lunerová, J., Vozárová, R., Buschmann, S., Olbricht, K., Groth, M., & Ritz, C. M. (2021). Power and Weakness of Repetition – Evaluating the Phylogenetic Signal From Repeatomes in the Family Rosaceae With Two Case Studies From Genera Prone to Polyploidy and Hybridization (Rosa and Fragaria). *Frontiers in Plant Science*, *12*(December). https://doi.org/10.3389/fpls.2021.738119

Holdich, D. M. (2002). Distribution of crayfish in Europe and some adjoining countries. *Bulletin Français de La Pêche et de La Pisciculture*, *367*, 611–650. https://doi.org/10.1051/kmae:2002055

https://github.com/fjruizruano/ngs-Protocols/blob/master/dimerator.py. (2025). *dimerator.py*. https://github.com/fjruizruano/ngs- protocols/blob/master/dimerator.py

Https://github.com/johnssproul/RepeatProfiler. (2025). *RepeatProfiler*. https://github.com/johnssproul/RepeatProfiler

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. *Frontiers in Ecology and Evolution*, *10*(August), 1–15. https://doi.org/10.3389/fevo.2022.930888

Ion, M. C., Ács, A.-R., Laza, A. V., Lorincz, I., Livadariu, D., Lamoly, A. M., Goia, B., Togor, A., Iorgu, E. I., Ştefan, A., Popa, O. P., & Pârvulescu, L. (2024). Conservation status of the idle crayfish Austropotamobius bihariensis Pârvulescu, 2019. *Global Ecology and Conservation*, *50*, e02847. https://doi.org/10.1016/j.gecco.2024.e02847

Ion, M. C., Bloomer, C. C., Bărăscu, T. I., Oficialdegui, F. J., Shoobs, N. F., Williams, B. W., Scheers, K., Clavero, M., Grandjean, F., Collas, M., Baudry, T., Loughman, Z., Wright, J. J., Ruokonen, T. J., Chucholl, C., Guareschi, S., Koese, B., Banyai, Z. M., Hodson, J., … Pârvulescu, L. (2024). World of Crayfish^{TM}: a web platform towards real-time global mapping of freshwater crayfish and their pathogens. *PeerJ*, *12*, e18229. https://doi.org/10.7717/peerj.18229

Jagannathan, M., Cummings, R., & Yamashita, Y. M. (2018). A conserved function for pericentromeric satellite DNA. *ELIFE*, *7*(e34122). https://doi.org/10.7554/eLife.34122.001

Jenkins, T. L., Ellis, C. D., & Stevens, J. R. (2019). SNP discovery in European lobster (Homarus gammarus) using RAD sequencing. *Conservation Genetics Resources*, *11*(3), 253–257. https://doi.org/10.1007/s12686-018-1001-8

João Da Silva, M., Gazoni, T., Haddad, C. F. B., & Parise-Maltempi, P. P. (2023). Analysis in Proceratophrys boiei genome illuminates the satellite DNA content in a frog from the Brazilian Atlantic forest. *Frontiers in Genetics*, *14*.

https://doi.org/10.3389/fgene.2023.1101397

Jussila, J., Edsman, L., Maguire, I., Diéguez-Uribeondo, J., & Theissinger, K. (2021). Money Kills Native Ecosystems: European Crayfish as an Example. *Frontiers in Ecology and Evolution*, *9*. https://doi.org/10.3389/fevo.2021.648495

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*(1), 474. https://doi.org/10.1186/1471-2105-7-474

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

Liu, Z., Zheng, J., Li, H., Fang, K., Wang, S., He, J., Zhou, D., Weng, S., Chi, M., Gu, Z., He, J., Li, F., & Wang, M. (2024). Genome assembly of redclaw crayfish (Cherax quadricarinatus) provides insights into its immune adaptation and hypoxia tolerance. *BMC Genomics*, *25*(1). https://doi.org/10.1186/s12864-024-10673-9

Louzada, S., Lopes, M., Ferreira, D., Adega, F., Escudeiro, A., Gama-carvalho, M., & Chaves, R. (2020). Architecture and Plasticity — An Evolutionary and Clinical A ffair. *Genes*.

Lower, S. S., McGurk, M. P., Clark, A. G., & Barbash, D. A. (2018). Satellite DNA evolution: old ideas, new approaches. In *Current Opinion in Genetics and Development* (Vol. 49, pp. 70–78). Elsevier Current Trends. https://doi.org/10.1016/j.gde.2018.03.003

Melters, D. P., Bradnam, K. R., Young, H. A., Telis, N., May, M. R., Ruby, J. G., Sebra, R., Peluso, P., Eid, J., Rank, D., Garcia, J. F., DeRisi, J. L., Smith, T., Tobias, C., Ross-Ibarra, J., Korf, I., & Chan, S. W. L. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology*, *14*(1). https://doi.org/10.1186/gb-2013-14-1-r10

Mlinarec, J., Mužić, 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*(12–13), 1497–1510. https://doi.org/10.1163/156854011X607015

Mlinarec, J., Skuhala, A., Jurković, A., Malenica, N., McCann, J., Weiss-Schneeweiss, H., Bohanec, B., & Besendorfer, V. (2019). The repetitive DNA composition in the natural pesticide producer Tanacetum cinerariifolium: Interindividual variation of subtelomeric tandem repeats. *Frontiers in Plant Science*, *10*(May), 1–14. https://doi.org/10.3389/fpls.2019.00613

Molina, W. F., Costa, G. W. W. F., Cunha, I. M. C., Bertollo, L. A. C., Ezaz, T., Liehr, T., & Cioffi, M. B. (2020). Molecular cytogenetic analysis in freshwater prawns of the genus macrobrachium (Crustacea: Decapoda: Palaemonidae). *International Journal of Molecular Sciences*, *21*(7), 1–12. https://doi.org/10.3390/ijms21072599

Negm, S., Greenberg, A., Larracuente, A. M., & Sproul, J. S. (2021). RepeatProfiler: A pipeline for visualization and comparative analysis of repetitive DNA profiles. *Molecular Ecology Resources*, *21*(3), 969–981. https://doi.org/10.1111/1755-0998.13305

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 Research*, *45*(12), e111–e111. https://doi.org/10.1093/nar/gkx257

Novák, P., Neumann, P., & Macas, J. (2020). Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nature Protocols*, *15*(11), 3745–3776. https://doi.org/10.1038/s41596-020-0400-y

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*(6), 792–793. https://doi.org/10.1093/bioinformatics/btt054

Otto, F. (1992). Preparation and staining of cells for high-resolution DNA analysis. In A. Radbruch (Ed.), *Flow cytometry and cell sorting* (pp. 101–104). Springer-Verlag.

Pajares, A. (2013). SIDIER: substitution and indel distances to infer evolutionary relationships. *Methods in Ecology and Evolution*, *4*, 1195–1200.

Palomeque, T., & Lorite, P. (2008). Satellite DNA in insects: a review. *Heredity*, *100*(6), 564–573. https://doi.org/10.1038/hdy.2008.24

Pathak, D., & Ali, S. (2012). Repetitive DNA: A Tool to Explore Animal Genomes/Transcriptomes. In *Functional Genomics*. InTech. https://doi.org/10.5772/48259

Petraccioli, A., Odierna, G., Capriglione, T., Barucca, M., Forconi, M., Olmo, E., & Biscotti, M. A. (2015). A novel satellite DNA isolated in Pecten jacobaeus shows high sequence similarity among molluscs. *Molecular Genetics and Genomics*, *290*(5), 1717–1725. https://doi.org/10.1007/s00438-015-1036-4

Piazza, A., Serero, A., Boulé, J.-B., Legoix-Né, P., Lopes, J., & Nicolas, A. (2012). Stimulation of Gross Chromosomal Rearrangements by the Human CEB1 and CEB25 Minisatellites in Saccharomyces cerevisiae Depends on G-Quadruplexes or Cdc13. *PLoS Genetics*, *8*(11), e1003033. https://doi.org/10.1371/journal.pgen.1003033

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*(1–2), 72–82. https://doi.org/10.1016/j.gene.2007.11.013

Plohl, M., Meštrović, N., & Mravinac, B. (2012). Satellite DNA evolution. *Genome Dynamics*, *7*(July), 126–152. https://doi.org/10.1159/000337122

Plohl, M., Petrović, V., Luchetti, A., Ricci, A., Šatović, E., Passamonti, M., & Mantovani, B. (2010). Long-term conservation vs high sequence divergence: the case of an extraordinarily old satellite DNA in bivalve mollusks. *Heredity*, *104*(6), 543–551. https://doi.org/10.1038/hdy.2009.141

R Core Team. (2024). *R: A language and environment for statistical computing* (4.4.1.).

Richard, G.-F., Kerrest, A., & Dujon, B. (2008). Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes. *Microbiology and Molecular Biology Reviews*, *72*(4), 686–727. https://doi.org/10.1128/mmbr.00011-08

Rico-Porras, J. M., Mora, P., Palomeque, T., Montiel, E. E., Cabral-de-Mello, D. C., & Lorite, P. (2024). Heterochromatin Is Not the Only Place for satDNAs: The High

Diversity of satDNAs in the Euchromatin of the Beetle Chrysolina americana (Coleoptera, Chrysomelidae). *Genes*, *15*(4). https://doi.org/10.3390/genes15040395

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. *Scientific Reports*, *6*(June), 1–14. 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*(8). https://doi.org/10.3390/genes14081627

Šatović-Vukšić, E., & Plohl, M. (2021). Classification Problems of Repetitive DNA Sequences. *DNA*, *1*(2), 84–90. https://doi.org/10.3390/dna1020009

Šatović-Vukšić, E., & Plohl, M. (2023). Satellite DNAs—From Localized to Highly Dispersed Genome Components. In *Genes* (Vol. 14, Issue 3). https://doi.org/10.3390/genes14030742

Schaper, E. (2014). The Evolution of Protein Tandem Repeats [Gothenburg University]. In *ETH Zurich*. https://doi.org/10.3929/ethz-a-010276597

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. *Conservation Genetics*, *18*(6), 1299–1315. https://doi.org/10.1007/s10592-017-0981-3

Shatskikh, A. S., Kotov, A. A., Adashev, V. E., Bazylev, S. S., & Olenina, L. V. (2020). Functional Significance of Satellite DNAs: Insights From Drosophila. *Frontiers in Cell and Developmental Biology*, *8*(May), 1–19. https://doi.org/10.3389/fcell.2020.00312

Silva, D. M. Z. de A., Utsunomia, R., Ruiz-Ruano, F. J., Daniel, S. N., Porto-Foresti, F., Hashimoto, D. T., Oliveira, C., Camacho, J. P. M., & Foresti, F. (2017). High-throughput analysis unveils a highly shared satellite DNA library among three species of fish genus Astyanax. *Scientific Reports*, *7*(1), 12726. https://doi.org/10.1038/s41598-017-12939-7

Smit, A., Hubley, R., & Green, P. (2013). *RepeatMasker Open-4.0*.

Suzuki, R., Terada, Y., & Shimodaira, H. (2019). *pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling* (R package version 2.2-0). https://cran.r-project.org/package=pvclust

Talbert, P. B., & Henikoff, S. (2022). The genetics and epigenetics of satellite centromeres. *Genome Research*, *32*(4), 608–615. https://doi.org/10.1101/gr.275351.121

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. *Frontiers in Genetics*, *11*(March). https://doi.org/10.3389/fgene.2020.00201

Thakur, J., Packiaraj, J., & Henikoff, S. (2021). Sequence, Chromatin and Evolution of Satellite DNA. *International Journal of Molecular Sciences*, *22*(9), 4309. https://doi.org/10.3390/ijms22094309

Theissinger, K., Edsman, L., Maguire, I., Diéguez-Uribeondo, J., & Jussila, J. (2022).

Nothing can go wrong–Introduction of alien crayfish to Europe. *PLOS Water*, *1*(11), e0000062. https://doi.org/10.1371/journal.pwat.0000062

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., … Zammit, G. (2023). How genomics can help biodiversity conservation. *Trends in Genetics*, *39*(7), 545–559. https://doi.org/10.1016/j.tig.2023.01.005

Ugarković, Đ. (2005). Functional elements residing within satellite DNAs. *EMBO Reports*, *6*(11), 1035–1039. https://doi.org/10.1038/sj.embor.7400558

Ugarković, Đ., & Plohl, M. (2002). Variation in satellite DNA profiles—causes and effects. *The EMBO Journal*, *21*(22), 5955–5959. https://doi.org/10.1093/emboj/cdf612

Ugarković, Đ., Sermek, A., Ljubić, S., & Feliciello, I. (2022). Satellite DNAs in Health and Disease. *Genes*, *13*(7), 1154. https://doi.org/10.3390/genes13071154

Utsunomia, R., Ruiz-Ruano, F. J., Silva, D. M. Z. A., Serrano, É. A., Rosa, I. F., Scudeler, P. E. S., Hashimoto, D. T., Oliveira, C., Camacho, J. P. M., & Foresti, F. (2017). A Glimpse into the Satellite DNA Library in Characidae Fish (Teleostei, Characiformes). *Frontiers in Genetics*, *8*. https://doi.org/10.3389/fgene.2017.00103

Utsunomia, R., Silva, D. M. Z. de A., Ruiz-Ruano, F. J., Goes, C. A. G., Melo, S., Ramos, L. P., Oliveira, C., Porto-Foresti, F., Foresti, F., & Hashimoto, D. T. (2019). Satellitome landscape analysis of Megaleporinus macrocephalus (Teleostei, Anostomidae) reveals intense accumulation of satellite sequences on the heteromorphic sex chromosome. *Scientific Reports*, *9*(1), 5856. https://doi.org/10.1038/s41598-019-42383-8

Veseljak, D., Despot-slade, E., Horvat, L., Vojvoda, T., & Mravinac, B. (2024). *Dynamic evolution of satellite DNAs drastically differentiates the genomes of Tribolium sibling species*. 1–49.

Vitales, D., Garcia, S., & Dodsworth, S. (2020). Reconstructing phylogenetic relationships based on repeat sequence similarities. *Molecular Phylogenetics and Evolution*, *147*, 106766. https://doi.org/10.1016/j.ympev.2020.106766

Wells, J. N., & Feschotte, C. (2020). A Field Guide to Eukaryotic Transposable Elements. *Annual Review of Genetics*, *54*, 539–561. https://doi.org/10.1146/annurev-genet-040620-022145

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. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1901), 1–10. https://doi.org/10.1098/rspb.2019.0079

Xu, Y., Tang, Y., Feng, W., Yang, Y., & Cui, Z. (2023). Comparative Analysis of Transposable Elements Reveals the Diversity of Transposable Elements in Decapoda and Their Effects on Genomic Evolution. *Marine Biotechnology*, *25*(6), 1136–1146. https://doi.org/10.1007/s10126-023-10265-w

Zhang, X., Yuan, J., Sun, Y., Li, S., Gao, Y., Yu, Y., Liu, C., Wang, Q., Lv, X., Zhang, X., Ma, K. Y., Wang, X., Lin, W., Wang, L., Zhu, X., Zhang, C., Zhang, J., Jin, S., Yu, K., … Xiang, J. (2019). Penaeid shrimp genome provides insights into benthic adaptation

and frequent molting. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-018-08197-4

Zhao, C., Zhang, X., Liu, C., Huan, P., Li, F., Xiang, J., & Huang, C. (2012). BAC end sequencing of Pacific white shrimp Litopenaeus vannamei: A glimpse into the genome of Penaeid shrimp. *Chinese Journal of Oceanology and Limnology*, *30*(3), 456–470. https://doi.org/10.1007/s00343-012-1159-y

Chapter IV

From DNA extraction to long read sequencing: workflow challenges of giant genomes of two non-model decapod species

Lena Bonassin, Christelle Rutz, Ljudevit Luka Boštjančić, Caterina Francesconi, Leonie Schardt, Carola Greve, Alexander Ben Hamadou, Damian Baranski, Charlotte Gerheim, Tassilo Erik Wollenweber, Barbara Feldmeyer, Sebastian Ploch, Arnaud Kress, Lucian Pârvulescu, Odile Lecompte, Kathrin Theissinger

To be submitted to BMC Genomics

Abstract

High-quality reference genomes are key for advancing evolutionary, ecological, and conservation research. Freshwater crayfish are ecologically important invertebrates, yet their genomic data remain limited. Currently, only four freshwater crayfish genome assemblies are publicly available, none of them belonging to the Astacidae family. Species of the Astacidae family are freshwater crayfish native to Europe, and are endangered by climate change, crayfish plague and competition from invasive non-native crayfish. The genome sequencing and assembly of crayfish species is challenging because of their large genome sizes, high number of repetitive sequences, and high heterozygosity. In this study we optimised protocols for long read sequencing of two freshwater crayfish species from the family Astacidae: the noble crayfish *Astacus astacus* and idle crayfish *Austropotamobius bihariensis*. We evaluated various DNA extraction methods, tissue types and library preparation strategies. We tested the suitability of six DNA extraction protocols (Qiagen MagAttract HMW DNA kit, Qiagen DNeasy blood and tissue kit, phenol-chloroform extraction, Nanobind Big DNA kit, salting out protocol, and sorbitol-wash coupled with salting-out protocol) and three tissue types (leg muscle, abdomen muscle, and claw tissue) to obtain high molecular weight (HMW) DNA. We assessed DNA quality and fragment length using fluorometric and spectrophotometric measurements, as well as pulsed-field capillary electrophoresis. The best performing protocol was chosen for the library preparation and sequencing. Specifically, the salting-out protocol combined with a sorbitol wash yielded the highest quantities (average 14 μg) of high-purity (A260/280 1.8 and A260/230 2.0-2.2), high-molecular-weight (HMW) DNA. We tested two sequencing platforms: PacBio and Nanopore using two library preparation approaches (with and without PCR amplification). PacBio outperformed Nanopore sequencing, especially when using the amplification-based library preparation approach. Our optimised workflow, encompassing the salting-out protocol with sorbitol wash and the use of abdomen or leg muscle tissue followed by PacBio amplification-based library preparation, provides a robust framework for generating high-quality HMW DNA and maximizing sequencing output from freshwater crayfish. With this approach the highest yield in a sequencing run was 30.3 Gb for *A. astacus* and 37 Gb for *A. bihariensis*. Overall, we successfully generated substantial genomic data, 640.26 Gb for *A. astacus* and 243.15 Gb for *A. bihariensis*, sufficient for *de novo* genome assemblies. This established workflow provides a valuable foundation for

accelerating genomic studies in freshwater crayfish and other challenging invertebrate species.

Keywords

HMW DNA extraction, long-read sequencing, freshwater crayfish, genomic DNA

4.1. Introduction

High quality reference genomes provide information on genetic and structural variants and elucidate complex genomic features such as satellite DNA (satDNA), mobile elements or segmental duplications (Theissinger et al., 2023). By thoroughly characterising these genomic features, the genomic resources can greatly assist biodiversity conservation and restoration efforts (Formenti et al., 2022; Theissinger et al., 2023). For instance, understanding genetic diversity and population structure through high-quality reference genomes can inform breeding programs, identify vulnerable populations, and guide reintroduction efforts for endangered species (Bonassin et al., 2024). In recent years, advances in long read sequencing technologies have allowed the genome assembly of many non-model species. Long read sequencing technologies simplify genome reconstruction and improve assembly contiguity, generating reads that can span tens of kilobases to several megabases (Espinosa et al., 2024). This offers the opportunity to investigate genomic structures and functions, phylogenetic relationships and evolutionary processes (Bein et al., 2025; Russo et al., 2022). Long-read sequencing is increasingly being applied to a wider range of species, leading to a rapidly growing number of high-quality reference genomes. However, there are still biological complexities that challenge genome sequencing and assembly processes. These include, but are not limited to large genome size, high heterozygosity levels, presence of mobile elements and/or tandem repeats and genomes with variable ploidy (Espinosa et al., 2024). A major step to overcome these sequencing challenges is to establish a reproducible high-quality laboratory workflow.

Due to their large proportion of repetitive elements, the production of reference genomes of Crustaceans is still particularly challenging. To date, there are only 45 published Crustacean genomes at chromosomal level, yet their contiguity remains low with short contig N50 (9 kb – 11 Mb) (NCBI Genome database, Yuan et al., 2023). Larger genomes require more sequencing data to obtain the necessary coverage and more computing resources to assemble a high-quality genome (Yuan et al., 2023). In Crustacea, genome sizes vary widely, with values ranging from 0.11 Gb in the ostracod *Xestolebris sp.* to 63.19 Gb in the amphipod

Ampelisca macrocephala (Gregory, 2025). Among species of the order Decapoda, genome size can reach up to 39.12 Gb in the polar shrimp *Sclerocrangon ferox* (Rees et al., 2008). In the freshwater crayfish family Astacidae there are notable large genomes, including the idle crayfish *Austroptamobius bihariensis* with 11.58 Gb (Chapter III), the noble crayfish *Astacus astacus* with 16.89 Gb (Chapter III), and the narrow-clawed crayfish *Pontastacus leptodactylus* with 18.7 Gb (Boštjančić et al., 2021). These genomes are also high in repetitive DNA content (Rutz et al., 2023) which poses difficulties in aligning and assembling reads due to the copies in multiple locations in the genome and the length of their repeat unit (Treangen & Salzberg, 2012). It has been shown that the genome size in Decapoda correlates with a higher number of transposable element (TE) copies, whose proportion in the genome can be up to 82% (Rutz et al., 2023, Chapter III). Along with a high proportion of TEs, satDNA occupies a large part of Decapoda genome. In the family Astacidae, satDNA represents the highest proportion of repetitive elements, with up to 55% in the *Pontastacus leptodactlyus* genome (Chapter III; Boštjančić et al., 2021). High heterozygosity also has a significant influence on sequencing and genome assembly resulting in assembly fragmentation, incorrect haplotype collapsing and increased error rates (T. Zhang et al., 2022). High heterozygosity was reported in several freshwater crayfish species (Gross et al., 2021; Lovrenčić et al., 2022) with a significant amount of genetic variation within populations (Bonassin et al., 2024; Wei et al., 2024). Therefore, genome assembly projects of freshwater crayfish are particularly challenging due to their complex genomic characteristics.

In addition to the degree of genomic complexity, the success of long read genome sequencing also depends on obtaining high quality and high molecular weight (HMW) DNA, ideally with a fragment length of 50 kb or more. Obtaining HMW DNA from crustaceans has several challenges, from DNA *in vivo* stability, to extraction, purification and library preparation: DNA is easily degraded by endonucleases (Abdelrahman et al., 2017; Angthong et al., 2020) and the purity is affected by high amounts of polysaccharides and polyphenolic proteins (Panova et al., 2016). Initial sample handling has significant impact on DNA integrity, therefore proper tissue preservation is needed to prevent enzymatic activity and oxidative degradation (Nagy, 2010). Flash-freezing of tissue in liquid nitrogen immediately after collection has been the benchmark for minimal DNA degradation and impact on fragment length (Dahn et al., 2022). Unlike short-read sequencing, which is compatible with DNA obtained from almost all standard DNA extraction methods, only a few extraction methods are suitable for HMW DNA (Dahn et al., 2022). These include bead-based, high-salt, agarose

plug and magnetic disk methods (Dahn et al., 2022). While the use of the appropriate extraction methods ensures the appropriate DNA integrity, DNA purity is an additional crucial parameter for long-read sequencing (Trigodet et al., 2022). Conventional protocols can be unreliable when applied to certain organisms with high contaminant levels still present in DNA extracts, which can interfere with downstream applications. To increase DNA purity, common DNA purification strategies include bead-based clean up, agarose gel electrophoresis, or buffer exchange. However, it is crucial that purification protocols preserve the integrity and length of the DNA as excessive shearing can compromise sequencing outcomes.

Pacific Biosciences' (PacBio) Single Molecule Real-Time (SMRT) sequencing is a long-read sequencing technology that generates highly accurate and long DNA reads (Rhoads & Au, 2015). PacBio Circular Consensus sequencing (CCS) strategy derives a consensus sequence from multiple passes of a single template molecule producing accurate reads. CCS reads with over 99% accuracy are called High Fidelity (HiFi) long reads (Schell et al., 2025). The choice of library preparation workflow, either amplification-free or whole-genome amplification-based, depends on the available DNA quantity. The standard amplification-free workflow is preferred due to sequencing native DNA but requires high amounts of DNA (minimum 400 ng recommended, PacBio, 2020). In contrast, the amplification-based workflow enables sequencing from samples of extremely limited DNA quantity (minimum 5 ng recommended, PacBio, 2020). However, whole genome amplification in this workflow can introduce risks such as errors or coverage biases when compared to the standard, amplification-free method (PacBio, 2020). In cases where no other option is available, these risks are acceptable if a high-quality reference genome is sought after. The second most widely used long-read sequencing technology is Oxford Nanopore Technologies (ONT) that uses change in ionic current as a DNA molecule passes through a nanopore to determine the sequence of nucleotides (Wang et al., 2021). The main strategies of library preparation for ONT are ligation-based protocols which involve enzymatic adapter ligation to native DNA fragments, transposase-based methods that utilise a transposase to fragment and attach adapters and PCR amplification-based workflows usually used for low input samples. Both PacBio and ONT offer considerable advantages compared to short read sequencing in generating the long reads critical for *de novo* assembly of complex, giant genomes.

In this study, we aim to identify the most effective combination of DNA extraction methods and purification strategies for long read sequencing of two freshwater crayfish species with

giant genomes, *A. astacus* and *A. bihariensis*. These species are characterised by large genomes which are impossible to assemble using short reads and which present significant challenges for long-read sequencing. However, due to their crucial ecological importance as keystone species in freshwater ecosystems and their endangered status, their reference genomes are more needed than ever. We hypothesise that a carefully optimised protocol, involving specific DNA extraction and purification steps, will yield DNA of sufficient quality to overcome the known difficulties in sequencing crayfish species with large and repetitive genomes. We assess DNA quantity and quality using fluorometric measurements and high sensitivity fragment length analysis. To test the suitability of the protocol for long-read sequencing technology, we prepared amplification-free as well as amplification-based libraries. Based on our findings, we propose a DNA extraction and library preparation protocol suitable for long read sequencing of freshwater crayfish. The insights gained in this study may also be applicable to other non-model invertebrate species.

4.2. Methods

4.2.1. Sampling of individuals and tissue

One adult male individual of *A. astacus* was obtained from the breeder Flusskrebszucht Frömel (Kavelstorf, Germany). One adult male individual of *A. bihariensis* was collected from the Valea Iadului river in Romania (46,7447 N 22,5597 E) with the necessary authorisation from the Romanian Academy (1/CJ/13.01.2021), the Romanian Ministry of Water and Forests (DGB/2/R5787/16.08.2022), the Apuseni Nature Park Administration (199/09.09.2022), the National Agency for Protected Areas (882/15.09.2022), and the Environmental Protection Agencies in the geographical area where the specimen was sampled (76/20.09.2022).

For both individuals, approximate 50 mg muscle tissue samples from abdomen, legs (pereopods) and claws (chelae) were dissected and flash frozen with liquid nitrogen before being stored at -80 °C until DNA extraction.

4.2.2. Genomic DNA extraction

Samples from both *A. astacus* and *A. bihariensis* were used in the extraction testing, and each protocol was tested at least in two replicates. For all DNA extractions, tissue disruption was performed in lysis buffer (described for each approach below) with a plastic pestle on ice. In total we tested five extraction protocols representative of the main DNA extraction principles used for obtaining HMW DNA, including three commercial protocols (Qiagen DNeasy Blood

and Tissue Kit, Qiagen MagAttract HMW DNA kit, Nanobind Tissue Big DNA Kit) and two non-commercial protocols (phenol-chloroform extraction and salting out protocol). The detailed procedure of each protocol is described in the text below.

4.2.2.1. Qiagen DNeasy Blood and Tissue Kit

The DNeasy Blood and Tissue Kit (Qiagen, Germany) was utilised according to the manufacturer's protocol for purification of total DNA from animal tissue, with the addition of an optional step: after lysis, RNase A was added to the lysate according to the protocol, and the solution was incubated for 2 min at room temperature. Finally, DNA was eluted in 100 µL of AE buffer preheated to 37 °C.

4.2.2.2. Qiagen MagAttract HMW DNA kit

The MagAttract HMW DNA kit (Qiagen, Germany) was used for genomic DNA extraction according to the manufacturer's protocol for purification of DNA from fresh or frozen tissue with the following modifications: to improve the lysis of the tissue, lysis was performed for 4 h, and all centrifugation steps were performed at 1200 rpm to reduce DNA shearing. Finally, DNA was eluted in 60 µL of AE buffer preheated to 37°C.

4.2.2.3. Nanobind Tissue Big DNA Kit

Genomic DNA extraction was performed using the Nanobind Tissue Big DNA Kit (Circulomics) following the manufacturer's protocol for isolation of HMW DNA from crab muscle (Circulomics Crab muscle Application Note v1 12/2019).

4.2.2.4. Salting out protocol

A modified salting out protocol (Jenkins et al., 2019) was implemented for DNA. Specific modifications included: tissue digestion performed for 4 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.

4.2.2.5. Phenol-chloroform extraction

Genomic DNA was extracted following the phenol-chloroform protocol by Sambrook & Russell, 2006. Tissue was homogenised in H1 buffer (50 mM Tris-Cl, 10 mM EDTA, 100 mM NaCl, 1% SDS and 1 mg/mL proteinase K) and lysed for 4 h at 65°C and 400 rpm on a Thermomixer. After lysis, 0.5 mg/mL RNase A was added, and samples incubated for 20 min at 37°C. After 300 µL of phenol-chloroform-isoamylalcohol (25:24:1) was added to the lysate, the mixture was centrifuged for 5 min at 5000 x g. 1x volume chloroform was added

to the aqueous phase and the mixture was centrifuged for 5 min at 5000 x g. DNA was precipitated by centrifugation, by adding 0,1x volume 3M sodium acetate and 2,5x volume 100% ethanol. The solution was incubated for 1 h at -20 °C and subsequently centrifuged for 10 min at 7500 x g. The pellet was washed twice, first with 1 mL 100% ethanol and again with 1 mL 80% ethanol, after an intermediate centrifugation for 10 min at 7500 x g. After the last centrifugation for 10 min at 7500 x g, the pellet was air dried and resuspended in 70 µL TE-buffer.

4.2.3. Evaluation of DNA quantity and quality

For all DNA extracts, DNA was quantified using a QuantiFluor® dsDNA System on the Quantus™ Fluorometer (Promega, USA). The DNA purity was estimated using a Nanophotometer P300 (Implen, Germany). The fragment size distribution was assessed using a Femto Pulse System and the Genomic DNA 165 kb Kit (Agilent, USA).

4.2.4. Pre-extraction sorbitol washing complex homogenate protocol

After analysing the results for DNA yield and fragment size, the two most suitable extraction methods MagAttract HMW DNA kit (section 4.2.2.2) and salting out protocol (section 4.2.2.4), evaluated based on the overall yield, fragment length and purity, were complemented with a sorbitol based purification protocol prior to DNA extraction to obtain higher purity of DNA. The sorbitol washing complex homogenate protocol (Jones et al., 2021) was modified as follows: tissue was grinded using a pestle in a 2 mL tube filled with the freshly prepared sorbitol wash solution. To the solution, 1% β-mercaptoethanol (v/v) was added and centrifuged for 5 min at 2500 x g. After removal of the wash solution, extractions were continued as described above.

4.2.5. Comparison of size selection protocols

4.2.5.1. AMPure PB beads size selection

To improve the fragment length of DNA used in library preparation, we tested the size selection using AMPure PB beads (40% v/v) following the protocol “Size selection with AMPure PB Beads” (101-730-400 Version 07, November 2021). DNA was incubated with beads at room temperature for 30 min. After removing the supernatant, the beads were washed with 80% ethanol twice and the DNA was eluted in 15 µL Elution buffer.

4.2.5.2. BluePippin size selection

We tested the size selection using the BluePippin (SageScience, USA). We followed the High-Pass^{TM} DNA Size Selection protocol with BLF7510 Cassette kit and 0.75% DF Marker

S1 high-pass 6-10kb vs3 Cassette Definition. Samples were eluted in 40 μL of BluePippin run buffer.

4.2.6. Oxford Nanopore Technology library preparation

For Nanopore libraries, we used DNA extracts with amount higher than 1 μg A260/280 purity ratio above 1.8 and A260/230 between 2.0 and 2.2. Nanopore libraries were prepared using the amplification-free Ligation gDNA sequencing kit SQK-LSK109 and sequenced on a MinION Flow Cell R9.4.1 FLO-MIN106D. To reduce costs, a set of libraries were prepared using the amplification-based Rapid barcoding gDNA sequencing kit SQK-RBK004 and sequenced on a Flongle Flow Cell R9.4.1 FLO-FLG001.

4.2.7. PacBio library preparation

The extraction and purification combinations that yielded the best results based on DNA quantity and quality and according to PacBio recommendations, were selected for library preparation and sequencing. We used DNA extracts with amount higher than 500 ng, A260/280 purity ratio above 1.8 and A260/230 between 2.0 and 2.2 and fragment length longer than 10 kb.

4.2.7.1. Low input protocol

Low input PacBio HiFi libraries for both species were 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). DNA was sheared using a long hydropore on a Megaruptor 2 (Diagenode, USA) set to 25 kb target length. After the final adapter ligation, nuclease treatment of the libraries was performed according to the protocol. The libraries were sequenced on Sequel IIe platforms at Novogene (UK) or Radboudumc Genome Technology Center (Netherlands) (Supplementary table IV-1).

4.2.7.2. Ultra-low input protocol

Library preparation for *Astacus astacus*

Ultra-low input PacBio HiFi libraries for the species *A. astacus* were prepared and sequenced at the West German Genome Centre (WGGC, Düsseldorf, Germany). Libraries were prepared with the ultra-low input WGS workflow using the Express Template Prep Kit 2.0 (PacBio) according to the 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 a Femto Pulse System (Agilent, USA) with the Genomic DNA 165 kb kit. 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 included 0.75 µL Takara LA Taq, 37.5 µL 2x GC buffer I, 12 µL dNTP mixture and 2µl PacBio amplification PCR primers. For reaction mix A, a 10 min extension time was used. Both PCR programs were initially run with 15 cycles. An additional 2-3 cycles were performed if the amount of PCR product was less than 250 ng per reaction. Concentration of the amplified libraries was measured with a Qubit fluorometer (Thermo Fischer Scientific, USA) and size distribution was analysed on a Fragment Analyzer (Agilent) with the DNF-464 large fragment kit. 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. Final libraries were size selected on a Blue Pippin (Sage Science, USA) with the following settings: 0,75% DF Marker S1 High-Pass 6-10kb vs3. Cut-off 6kb. Library concentration was measured with Qubit dsDNA HS (Thermo Fischer Scientific, USA) and size distribution was analysed 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/Sequel IIe instrument using 20 8M Sequel II 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.

Library preparation for *Austropotamobius bihariensis*

Ultra-low input PacBio HiFi libraries for the species *A. bihariensis* were prepared according to the protocol described in (Bein et al., 2025). The libraries were prepared with the SMRTbell Express Template Prep Kit 2.0 according to the "Procedure & Checklist— Preparing HiFi SMRTbell® Libraries from Ultra-Low DNA Input" (PN 101–987-800 Version 02). Whole genome amplification was performed using the Polymerase C (KOD Xtreme™ Hot Start DNA Polymerase, Merck PN 71975). Libraries were sequenced on the Revio platform by Bioscientia Healthcare GmbH (Ingelheim, Germany).

4.2.8. Statistical analyses

Statistical analyses were conducted in R version 4.2.1 (R Core Team, 2021). For all analyses, the significance level α=0.05 was used. The assumption of normality was assessed using the Shapiro-Wilk test, and homogeneity of variances was evaluated with Levene's test. Statistical analyses were done on extraction methods with more than 5 replicates. Differences in DNA yield across extraction methods and tissue types were analysed using the Aligned Rank Transform (ART) ANOVA. Significant main effects were investigated using Tukey post-hoc pairwise comparisons to determine specific group differences. Differences in A260/280 and A260/230 purity ratio across extraction methods and tissue types were analysed using ANOVA. Significant main effects were investigated using Tukey post-hoc pairwise comparisons to determine specific group differences. Wilcoxon signed-rank test was used to assess differences in PacBio HiFi read length between library preparation methods, while t-test was used for testing differences in PacBio HiFi yield. Kruskal Wallis test was used to assess differences in PacBio HiFi read length and PacBio HiFi yield between sample replicates. PacBio sequencing runs were evaluated comparing productivity values P0, P1 and P2, which represent a percentage of the total number of ZMWs (Zero Mode Waveguide) on a SMRT cell. Spearman's correlation was used to calculate the correlation between P1 productivity values and HiFi read length, and between P1 productivity values and HiFi yield. Differences between P0 and P1 against different library preparation approaches were assessed using a t-test.

4.3. Results

In this study we employed multiple tissues, extraction methods and library preparation strategies to identify the protocol that produces highest amounts of DNA suitable for sequencing. Figure IV-1. presents a comprehensive visualisation of the sample processing workflow. After DNA extraction and library preparation, quality controls were performed and samples that failed to meet DNA quality and quantity thresholds were not advanced to subsequent processing steps. Details about samples, extraction protocol used, quantity and quality of DNA are provided in Supplementary table IV-1.

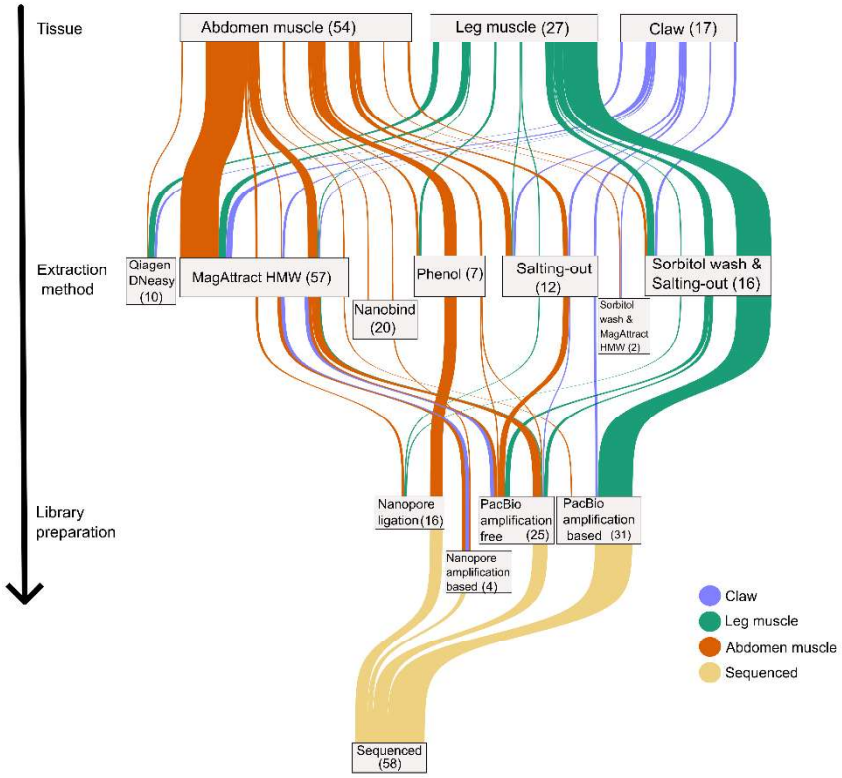


Figure IV-1. Workflow of sample processing from tissue to sequencing. The Sankey diagram illustrates the flow of samples through various stages of processing for different tissue types. Ribbons represent the sample pathways, originating from three tissue types (Claw, Leg muscle, Abdomen muscle). Samples then proceed through different DNA extraction methods (Sorbitol wash & Salting-out, Salting-out, MagAttract HMW, Phenol-chloroform, Nanobind Big DNA kit, Qiagen DNeasy). Following extraction, samples proceed to library preparation methods (PacBio amplification-based kit, PacBio amplification free kit, Nanopore transposase-based kit and Nanopore ligation kit), leading to sequencing. The width of the ribbons is proportional to the number of samples following that specific path. The numbers in parenthesis indicate the number of samples.

4.3.1. Total DNA concentration, purity and fragment length

DNA yield, purity and fragment length were assessed across different extraction methods and tissue types. DNA yield was significantly different among the various DNA extraction methods (ART ANOVA, F = 12.012, p = 3.4811 e-05, Supplementary table IV-2). Specifically, the salting-out protocol and sorbitol wash coupled with salting-out protocol produced the highest yields, with an average of 13 µg of DNA (Supplementary table IV-3, Figure IV-2). The lowest DNA yield was produced by sorbitol wash coupled with Qiagen MagAttract HMW kit, with an average of 311 ng (Supplementary figure IV-1). Tissue type also exerted a significant effect on DNA yield (ART ANOVA, F = 12.887, p = 1,8397 e-05,

Supplementary table IV-2), with abdomen and leg muscle yielding more DNA than claw tissue (Supplementary table IV-3).

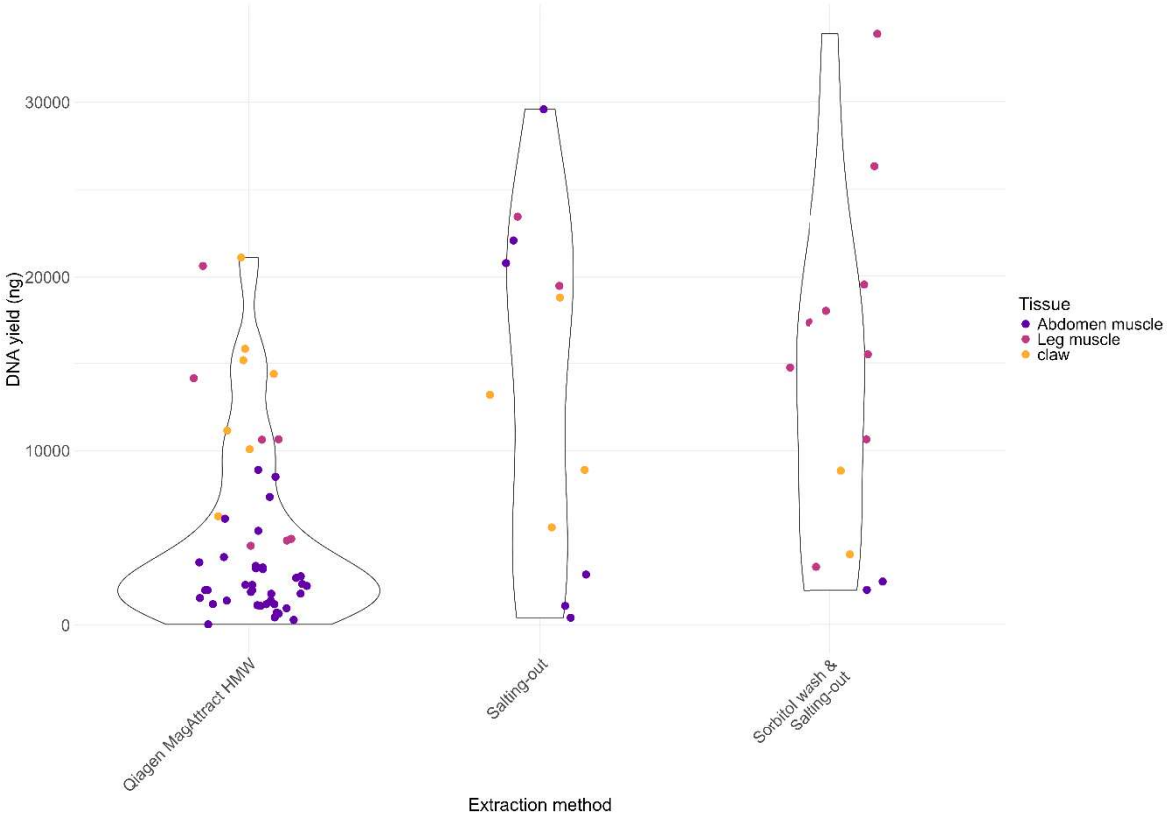


Figure IV-2. Violin dot plot of total DNA yield (ng) from three extraction methods. Colour indicates tissue types used for DNA extraction.

DNA purity, as indicated by the A260/280 ratio, varied significantly across extraction methods (ANOVA, F = 4.031, p = 0.0235, Supplementary table IV-4). Sorbitol washing coupled with the salting-out protocol resulted in A260/280 ratios in the 1.8 to 2.0 range, while Qiagen MagAttract HMW kit extracts showed values lower than 1.8 (Figure IV-3, Supplementary table IV-5). Tissue type did not have an influence on A260/280 values (Supplementary table IV-4). Moreover, there was no difference in A260/230 values between extraction methods and tissue types (Supplementary table IV-6). The lowest purity values were obtained with the Nanobind Big DNA kit and sorbitol wash coupled with the Qiagen MagAttract HMW kit (Supplementary figure IV-2.)

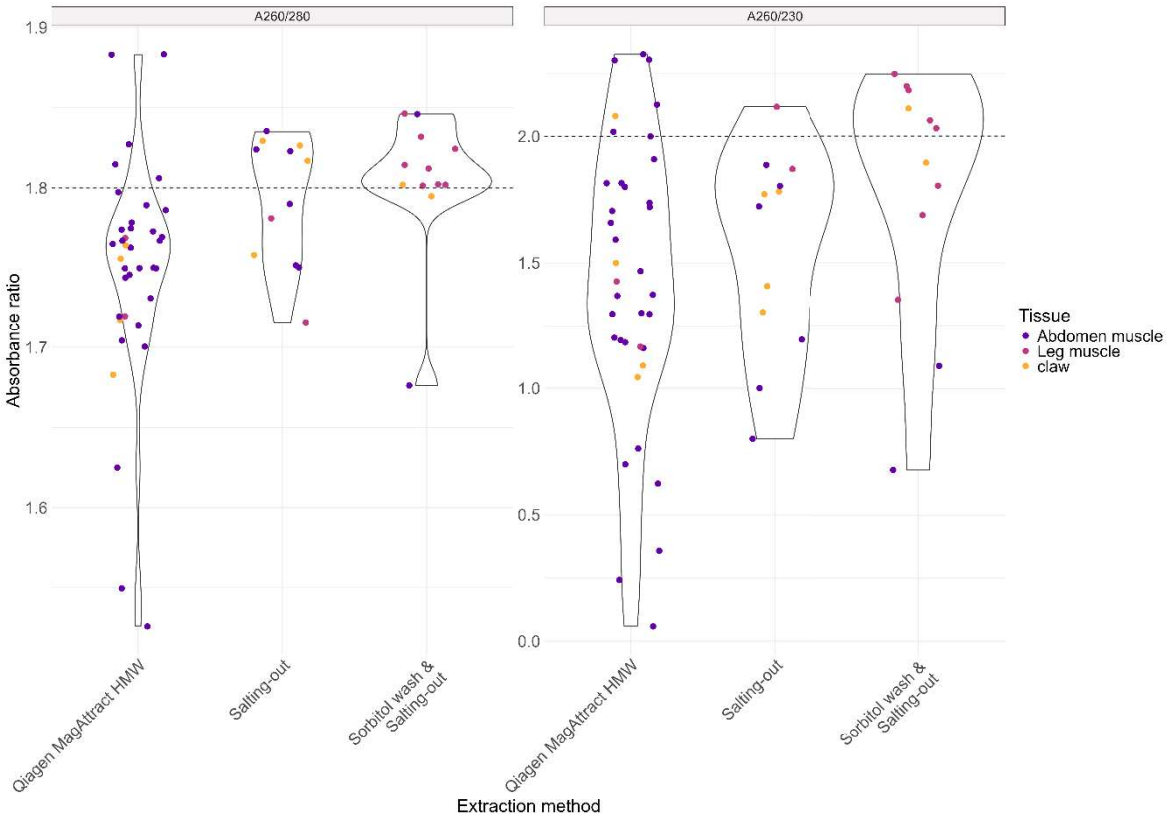


Figure IV-3. Absorbance ratio A260/280 (left) and A260/230 (right) for three extraction methods. Colour indicates tissue types used for DNA extraction. Dashed lines indicate optimal absorbance ratio values.

Fragment size distribution revealed that both extraction method and tissue type had significant impact on the fragment length (Supplementary table IV-7). Extraction methods had a significant impact (ART ANOVA, F = 6.1652, p= 0.00373), with the sorbitol wash coupled with salting out protocol yielding DNA with fragments longer than Qiagen MagAttract HMW and salting-out alone (Supplementary table IV-8, Figure IV-4). Tissue type also exerted a significant influence on DNA fragment length (ART ANOVA, F = 3.3711, p = 0.0106), however, post-hoc analyses did not reveal specific differences among tissue types. Among all extraction methods the shortest DNA fragments were obtained with phenol-chloroform and Qiagen DNeasy kit (Supplementary figure IV-3).

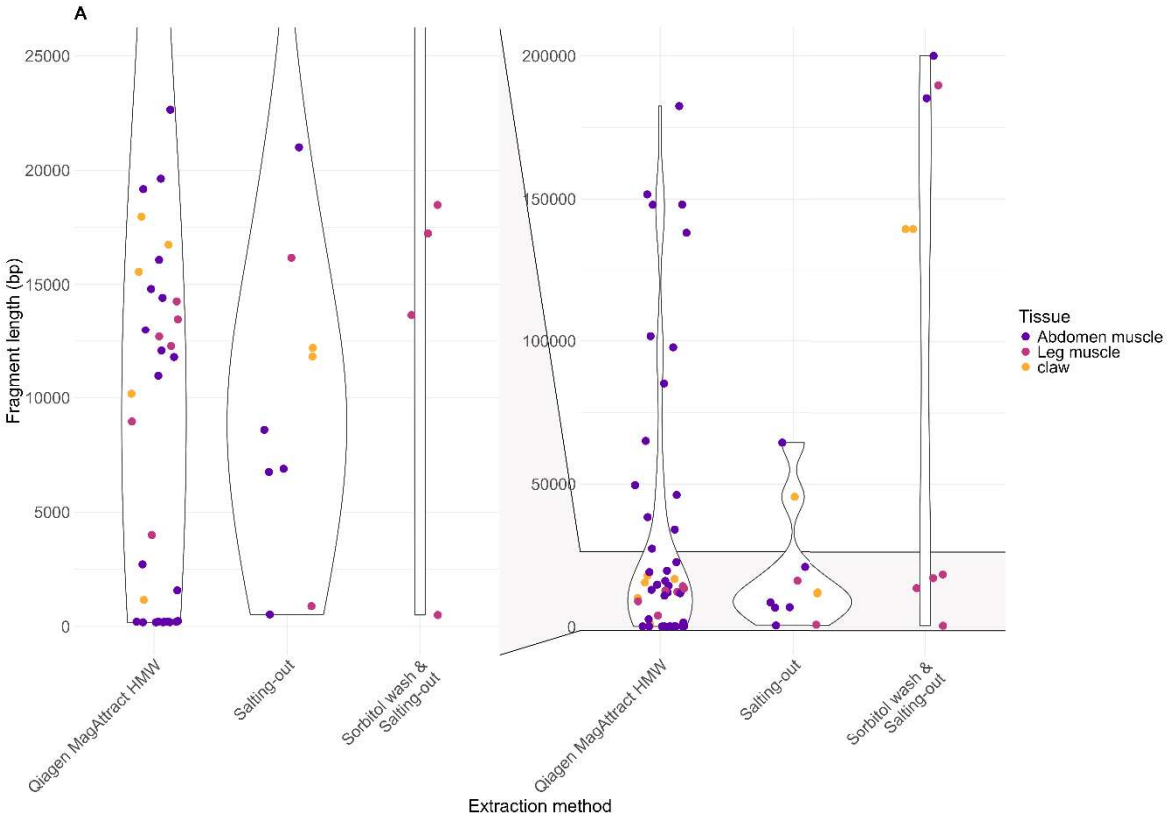


Figure IV-4. DNA fragment length (bp) at which the highest concentration of DNA extract was observed for three extraction method. A) Zoomed in y-axis 0-25000 bp and B) y axis 0-200000 bp. Colour indicates tissue types used for DNA extraction.

4.3.2. Size selection

To improve the fragment length of the libraries we evaluated the AMPure PB bead (40 % v/v) and BluePippin size selection methods using an exemplary DNA extract from *A. astacus* leg muscle obtained with sorbitol wash and salting out extraction method. The original DNA sample (Figure IV-5) exhibited a broad fragment distribution, with a highest concentration peak around 11 kb. Following AMPure PB bead size selection (Figure IV-5), the fragment distribution showed a peak around 12 kb, however it was still broad. In contrast, the BluePippin size selection (Figure IV-5) effectively removed the low molecular weight DNA fragments, with the apex of the biggest peak at 16 kb.

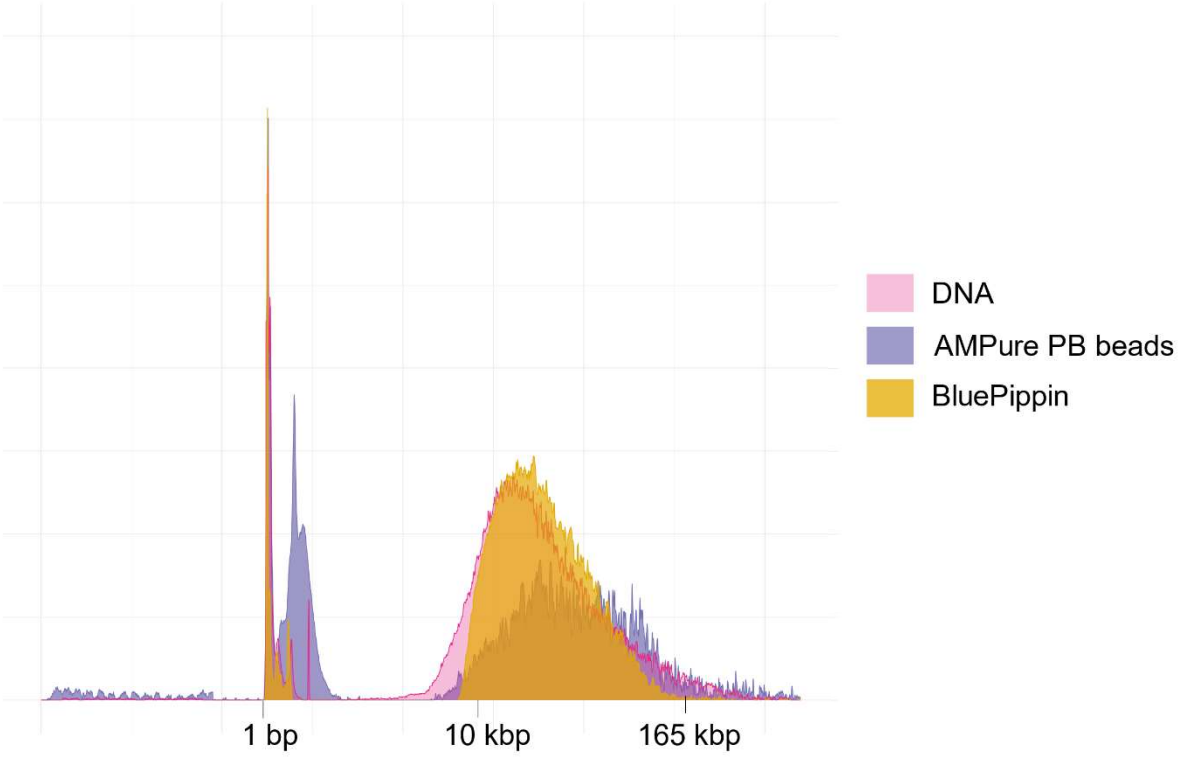


Figure IV-5. DNA fragment distribution electropherograms from a Femto pulse run using the genomic DNA 165 kb kit (lower marker at 1 bp, ladder range from 1 bp to 165 kb). Colours indicate original DNA sample (pink), AMPure PB bead (purple) and BluePippin (yellow) size selected sample.

4.3.3. DNA sequencing

Based on comprehensive quality metrics, the DNA extraction that yielded optimal DNA quality (high yield, purity, and large fragment size) were selected for PacBio HiFi library preparation and subsequent sequencing. Some DNA extracts were used for preparing multiple HiFi libraries. As shown in Figure IV-1, 58 out of the 76 libraries were successfully sequenced. The remaining libraries could not be sequenced primarily due to insufficient DNA amounts or wide fragment distribution, which failed to meet Nanopore and PacBio's sequencing requirements.

To evaluate the Nanopore sequencing we assessed the read N50 and the total sequencing yield across two library preparation approaches: ligation based and amplification-based approach. The overall read N50 ranged from 2525 bp to 363690 bp, and the yield from to 4.3 Mbp to 1800 Mbp. The phenol-chloroform DNA extraction showed longer read lengths and higher sequencing yields than the Qiagen MagAttract HMW kit (Supplementary figure IV-4). Overall, the ligation-based kit produced longer fragments and higher sequencing yield than

the transposase-based kit (Supplementary figure IV-4). Considering the generally lower yields observed with Nanopore sequencing in this study, most of the libraries for subsequent sequencing were prepared and run using PacBio technology.

PacBio sequencing success was evaluated with HiFi read length and HiFi yield values in relation to the two library preparation approaches: the low input and the ultra-low input (Figure IV-6). The HiFi read length differed between the two library preparation methods, with ultra-low input producing longer reads (Wilcoxon Rank- Sum test, W = 53, p = 0.0408, Supplementary table IV-9). Similarly, HiFi yield was significantly higher for the ultra-low input approach (t-test, t = -18.13714, p = 6.66e-21). Furthermore, analyses of within-sample replicates consistently showed no significant differences in neither HiFi yield nor HiFi read length (Kruskal-Wallis H-tests: p > 0.05 for all comparisons, Supplementary table IV-10).

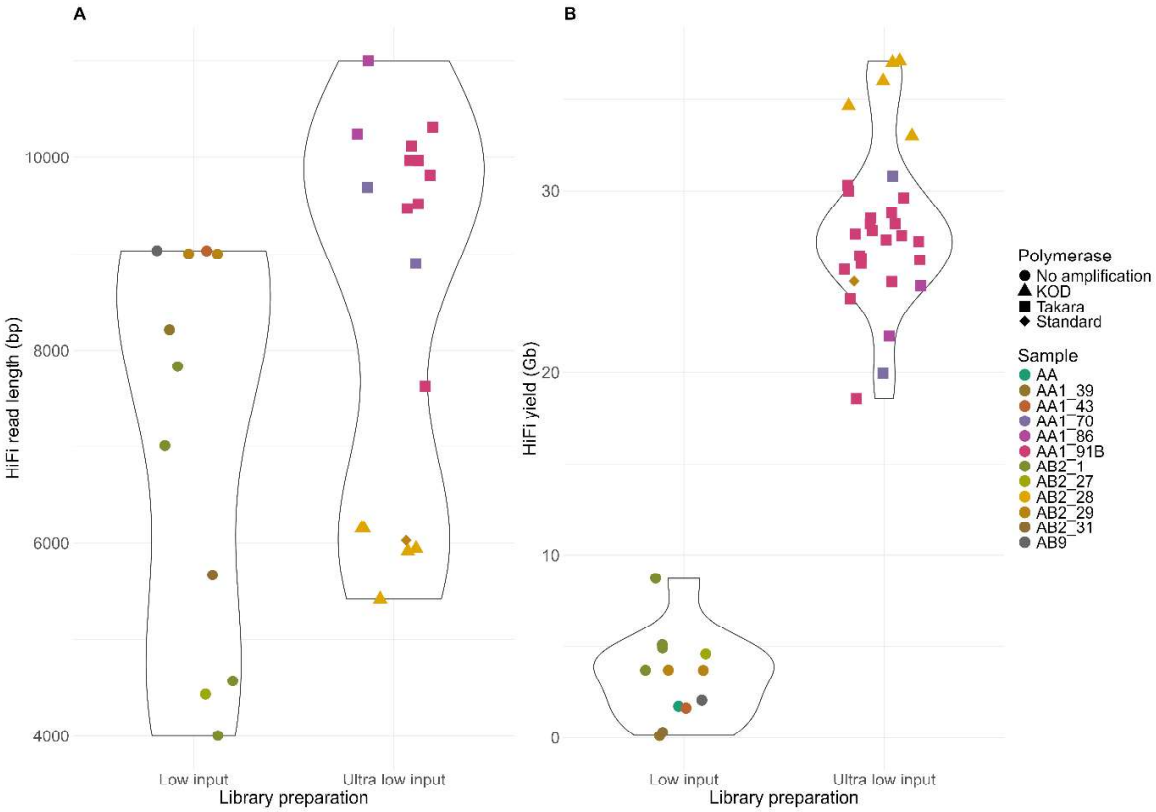


Figure IV-6. HiFi read length (A) and HiFi yield (B) separated by library preparation method. Colours indicate samples used for library preparation, while shape indicates the polymerase used in the ultra-low input library preparation.

Productivity values P0, P1 and P2 were evaluated for PacBio sequencing success. These values are co-dependant and together equal P0 + P1 + P2 = 1. For all samples, the P2 value was low (< 6.26, mean = 2.11), while an increase in P0 would be reflected as a decrease in

P1. There was a significant difference in P0 and P1 values between the two library preparation approaches (Supplementary table IV-11). Amplification based libraries exhibited P1 values ranging from 46.54 to 85.66 (mean = 70,42), which were overall higher than those observed in amplification-free libraries, ranging from 7.02 to 58.94 (mean = 31.90) (Figure IV-7A, C). Spearman's correlation revealed a significant positive relationship between P1 productivity values and HiFi yield for all samples prepared with ultra-low and low input protocols (R = 0.76, p = 3.4 e-06, Figure IV-7B). A positive relationship was also identified between P1 and HiFi Yield for amplification-free libraries (R = 0.67, p = 0.024), but correlation was not observed for amplification-based libraries (Figure IV-7B). No correlation was found between P1 values and HiFi read length for all samples (Figure IV-7D).

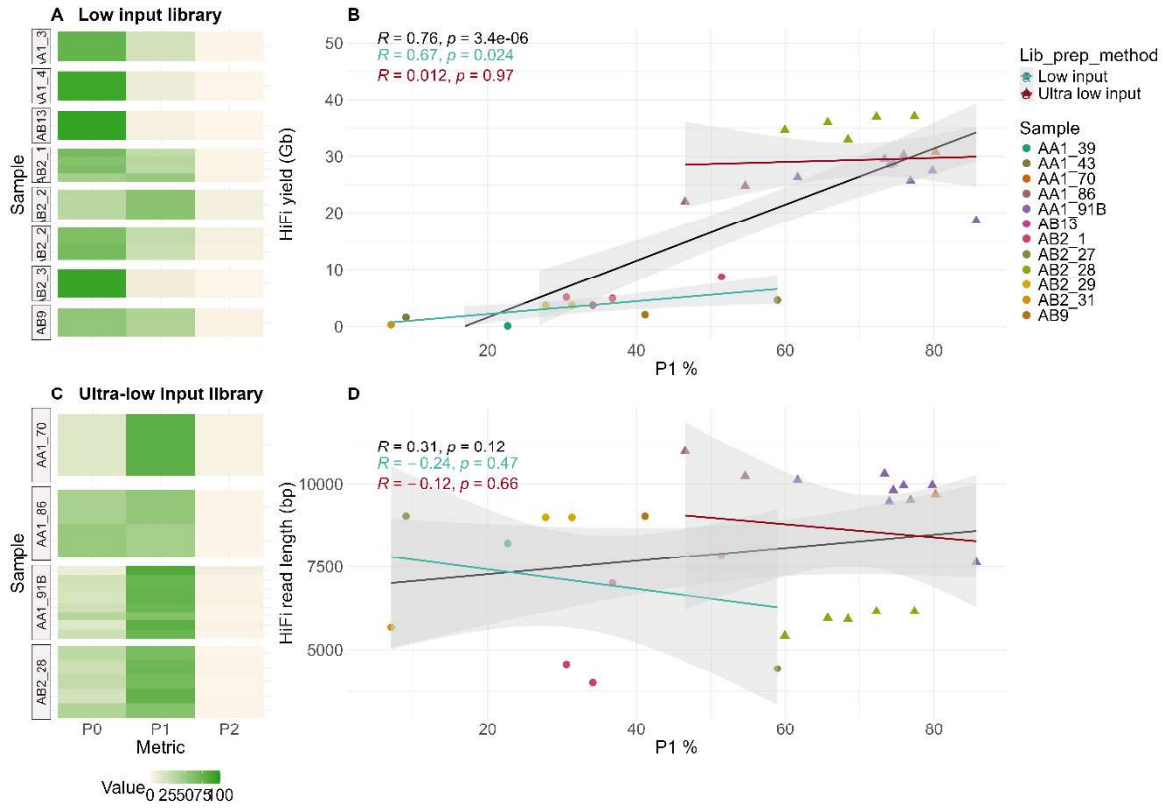


Figure IV-7. Sequencing metrics P0, P1 and P2 and correlation of P1 % with HiFi yield (Gb) and HiFi read length (bp) per sample for Low input and Ultra low input libraries. Panels A and C display heatmaps of sequencing metrics P0, P1, and P2 (%) for low input and ultra-low input libraries, respectively. Panels B and D show correlation of P1 % with HiFi sequencing yield (in Gb) and HiFi read length (bp), respectively.

4.4. Discussion

The objective of this study was to evaluate various DNA extraction methods, tissue types and library preparation strategies to identify the optimal workflow for generating high quality HMW DNA for long read sequencing of freshwater crayfish species. The results of our study

demonstrate that both the choice of extraction protocol and tissue used for library preparation plays important roles in long read sequencing success. Moreover, we identified that the PacBio HiFi sequencing and its amplification-based library preparation approach is better suited for obtaining high yields of sequencing data.

The complexities of *de novo* genome assembly are particularly pronounced in freshwater crayfish, characterised by large, repetitive, genomes with a large number of chromosomes. Successful assemblies necessitate substantial quantities of high-molecular-weight (HMW) DNA (Dahn et al., 2022). Adding to these challenges, high levels of heterozygosity in crayfish can significantly complicate genome assembly and introduce errors (Xu et al., 2021; You et al., 2020). Therefore, it is important that all DNA used for sequencing originates from a single individual. Maximising the DNA yield from a single specimen is a crucial factor making efficient extraction protocols paramount for achieving the necessary input for long read sequencing technologies. To date, only four freshwater crayfish genomes have been assembled. These previous efforts used hybrid strategies combining Illumina short reads for high accuracy with PacBio and/or Nanopore long reads and Hi-C data for scaffolding (Austin et al., 2022; Gutekunst et al., 2018; Liao et al., 2024; Tan et al., 2020; Xu et al., 2021). Across these studies, DNA extraction relied on commercial kits (Qiagen DNeasy, E.Z.N.A Tissue DNA kit, Zymo Quick gDNA) from different tissue types (hepatopancreas, ovaries, abdomen muscle). Given the large crayfish genomes (3.5 Gb – 4.6 Gb) and high repeat content (27.5% - 79.6%) these studies produced large amounts of sequencing data requiring numerous sequencing runs, e.g., 24 PacBio libraries for the *Procambarus clarkii* (Liao et al., 2024) and 20 R9.4.1 Nanopore flowcells for the *Cherax destructor* (Austin et al., 2022). These studies collectively highlight that achieving high-quality, contiguous genome assemblies in freshwater crayfish necessitates efficient DNA extraction and, advanced sequencing technologies strategies to overcome the inherent genomic complexities of this group. However, freshwater crayfish genome studies do not report DNA extraction quantity and quality metrics. This lack of reported data makes it difficult to compare results between studies and establish a standard for high-quality DNA extraction in this group of organisms. It highlights a gap in the existing literature and underscore the need for reporting methodological details, including DNA quality metrics, to ensure reproducibility and facilitate future genomic research in freshwater crayfish.

In our study, DNA yield proved to be highly dependent on both the extraction and tissue type (Figure IV-2, Supplementary table IV-2). The salting out protocol, particularly when coupled

with the sorbitol wash purification, consistently produced highest DNA yields (Supplementary table IV-3). Other DNA extraction kits, based on column extraction (Qiagen DNeasy) and magnetic beads (Qiagen MagAttract HMW) produced lower yields (Supplementary figure IV-1), possibly because of limited binding capacity to the silica column or beads, the column being clogged with tissue fibres or excessive DNA loss during washing steps (Qiagen, 2023). Tissue type also played a significant role, with abdomen and leg muscle tissues consistently yielding more DNA than claw tissue (Supplementary table IV-3). This is likely attributed to cellular density and toughness, as claw muscle contains long fibres and can be integrated with connective tissue (Fukuzawa, 2001), making it lowly abundant in cells and tougher to homogenise. A tissue type commonly used for DNA extraction in crustaceans is the hepatopancreatic tissue, which is, however rich in nuclease enzymes which can affect DNA integrity, contains high levels of lipids and other metabolites which can inhibit downstream reactions, and can be a source of microbial contamination.

All enzymatic reaction involved in library preparation and sequencing can be affected by DNA purity and leftover contaminants. In common laboratory practices, DNA samples with A260/280 ratio above 1.8 and A260/230 ratio 2.0 – 2.2 are considered suitable for downstream applications (Glasel, 1995). Chemicals like ethanol or phenol used in DNA extraction protocols, proteins and/or other organic compounds from the animal tissue can compromise DNA purity (Russo et al., 2022). Low A260/280 values may indicate protein and phenol presence in the DNA extract, while low A260/230 values indicate to a presence of carbohydrates, buffer salts from DNA extraction, ethanol or EDTA (Russo et al., 2022). Common contaminants reducing A260/230 values are polysaccharides and polyphenols, often found in crustacean tissue (Panova et al., 2016). Our findings demonstrated that the sorbitol wash coupled with the salting-out protocol delivered DNA with A260/280 ratios within the ideal range, indicative of high purity (Figure IV-3, Supplementary table IV-5). Salting-out protocols are efficient in removing proteins, as high salt concentration reduce the solubility of the protein molecules, precipitating them from the solution. A260/230 ratios did not show significant variations across methods or tissue types (Supplementary table IV-6), however, only the sorbitol wash coupled with the salting out protocol resulted in values above the optimal 2.0 threshold (Figure IV-3). The sorbitol wash performed before lysis of the tissue included sorbitol, polyvinylpyrrolidone and β-mercaptoethanol (section 4.2.4, Jones et al., 2021). Sorbitol forms a hypertonic environment that causes water and cytoplasmic content to leak out of the cell (Lang et al., 2014), PVP binds to phenolic compounds that can oxidise

into quinones and bind to DNA and proteins (Pérez-Pérez et al., 2025). Lastly, ß-mercaptoethanol reduces disulphide bonds in proteins and inhibits the oxidation of polyphenols (Price et al., 1969). High purity ensures that DNA is free from contaminants that could inhibit polymerase activity during library amplification or sequencing, thereby improving overall library quality and sequencing success.

For long-read sequencing technologies, the integrity of DNA, particularly its fragment length, is a critical determinant of sequencing success and read length. Both extraction method and tissue type significantly influenced DNA fragment length (Figure IV-4, Supplementary table IV-7). The sorbitol wash coupled with the salting-out protocol produced DNA with notably longer fragments compared to the Qiagen MagAttract HMW kit and salting-out alone (Supplementary table IV-8). The Nanobind Big DNA kit also produced long fragments in the HMW range. Conversely, phenol chloroform and the Qiagen DNeasy kit resulted in the shortest DNA fragments (Supplementary figure IV-3). While tissue type also had a significant impact, specific post-hoc differences among tissue types were not elucidated (Supplementary table IV-8). These findings highlight the importance of careful protocol selection, especially in the presence of challenging biological factors, such as high concentrations of DNases (Angthong et al., 2020), which can severely impact DNA integrity. For instance, the effect of ß-mercaptoethanol, disrupting disulphide bonds, can inactivate DNases present in high amounts in crustacean tissue that fragment DNA (Russo et al., 2022). The choice of extraction mechanism is equally crucial for obtaining HMW DNA of appropriate integrity (Trigodet et al., 2022). Nanobind kits are utilising magnetic disks to bind DNA with minimal shearing and are specifically designed for HMW DNA extraction for PacBio sequencing (PacBio, 2024; Y. Zhang et al., 2016). However, we observed an underperformance in both DNA yield and purity (Supplementary figure IV-1, 2). This, coupled with their higher cost compared to non-kit protocols, should present significant considerations for their application in large-scale genomic sequencing efforts.

The primary measure of an extraction protocol's efficiency is its ability to yield DNA suitable for downstream library preparation and subsequent sequencing. In our study, 76% of the prepared libraries were successfully sequenced (Figure IV-1). Failures were predominantly attributed to insufficient DNA quantity or suboptimal fragment distribution. Even when extraction protocols yielded DNA of apparent good quality, issues during the library preparation led to failure: low yields during library preparation can result from the nuclease treatment step that removes damaged or un-ligated SMRTbell templates (in PacBio library

preparation) or from low recovery during AMPure clean up steps (both PacBio and Nanopore) (Bronner et al., 2025). This could indicate the presence of DNA damage and contaminants not detected during the DNA quality controls. This underscores the importance of rigorous quality control at early stages and the implementation of an optimised workflow.

The Nanopore libraries sequenced in our study yielded an overall low amount of data and short fragment lengths (Supplementary figure IV-4), which contrasted with typical expectations for Nanopore technology. The ligation-based kit generally produced higher sequencing output than the transposase-based kit (Supplementary figure IV-4). This aligns with expectations, as transposase kits introduce additional fragmentation, making them less suitable for maximising read length (Oxford Nanopore Technologies, 2024). However, even when shorter fragments were anticipated, low yields remained the main issue. The highest yield in a run obtained in our study was 1.8 Gb, while usual expected throughput ranges between 10 and 15 Gb per run (Wang et al., 2021). In the *C. quadricarinatus* genome study (Tan et al., 2020) the Nanopore output was 36 Gb (7x coverage) with an average length of 3419 bp. In the *C. destructor* genome study (Austin et al., 2022), 20 Nanopore flowcells were used with 106 Gb output and 6705 bp average length. These values are still below the expected Nanopore throughput, especially in fragment length. In our study, across all runs, most sequencing failed within a few hours, with pores becoming unavailable for sequencing (data not shown), indicating pore blockage, presumably by contaminants (Oxford Nanopore Technologies, 2024). An additional issue could stem from the long homopolymer stretches and short tandem repeats present in repetitive genomes, which negatively influence base detection as the DNA molecule passes through the nanopore (Fang et al., 2022). While R10 flowcells, with longer pore head, could potentially mitigate the issue (Sanderson et al., 2023), this was not pursued during this study. Instead, we focused our efforts on optimising PacBio sequencing.

With the aim of maximising the sequencing output, we compared PacBio amplification-free and amplification-based library preparation protocols. Significant differences were observed, with both HiFi read length and HiFi yield being higher for the amplification-based approach (Figure IV-6, Supplementary table IV-9). Ultra-low input amplification-based libraries yielded between 10 and 30 Gb, which is the expected output for the Sequel IIe system, while the Revio system can produce up to 120 Gb per SMRTcell (PacBio, 2025). A significant difference in P0 and P1 values was observed between amplification-based and amplification-free library preparation approaches (Supplementary table IV-10). Amplification-based

libraries generally exhibited higher P1 values, indicating higher percentage of occupancy and sequencing wells producing high quality sequencing. Overall, P1 correlates with higher sequencing yield for all libraries. However, when observed separately, only amplification free libraries show a correlation in P1 values and yield (Figure IV-7B). Low P1 values indicate sample quality issues, therefore higher P1 values and higher sequencing yield originates from samples with better DNA quality (Pacific Biosciences, 2020). The high P1 value for all amplification-based libraries, but no correlation with sequencing yield, could be indicative for an overall high DNA quality without contaminants that could interfere the sequencing. Therefore, the DNA sequence itself becomes the metric to evaluate the quality of the sequencing.

The evaluation of library preparation and sequencing is important especially when working with complexities of large genomes. PacBio recommends the amplification-based protocol as suitable for genomes up to 0.5 Gb, however this approach has been successfully applied to genomes up to 3.1 Gb (Bein et al., 2025). It has been shown that an inclusion of a PCR pre-amplification step improved the sequencing yield in samples with poor sequencing yield, and the data is adequate for genome assemblies when combined with data from amplification-free libraries (Bronner et al., 2025). Furthermore, a PCR amplification step is usual in Illumina library preparation protocols, which sequencing is successful for freshwater crayfish (Liao et al., 2024; Liu et al., 2024). For large genomes like freshwater crayfish, obtaining large amounts of data, and thus combining data from multiple libraries is crucial for successful assembly. In total we generated 640.26 Gb of data for *A. astacus* and 243.15 Gb of data for *A. bihariensis*, corresponding to 38x and 21x coverage, respectively. This highlights how different library preparation strategies are essential to overcome the challenges of large and complex genomes.

4.5. Conclusion

Our study evaluated various DNA extraction methods, tissue types, and library preparation strategies to establish an optimal workflow for generating high-quality high-molecular-weight (HMW) DNA suitable for PacBio HiFi sequencing of freshwater crayfish species. We demonstrate that both the chosen extraction protocol and tissue type are critical determinants of long-read sequencing success. The salting out protocol coupled with a sorbitol wash performed on muscle tissue from abdomen or legs proved to be suitable for library preparation. Furthermore, our investigation into library preparation revealed that the PacBio

amplification-based approach is suited for maximising sequencing yield. The variable sequencing success emphasizes the importance of quality control throughout the workflow.

This established workflow provides a foundation for future genomics studies in freshwater crayfish and other challenging invertebrate species. Future research could explore further optimisation of these protocols, as well as deeper understanding of the specific contaminants or DNA damage present in the sample. The field of genomics is highly dynamic, with DNA extraction and library preparation protocols, as well as sequencing technologies, continually being improved. Optimised approaches are key for accelerating genomic studies across a broad range of biodiversity, including often overlooked complex non model species, thereby contributing to an understanding of evolution, ecology, and conservation.

Funding

This work was funded by the Agence Nationale de la Recherche (GEODE: ANR-21-CE02-0028), the Deutsche Forschungsgemeinschaft (GEODE: DFG 490760095/TH 1807/7-1). Kathrin Theissinger is supported through the DFG Heisenberg programme (534452071/TH 1807/10-1). Ljudevit Luka Boštjančić is supported through the Deutsche Bundesstiftung Umwelt (39954/01). Lucian Pârvulescu is supported through the grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI–UEFISCDI, project number PN-III-P4-ID-PCE-2020-1187, within PNCDI III.

Acknowledgment

We thank the Radboudumc Genome Technology Center for the use of the Sequencing Core Facility (Nijmegen, Netherlands), which provided the PacBio SMRT sequencing service on the Sequel IIe platform. We also thank the Bioscientia Institut für Medizinische Diagnostik GmbH for providing the PacBio SMRT sequencing service on the PacBio Revio platform.

References

Abdelrahman, H., ElHady, M., Alcivar-Warren, A., Allen, S., Al-Tobasei, R., Bao, L., Beck, B., Blackburn, H., Bosworth, B., Buchanan, J., Chappell, J., Daniels, W., Dong, S., Dunham, R., Durland, E., Elaswad, A., Gomez-Chiarri, M., Gosh, K., Guo, X., … Zhou, T. (2017). Aquaculture genomics, genetics and breeding in the United States: current status, challenges, and priorities for future research. *BMC Genomics 2017 18:1*, *18*(1), 1–23. https://doi.org/10.1186/S12864-017-3557-1

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, 1–18. https://doi.org/10.7717/peerj.10340

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. *Frontiers in Genetics*, *12*(January), 1–8. https://doi.org/10.3389/fgene.2021.695763

Bein, B., Chrysostomakis, I., Arantes, L. S., Brown, T., Gerheim, C., Schell, T., Schneider, C., Leushkin, E., Chen, Z., Sigwart, J., Gonzalez, V., Wong, N. L. W. S., Santos, F. R., Blom, M. P. K., Mayer, F., Mazzoni, C. J., Böhne, A., Winkler, S., Greve, C., & Hiller, M. (2025). Long-read sequencing and genome assembly of natural history collection samples and challenging specimens. *Genome Biology*, *26*(1), 25. https://doi.org/10.1186/s13059-025-03487-9

Bonassin, L., Pârvulescu, L., Boštjančić, L. L., Francesconi, C., Paetsch, J., Rutz, C., Lecompte, O., & Theissinger, K. (2024). Genomic insights into the conservation status of the Idle Crayfish Austropotamobius bihariensis Pârvulescu, 2019: low genetic diversity in the endemic crayfish species of the Apuseni Mountains. *BMC Ecology and Evolution*, *24*(1), 78. https://doi.org/10.1186/s12862-024-02268-5

Boštjančić, L. L., Bonassin, L., Anušić, L., Lovrenčić, L., Besendorfer, V., Maguire, I., Grandjean, F., Austin, C. M., Greve, C., Hamadou, A. Ben, & Mlinarec, J. (2021). The Pontastacus leptodactylus (Astacidae) Repeatome Provides Insight Into Genome Evolution and Reveals Remarkable Diversity of Satellite DNA. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.611745

Bronner, I. F., Dawson, E., Park, N., Piepenburg, O., & Quail, M. A. (2025). Evaluation of controls, quality control assays, and protocol optimisations for PacBio HiFi sequencing on diverse and challenging samples. *Frontiers in Genetics*, *15*(January), 1–17. https://doi.org/10.3389/fgene.2024.1505839

Dahn, H. A., Mountcastle, J., Balacco, J., Winkler, S., Bista, I., Schmitt, A. D., Pettersson, O. V., Formenti, G., Oliver, K., Smith, M., Tan, W., Kraus, A., Mac, S., Komoroske, L. M., Lama, T., Crawford, A. J., Murphy, R. W., Brown, S., Scott, A. F., … Fedrigo, O. (2022). Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing. *GigaScience*, *11*, 1–13. https://doi.org/10.1093/gigascience/giac068

Espinosa, E., Bautista, R., Larrosa, R., & Plata, O. (2024). Advancements in long-read genome sequencing technologies and algorithms. *Genomics*, *116*(3), 110842. https://doi.org/10.1016/J.YGENO.2024.110842

Fang, L., Liu, Q., Monteys, A. M., Gonzalez-Alegre, P., Davidson, B. L., & Wang, K. (2022). DeepRepeat: direct quantification of short tandem repeats on signal data from nanopore sequencing. *Genome Biology*, *23*(1), 108. https://doi.org/10.1186/s13059-022-02670-6

Formenti, G., Theissinger, K., Fernandes, C., Bista, I., Bombarely, A., Bleidorn, C., Ciofi, C., Crottini, A., Godoy, J. A., Höglund, J., Malukiewicz, J., Mouton, A., Oomen, R. A., Paez, S., Palsbøll, P. J., Pampoulie, C., Ruiz-López, M. J., Svardal, H., Theofanopoulou, C., … Zammit, G. (2022). The era of reference genomes in conservation genomics. *Trends in Ecology & Evolution*, *37*(3), 197–202. https://doi.org/10.1016/J.TREE.2021.11.008

Fukuzawa, A. (2001). Invertebrate connectin spans as much as 3.5 microm in the giant sarcomeres of crayfish claw muscle. *The EMBO Journal*, *20*(17), 4826–4835.

https://doi.org/10.1093/emboj/20.17.4826

Glasel, J. A. (1995). Validity of nucleic acid purities monitored by 260nm/280nm absorbance ratios. *BioTechniques*, *18*(1), 62–63. http://www.ncbi.nlm.nih.gov/pubmed/7702855

Gregory, T. R. (2025). *Animal Genome Size Database*. http://www.genomesize.com

Gross, R., Lovrenčić, L., Jelić, M., Grandjean, F., Đuretanović, S., Simić, V., Burimski, O., Bonassin, L., Groza, M.-I., & Maguire, I. (2021). Genetic diversity and structure of the noble crayfish populations in the Balkan Peninsula revealed by mitochondrial and microsatellite DNA markers. *PeerJ*, *9*(August), e11838. https://doi.org/10.7717/peerj.11838

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. *Nature Ecology & Evolution*, *2*(3), 567–573. https://doi.org/10.1038/s41559-018-0467-9

Jenkins, T. L., Ellis, C. D., & Stevens, J. R. (2019). SNP discovery in European lobster (Homarus gammarus) using RAD sequencing. *Conservation Genetics Resources*, *11*(3), 253–257. https://doi.org/10.1007/s12686-018-1001-8

Jones, A., Torkel, C., Stanley, D., Nasim, J., Borevitz, J., & Schwessinger, B. (2021). High-molecular weight DNA extraction, clean-up and size selection for long-read sequencing. *PLOS ONE*, *16*(7), e0253830. https://doi.org/10.1371/journal.pone.0253830

Lang, I., Sassmann, S., Schmidt, B., & Komis, G. (2014). Plasmolysis: Loss of Turgor and Beyond. *Plants*, *3*(4), 583–593. https://doi.org/10.3390/plants3040583

Liao, M., Xu, M., Hu, R., Xu, Z., Bonvillain, C., Li, Y., Li, X., Luo, X., Wang, J., Wang, J., Zhao, S., & Gu, Z. (2024). The chromosome-level genome assembly of the red swamp crayfish Procambarus clarkii. *Scientific Data*, *11*(1), 1–8. https://doi.org/10.1038/s41597-024-03718-x

Liu, Z., Zheng, J., Li, H., Fang, K., Wang, S., He, J., Zhou, D., Weng, S., Chi, M., Gu, Z., He, J., Li, F., & Wang, M. (2024). Genome assembly of redclaw crayfish (Cherax quadricarinatus) provides insights into its immune adaptation and hypoxia tolerance. *BMC Genomics*, *25*(1), 746. https://doi.org/10.1186/s12864-024-10673-9

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. *Scientific Reports*, *12*(1), 2040. https://doi.org/10.1038/s41598-022-06027-8

Nagy, Z. T. (2010). A hands-on overview of tissue preservation methods for molecular genetic analyses. *Organisms Diversity & Evolution*, *10*(1), 91–105. https://doi.org/10.1007/s13127-010-0012-4

Oxford Nanopore Technologies. (2024). *Chemistry Technical Document*. *CHTD_500_v1_revAR_25Nov2024*.

PacBio. (2020). *Considerations for using the low and ultra-low DNA input workflows for whole genome sequencing PN 101-995-900*.

PacBio. (2024). *Nanobind® CBB kit For extraction of HMW (50–300+ kb) genomic DNA from cultured cells, cultured bacteria, and blood. Guide & overview*. *102-572-200 REV06*. https://www.pacb.com/wp-content/uploads/Guide-overview-Nanobind-CBB-kit.pdf#page=4.34

PacBio. (2025). *PacBio complete biological insights for confident decisions*. *102-326-857*

REV04. https://www.pacb.com/wp-content/uploads/HiFi-systems-brochure.pdf

Pacific Biosciences. (2020). *Guide - Step-By-Step Run Performance Evaluation*. *101-993-600 Version 01*.

Panova, M., Aronsson, H., Cameron, R. A., Dahl, P., Godhe, A., Lind, U., Ortega-Martinez, O., Pereyra, R., Tesson, S. V. M., Wrange, A.-L., Blomberg, A., & Johannesson, K. (2016). DNA Extraction Protocols for Whole-Genome Sequencing in Marine Organisms. In *Methods in Molecular Biology* (Vol. 1452, pp. 13–44). https://doi.org/10.1007/978-1-4939-3774-5_2

Pérez-Pérez, R., Pinski, A., Zaranek, M., Beckmann, M., Mur, L. A. J., Nowak, K., Rojek-Jelonek, M., Kostecka-Gugała, A., Petryszak, P., Grzebelus, E., & Betekhtin, A. (2025). Effect of potent inhibitors of phenylalanine ammonia-lyase and PVP on in vitro morphogenesis of Fagopyrum tataricum. *BMC Plant Biology*, *25*(1), 469. https://doi.org/10.1186/s12870-025-06440-x

Price, P. A., Stein, W. H., & Moore, S. (1969). Effect of Divalent Cations on the Reduction and Re-formation of the Disulfide Bonds of Deoxyribonuclease. *Journal of Biological Chemistry*, *244*(4), 929–932. https://doi.org/10.1016/S0021-9258(18)91875-2

Qiagen. (2023). *DNeasy® Blood & Tissue Handbook*.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Rees, D. J., Belzile, C., Glémet, H., & Dufresne, F. (2008). Large genomes among caridean shrimp. *Genome*, *51*(2), 159–163. https://doi.org/10.1139/G07-108

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and its Applications. *Genomics, Proteomics & Bioinformatics*, *13*(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Russo, A., Mayjonade, B., Frei, D., Potente, G., Kellenberger, R. T., Frachon, L., Copetti, D., Studer, B., Frey, J. E., Grossniklaus, U., & Schlüter, P. M. (2022). Low-Input High-Molecular-Weight DNA Extraction for Long-Read Sequencing From Plants of Diverse Families. *Frontiers in Plant Science*, *13*(May), 1–12. https://doi.org/10.3389/fpls.2022.883897

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*(8). https://doi.org/10.3390/genes14081627

Sambrook, J., & Russell, D. W. (2006). Purification of Nucleic Acids by Extraction with Phenol:Chloroform. *Cold Spring Harbor Protocols*, *2006*(1), pdb.prot4455. https://doi.org/10.1101/pdb.prot4455

Sanderson, N. D., Kapel, N., Rodger, G., Webster, H., Lipworth, S., Street, T. L., Peto, T., Crook, D., & Stoesser, N. (2023). Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction. *Microbial Genomics*, *9*(1). https://doi.org/10.1099/mgen.0.000910

Schell, T., Greve, C., & Podsiadlowski, L. (2025). Establishing genome sequencing and assembly for non-model and emerging model organisms: a brief guide. *Frontiers in Zoology*, *22*(1), 7. https://doi.org/10.1186/s12983-025-00561-7

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. *Frontiers in Genetics*, *11*(March). https://doi.org/10.3389/fgene.2020.00201

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., … Zammit, G. (2023). How genomics can help biodiversity conservation. *Trends in Genetics*, *39*(7), 545–559. https://doi.org/10.1016/j.tig.2023.01.005

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, *13*(1), 36–46. https://doi.org/10.1038/nrg3117

Trigodet, F., Lolans, K., Fogarty, E., Shaiber, A., Morrison, H. G., Barreiro, L., Jabri, B., & Eren, A. M. (2022). High molecular weight DNA extraction strategies for long-read sequencing of complex metagenomes. *Molecular Ecology Resources*, *22*(5), 1786–1802. https://doi.org/10.1111/1755-0998.13588

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, *39*(11), 1348–1365. https://doi.org/10.1038/s41587-021-01108-x

Wei, J., Huang, C., Nie, X., Wang, Y., Hong, K., Su, Q., Liu, M., Zhou, Q., Mai, Z., Liu, F., Li, H., Liu, C., Zeng, Z., Zhu, X., & Yu, L. (2024). Analysis of seven populations of cultured redclaw crayfish, Cherax quadricarinatus, using newly developed microsatellite markers. *Aquaculture Reports*, *35*, 102024. https://doi.org/10.1016/J.AQREP.2024.102024

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*(5), 3274–3284. https://doi.org/10.1016/j.ygeno.2021.07.017

You, X., Shan, X., & Shi, Q. (2020). Research advances in the genomics and applications for molecular breeding of aquaculture animals. *Aquaculture*, *526*, 735357. https://doi.org/10.1016/j.aquaculture.2020.735357

Yuan, J., Yu, Y., Zhang, X., Li, S., Xiang, J., & Li, F. (2023). Recent advances in crustacean genomics and their potential application in aquaculture. *Reviews in Aquaculture*, *July 2022*, 1501–1521. https://doi.org/10.1111/raq.12791

Zhang, T., Zhou, J., Gao, W., Jia, Y., Wei, Y., & Wang, G. (2022). Complex genome assembly based on long-read sequencing. *Briefings in Bioinformatics*, *23*(5). https://doi.org/10.1093/bib/bbac305

Zhang, Y., Zhang, Y., Burke, J. M., Gleitsman, K., Friedrich, S. M., Liu, K. J., & Wang, T. (2016). A Simple Thermoplastic Substrate Containing Hierarchical Silica Lamellae for High-Molecular-Weight DNA Extraction. *Advanced Materials*, *28*(48), 10630–10636. https://doi.org/10.1002/adma.201603738

Chapter V

Genomic insights into the conservation status of the Idle Crayfish *Austropotamobius bihariensis* Pârvulescu, 2019: low genetic diversity in the endemic crayfish species of the Apuseni Mountains

Lena Bonassin, Lucian Pârvulescu, Ljudevit Luka Boštjančić, Caterina Francesconi, Judith Paetsch, Christelle Rutz, Odile Lecompte, Kathrin Theissinger

Published in BMC Ecology and Evolution (2024) 24:78, doi: 10.1186/s12862-024-02268-5

Abstract

Background: Biodiversity in freshwater ecosystems is declining due to an increased anthropogenic footprint. Freshwater crayfish are keystone species in freshwater ecosystems and play a crucial role in shaping the structure and function of their habitats. The Idle Crayfish *Austropotamobius bihariensis* is a native European species with a narrow distribution range, endemic to the Apuseni Mountains (Romania). Although its area is small, the populations are anthropogenically fragmented. In this context, the assessment of its conservation status is timely.

Results: Using a reduced representation sequencing approach, we identified 4875 genomic SNPs from individuals belonging to 13 populations across the species distribution range. Subsequent population genomic analyses highlighted low heterozygosity levels, low number of private alleles and small effective population size. Our structuring analyses revealed that the genomic similarity of the populations is conserved within the river basins.

Conclusion: Genomic SNPs represented excellent tools to gain insights into intraspecific genomic diversity and population structure of the Idle Crayfish. Our study highlighted that the analysed populations are at risk due to their limited genetic diversity, which makes them extremely vulnerable to environmental alterations. Thus, our results emphasize the need for conservation measures and can be used as a baseline to establish species management programs.

Keywords

ddRAD-seq, endemic species, Idle Crayfish, genetic diversity, Apuseni Mountains

5.1. Background

Freshwaters are one of the most diverse ecosystems on the planet with exceptionally high levels of endemism (Williams-Subiza & Epele, 2021). Unfortunately, freshwater biodiversity is declining rapidly, faster than that of terrestrial or marine, with small endemic populations with limited distribution being especially affected (Tickner et al., 2020). Among the other taxa, freshwater crayfish are keystone species with a fundamental role in determining the structure and function in freshwater ecosystems (Reynolds et al., 2013). Moreover, crayfish have a high cultural significance, especially in Europe (Swahn, 2004). However, many native

crayfish populations are in decline and nearing extinction (Jussila et al., 2021). Native crayfish species richness in Europe is relatively low, with only six native species present (Kouba et al., 2014). Nevertheless, the species' genetic diversity is high, especially within the genus *Austropotamobius* (Jelić et al., 2016; Lovrenčić et al., 2020; Pârvulescu et al., 2019). The Idle Crayfish, *Austropotamobius bihariensis,* Pârvulescu, 2019, is an endemic freshwater crayfish species with the smallest distribution range restricted to the Apuseni Mountains in Romania (Pârvulescu, 2019). The small distribution range covers tributaries of the three Criş rivers in the Apuseni Mountains characterised by habitats with clean waters in the mountainous and sub-mountainous regions (Pârvulescu, 2019). Being a recently described species, the conservation status of *A. bihariensis* is not yet finally determined (Ion et al., 2024). With the other *Austropotamobius* species, it is the most vulnerable among native European freshwater crayfish species, being threatened by water quality deterioration, urbanisation, and other anthropogenic influences (Pârvulescu et al., 2020; Tarandek et al., 2023). Compared to other native species, the genus *Austropotamobius* has lower dispersal capacity, lower reproductive output and higher oxygen demand (Kozák et al., 2015; Pârvulescu et al., 2011). Moreover, the invasive spiny-cheek crayfish *Faxonius limosus* (Rafinesque, 1817) is spreading through the Romanian rivers, carrying the crayfish plague pathogen *Aphanomyces astaci* (Groza et al., 2021; Pacioglu et al., 2020; Pârvulescu et al., 2012; Ungureanu et al., 2020). This pathogen has already caused the devastation of several native crayfish populations throughout Europe, and its presence has been recently confirmed among *A. bihariensis* populations (Satmari et al., 2023; Theissinger et al., 2022).

Species with restricted distribution and limited dispersal capabilities are threatened by extinction due to loss of habitat, genetic variation, and invasive species (Allendorf & Lundquist, 2003). Endemic species present in a narrow geographical range are particularly vulnerable and are often characterised by a small population size (Jamieson, 2007). In small populations, genetic diversity is quickly declining due to genetic forces, posing a threat to the long-term survival of the species (Willi et al., 2022). Specifically, genetic variability is important for preserving the adaptive potential of a species, its reproductive success and disease resistance, especially in response to environmental changes (Çilingir et al., 2022; Reed & Frankham, 2003). Genetic characterisation of individuals within and amongst populations allows the identification of genetic population structure and gene flow for defining genetically similar conservation units (Woodruff, 2001). This information can then

be used to conduct informed management actions such as habitat restoration or species translocation (Zimmerman et al., 2020).

The assessment of genetic diversity is an initial step towards conservation actions. Genomic data allows to characterise and monitor genetic diversity, maximising the information obtained from each individual (Theissinger et al., 2023). Previous studies on *A. bihariensis* and other European freshwater crayfish taxa focused on the phylogenetic analysis of mitochondrial DNA markers (Pârvulescu et al., 2019), and the population diversity was assessed based on a small number of microsatellite loci (Pârvulescu et al., 2020). A small number of loci can limit the characterisation of the genetic diversity of species. This limitation can be overcome by using genome-wide assessments. (Theissinger et al., 2023). When a reference genome is unavailable, a reduced representation DNA sequencing (RRS), which sequences a random fraction across the entire genome, is a suitable approach to provide key insights into the genomic structure of a population. In particular, ddRADseq (double digest DNA restriction-site-associated DNA sequencing) provides a large single nucleotide polymorphism (SNP) dataset from a subset of the genome (Zimmerman et al., 2020). Unlike microsatellites, SNPs are more abundant and uniformly distributed across the genome, being found in both non-coding and coding regions of the genome, and thus increase the sensitivity of the analysis and robustness of population genetic estimates compared to microsatellite markers. Therefore, SNPs are appropriate markers for the assessment of demographic as well as functional processes (Zimmerman et al., 2020).

Here, we conducted the first population genomic analyses of the endemic freshwater crayfish species *A. bihariensis* using a large SNP dataset to aid the assessment of the species conservation status. We performed ddRAD sequencing to obtain an insight into the genomic variants and genetic diversity present in 13 populations, belonging to five river basins across the entire distribution range of this endemic species. Due to the low dispersal capability of the species, we hypothesised that the population structure is reflected by the river basins. Based on the identified SNPs, we also aimed to uncover unique genomic variants to identify populations of the highest conservation priority.

5.2. Methods

5.2.1. Sample collection and DNA extraction

Tissue samples from 235 individuals were obtained by collecting one pereopod from each individual and tissue was stored in 96% ethanol at 4 °C until DNA extraction. Sample

collection was conducted at 13 locations belonging to five river basins (Supplementary table V-1) across the species distribution range (Figure V-1), obtaining between 10 and 20 individuals per population (Supplementary table V-1). Considering the high number of individuals needed for population genetic studies, the sampling was not conducted in populations known to have a low number of individuals. DNA was extracted 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).

5.2.2. ddRAD sequencing

ddRAD libraries were produced by IGA Technology Services (Udine, Italy) using a custom protocol with minor modifications with respect to Peterson's double digest restriction-site associated DNA preparation (Peterson et al., 2012). The enzyme pair was selected based on *in silico* analysis of 24 Gb of PacBio HiFi reads of the species (unpublished data). Genomic DNA was fluorometrically quantified using Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA, USA) and normalised to a uniform amount. It was then double digested with 2.4 U of both PstI and EcoRI endonucleases (New England BioLabs, USA) in 30 µL reaction supplemented with CutSmart Buffer and incubated at 37°C for 90 min, then at 75°C for 20 min. Fragmented DNA was subsequently ligated with 180 U of T4 DNA ligase (New England BioLabs, USA) and 2.5 pmol of overhang barcoded adapters for both cut sites in a 50 µL reaction incubated at 23°C for 60 min and at 20°C for 60 min, followed by 20 min at 65°C. Samples were pooled on multiplexing batches and purified with 1.5 volumes of AMPureXP beads (Agencourt). For each pool, targeted fragment distributions were collected using BluePippin (Sage Science Inc., USA) with a set range 400 bp - 550 bp. The gel eluted fraction was amplified with indexed primers using Phusion High-Fidelity PCR Master Mix (New England BioLabs, USA) in a final volume of 50 µL and subjected to the following thermal protocol: [95°C, 3 min] - [95°C, 30 s - 60°C, 30 s - 72°C, 45 s] x 10 cycles - [72°C, 2 min]. PCR products were purified with 1 volume of AMPureXP beads (Agencourt). The resulting libraries were checked with both Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA) and Bioanalyzer DNA assay (Agilent technologies, Santa Clara, CA). Libraries were sequenced

with 150 cycles in paired-end mode on a NovaSeq 6000 instrument following the manufacturer's instructions (Illumina, San Diego, CA).

5.2.3. ddRADseq data processing

Demultiplexing of raw Illumina reads was performed using the process_radtags utility included in Stacks v2.61 (Catchen et al., 2013). Sequence quality was assessed using FastQC v0.11.9 (Andrews, 2010) and MultiQC v1.9 (Ewels et al., 2016). *De novo* assembly was performed in Stacks v2.62 and the parameters for the assembly were selected following the recommendations by (Paris et al., 2017). For the *de novo* building of loci, creation of a catalog of loci and SNP calling, the pipeline module denovo_map.pl included in Stacks v2.62 was used with the following parameters: -m 6, -M 2 and -n 2. Using Plink v1.90 (Purcell et al., 2007), SNPs and individuals with a missing call frequency greater than 0.1, and SNPs with minor allele frequency lower than 0.05 were filtered out. The filtered SNPs and individuals were used to create a whitelist and run the populations module in Stacks v2.62.

5.2.4. Population genetic diversity

The population genetic diversity was assessed by calculating the percentage of polymorphic loci (P), number of private alleles, observed heterozygosity (H_O), expected heterozygosity (H_E), and inbreeding coefficient (F_{IS}) based on SNPs in the populations module in Stacks v2.62. The analysis was first done by assigning individuals to populations and then by assigning individuals to river basins. The effective population size (N_e) was estimated using NeEstimator v2 (Do et al., 2014). To obtain more reliable results, N_e was calculated using the LD method and heterozygote excess method, and 95% Cis were calculated by a jackknife-on-samples method. To avoid bias caused by rare allele presence, N_e was calculated at a Minor Allele Frequency (MAF) equal or smaller than 0.02 and 0.01.

5.2.5. Population genetic structure

The population differentiation was estimated by pairwise comparisons of the fixation coefficient (F_{ST}) calculated in the populations module in Stacks v2.62. The genetic structure was assessed with principal component analysis (PCA) and Bayesian clustering algorithm implemented in fastStructure (Raj et al., 2014). PCA was performed using Plink v1.90 and plotted using the ggplot2 R package (Wickham, 2016). We performed the fastStructure analysis for K values between 1 and 12 to determine the most likely value for K determined by marginal likelihood. The results were visualised using the pophelper R package (Francis, 2017). FineRADstructure (Malinsky et al., 2018) was used to observe co-ancestry among

individuals and populations based on haplotypes with default parameters. Final editing of the resulting graphics was done in Inkscape 1.2.1 (Inkscape, 2020).

5.2.6. Ethical statement

All tissue samples involved in this study were taken in accordance with international ethical guidelines. No animal was killed, and after collecting the sample, the animal was released exactly where it was caught. Also, the necessary approvals were requested and obtained, according to the legislation in force in the area: Romanian Academy (1/CJ/13.01.2021), Romanian Ministry of Water and Forests (DGB/2/R5787/16.08.2022), Apuseni Nature Park Administration (199/09.09.2022), National Agency for Protected Areas (882/15.09.2022), Environmental Protection Agencies in the geographical area (8027/26.07.2022, 76/20.09.2022, 53/20.09.2022, 29/27.10.2022, 77/28.10.2022).

5.3. Results

5.3.1. ddRAD data assembly

We sequenced ddRAD libraries from 235 individuals across 13 populations (Figure V-1, Supplementary table V-1). In total 3 371 654 698 reads were obtained with a length of 135 bp. After quality filtering of the reads, in total 3 265 444 937 reads were retained, ranging from 872 095 to 92 642 412 reads per individual, with a mean of 13 895 510 reads (Supplementary table V-2). In total, 2 042 818 loci were assembled with a mean length of 260.78 bp and 1 381 639 SNPs were identified. After SNP and individual missingness filtering, RAN population was removed due to too much missing data. The final dataset consisted of 4 875 SNPs and 205 individuals from 12 populations (Figure V-1, Table V-1).

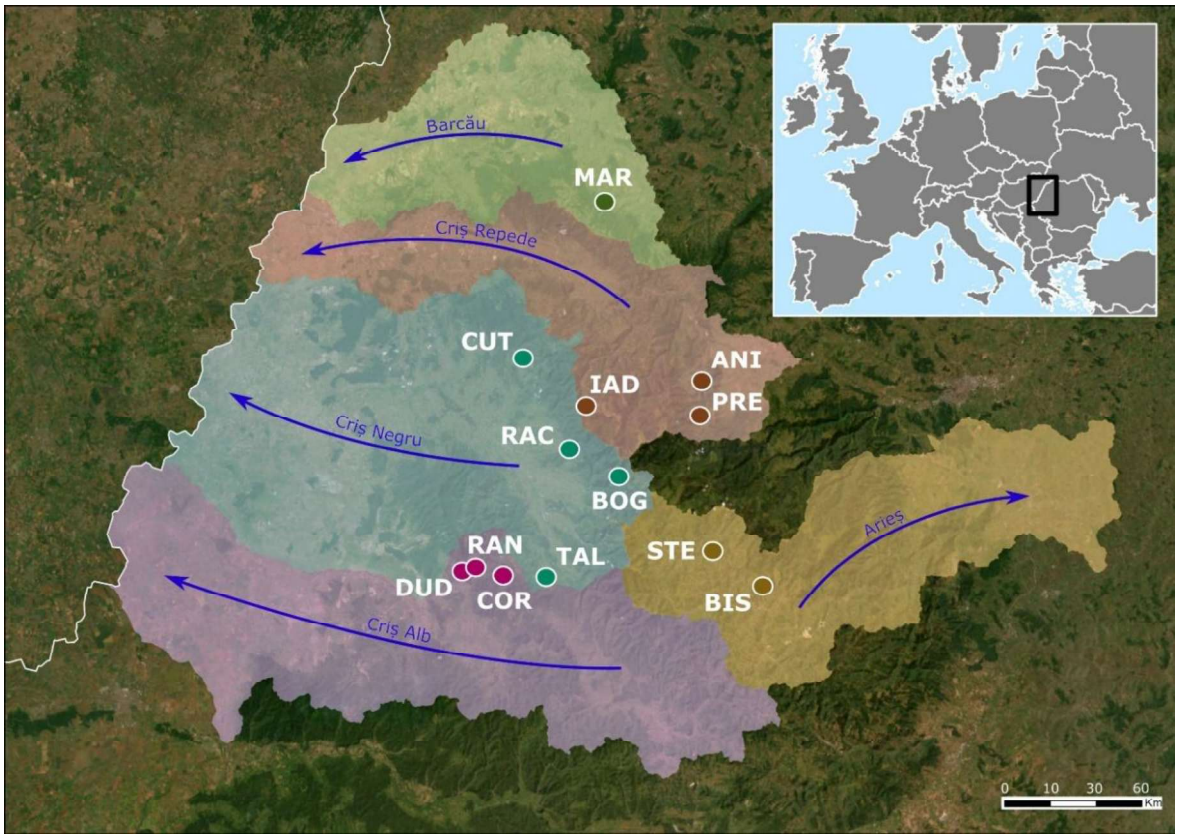


Figure V-1. Distribution map of the sampled locations. Colours denote different river basins and arrows the direction of river flow. Population acronyms: DUD – Dudușoaia, COR – Corbului, TAL – Tâlniciorii, BOG – Boga, RAC – Racu, CUT – Cuților, IAD – Iadei, PRE – Preluca, ANI – Anișelului, MAR – Mare, STE - Starpă, BIS – Bistrii. The base map layout is provided by Earthstar Geographics (https://www.terracolor.net) and river basin boundaries were delineated using the HydroBASINS database (https://www.hydrosheds.org).

5.3.2. Population genetic diversity

The observed overall genetic diversity of *A. bihariensis* populations was similar on a population level (Table V-1) and at the river basin level (Table V-2). Considering all populations, the percentage of polymorphic loci ranged between 0.212% and 0.464%. On the population level, only DUD had private alleles (n=12). On the river basin level, private alleles were identified in the Alb (n=77) and Negru (n=10) river basins. The values for observed heterozygosity (H_O) ranged from 0.164 to 0.311, with the lowest value for the ANI population and the highest for TAL. The H_O values for river basins ranged from 0.178 (Criș Repede basin) to 0.278 (Arieș basin). The values of expected heterozygosity (H_E) ranged from 0.149 (ANI) to 0.294 (RAC). H_E values for river basins ranged from 0.177 (Criș Repede basin) to 0.313 (Criș Alb basin). The inbreeding coefficient (F_{IS}) ranged from -0.058 to 0.011 for the populations and from -0.058 to 0.109 for the river basins, respectively.

Table V-1. River basins and populations used in genetic analyses, number of individuals per population, percentage of polymorphic loci, number of private alleles, observed (H_O) and expected (H_E) heterozygosity and inbreeding coefficient (F_{IS})

River basin	Population	Acronym	N	Polymorphic loci %	Private alleles	H_O	H_E	F_{IS}
Crișul Alb	Dudușoaia	DUD	20	0.409	12	0.273	0.259	-0.001
Crișul Alb	Corbului	COR	10	0.435	0	0.313	0.292	0.002
Crișul Negru	Racu	RAC	17	0.459	0	0.321	0.294	-0.031
Crișul Negru	Tâlniciorii	TAL	18	0.464	0	0.331	0.290	-0.058
Crișul Negru	Cuților	CUT	19	0.303	0	0.210	0.198	-0.011
Crișul Negru	Boga	BOG	20	0.342	0	0.238	0.232	0.011
Crișul Repede	Iadei	IAD	18	0.266	0	0.181	0.173	-0.002
Crișul Repede	Preluca	PRE	19	0.317	0	0.179	0.172	-0.001
Crișul Repede	Anișelului	ANI	7	0.212	0	0.164	0.149	0.007
Barcău	Mare	MAR	17	0.383	0	0.251	0.217	-0.058
Arieș	Bistrii	BIS	20	0.381	0	0.210	0.190	-0.022
Arieș	Starpă	STE	20	0.309	0	0.203	0.190	-0.007

Table V-2. Number of individuals per river basin, percentage of polymorphic loci, number of private alleles, observed (H_O) and expected (H_E) heterozygosity and inbreeding coefficient (F_{IS}) of individuals grouped by river basins.

River basin	N	Polymorphic loci %	Private alleles	H_O	H_E	F_{IS}
Crișul Alb	30	0.475	77	0.287	0.313	0.096
Crișul Negru	72	0.498	10	0.272	0.307	0.109
Crișul Repede	42	0.339	0	0.178	0.177	0.010
Barcău	17	0.383	0	0.252	0.217	-0.058
Arieș	40	0.408	0	0.207	0.205	0.017

The results of population size estimation are shown in Table V-3. Across the majority of the investigated populations, the effective population size estimated for 0.02 and 0.01 MAF with LD method ranged from 3.1 to 86.3, while estimated with heterozygote excess method ranged from 6.9 to 1033.3. The estimates were indefinable (∞) for the populations BOG, ANI and

COR estimated with both methods. For the majority of the populations, 95% CIs were wide, with the upper limit indefinable (∞).

Table V-3. Effective population size estimation based on linkage diseqilibrium (N_eLD) and heterozygote excess (N_eb) for 0.02 and 0.01 minor allele frequency (MAF) and 95% CI based on jackknifing method – N_e estimator. Population acronyms: DUD – Duduşoaia, COR – Corbului, TAL – Tâlniciorii, BOG – Boga, RAC – Racu, CUT – Cuţilor, IAD – Iadei, PRE – Preluca, ANI – Anişelului, MAR – Mare, STE – Starpă, BIS – Bistrii.

Population	N_eLD	95% CI	N_eb	95% CI
DUD	21.0	6.8 - ∞	1033.3	48.0 - ∞
COR	∞	23.4 - ∞	∞	56.8 - ∞
BIS	3.1	1.0 – 78.8	17.3	13.0 – 25.9
STE	72.6	18.7 - ∞	47.4	22.2 - ∞
MAR	20.1	10.7 – 21.1	6.9	6.2 – 8
RAC	86.3	27.3 – 88.1	14.9	11.9 – 20.1
TAL	26.2	8.3 – 99.2	8.3	7.3 – 9.6
CUT	55.9	21.8 - ∞	28.9	17.0 – 102.5
BOG	∞	∞ - ∞	∞	∞ - ∞
IAD	39.2	10.8 - ∞	139.9	28.4. - ∞
PRE	9.6	3.3 – 22.4	449.4	39.3 - ∞
ANI	∞	∞ - ∞	∞	336.9 - ∞

5.3.3. Population genetic structure

The pairwise fixation index (F_{ST}) ranged from 0.025 (PRE – IAD) to 0.29 (DUD – IAD) (Figure V-2). The highest F_{ST} is present in the DUD population, while the lowest differentiation is seen in the MAR population.

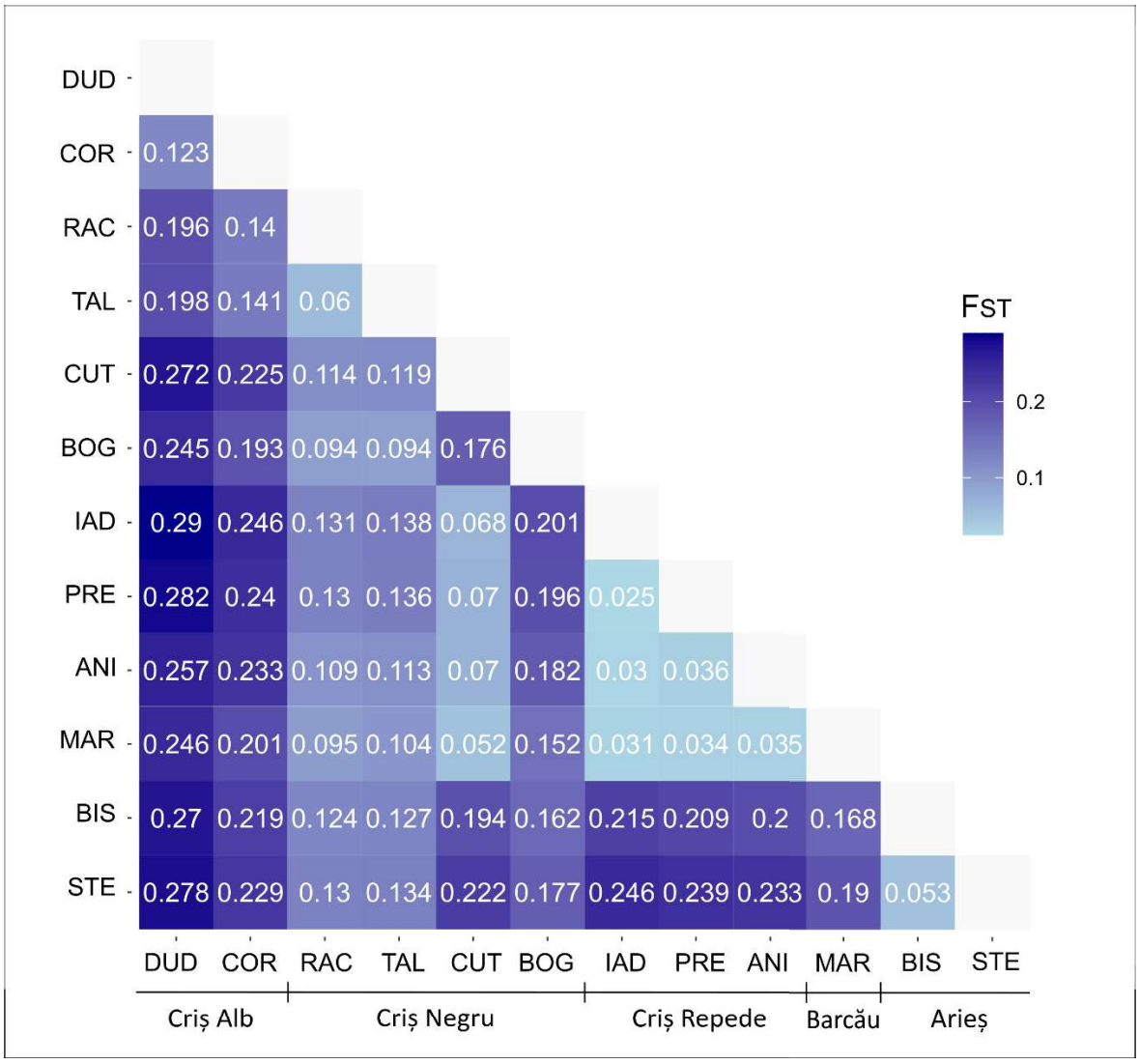


Figure V-2. Fixation coefficient (F_{ST}) between each population pair. Darker blue indicates higher F_{ST} values. Population acronyms: DUD – Duduşoaia, COR – Corbului, TAL – Tâlniciorii, BOG – Boga, RAC – Racu, CUT – Cuţilor, IAD – Iadei, PRE – Preluca, ANI – Anişelului, MAR – Mare, STE – Starpă, BIS – Bistrii.

The structuring of population with K=5 (number of river basins), K=8 (highest marginal likelihood revealed by fastStructure analysis) and K=12 (number of populations) is shown in Figure V-3. With K=5, the DUD and COR populations shared one cluster, in accordance with their belonging to the Criș Alb river basin. A second distinct cluster was formed by the populations of the Arieș river basin. RAC and TAL populations, originating from Criș Negru river basin, form another distinct cluster with the largest number of admixed individuals. All individuals from BOG population belonged to one distinct cluster. CUT, MAR, IAD, PRE and ANI populations belonged to one other cluster combining populations from river basins Criș Negru (CUT), Barcău (MAR) and Criș Repede (IAD, PRE, ANI) (Figure V-3). With

K=8, the individuals from DUD, COR populations (Criș Alb basin), and BOG, RAC and TAL (Criș Negru basin) formed each their own unique cluster (Figure V-3). Some individuals from the RAC and TAL populations showed genetic admixture and belonging to multiple clusters. The BIS and STE populations together formed one cluster in accordance with their geographical location, both belonging to the Arieș river basin. The MAR population (Barcău basin), and ANI, PRE, IAD populations from Criș Repede basin formed one cluster, shared with all the individuals from the CUT population (Criș Negru basin). With K=12, the clustering of the populations was the same as for K=8, except for some individuals from populations CUT (Criș Negru basin), MAR (Barcău basin), IAD, PRE, ANI (Criș Repede basin) which split into a separate cluster.

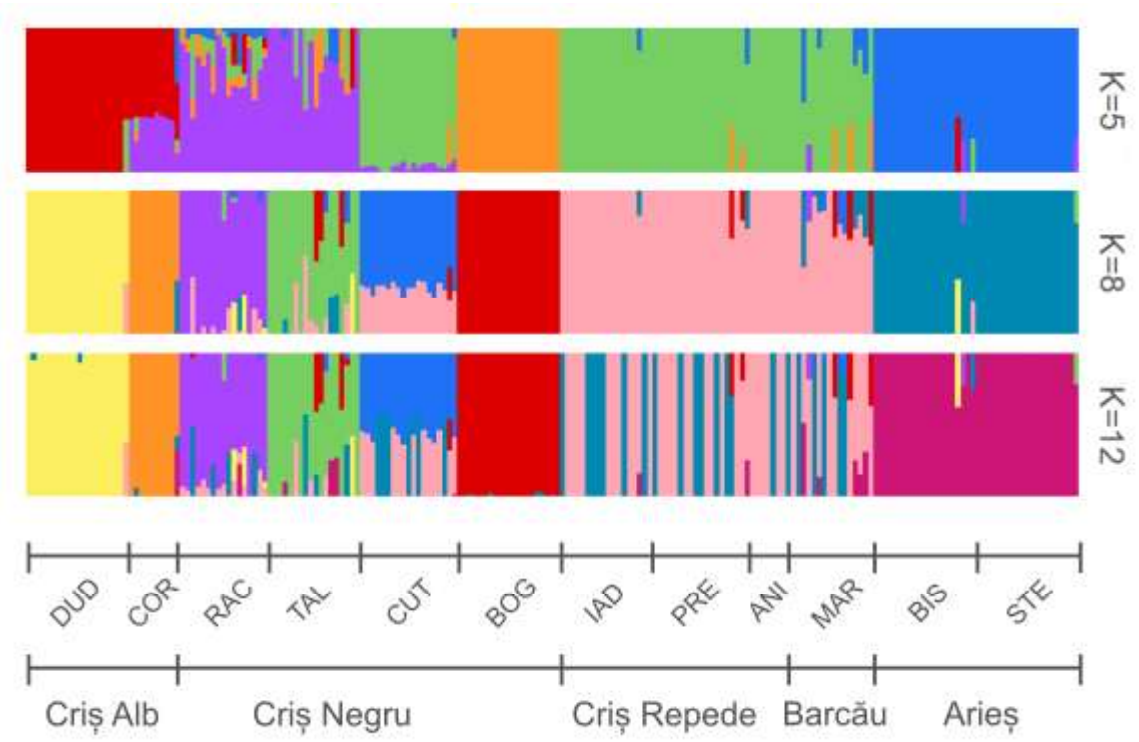


Figure V-3. Population structure based on fastStructure analysis for K=5, 8, and 12. Different colours represent different genetic clusters. Each column represents one individual. Population acronyms: DUD – Dudușoaia, COR – Corbului, TAL – Tâlniciorii, BOG – Boga, RAC – Racu, CUT – Cuților, IAD – Iadei, PRE – Preluca, ANI – Anișelului, MAR – Mare, STE – Starpă, BIS – Bistrii. Columns with different colours indicate admixture of populations.

The PCA analysis showed congruent results to structure analysis with K=8, with the first two components explaining 57.3% of the variance (Figure V-4A), while PC3 explains 10.1% of

the variance (Figure V-4B). Based on the variation represented in PC1, populations belonging to the river basin Criș Alb grouped separately from the rest of the analysed populations. Based on the variation from PC2, the populations of the Criș Repede river basins grouped separately, while BOG (Criș Negru basin) separated based on the variation from PC3.

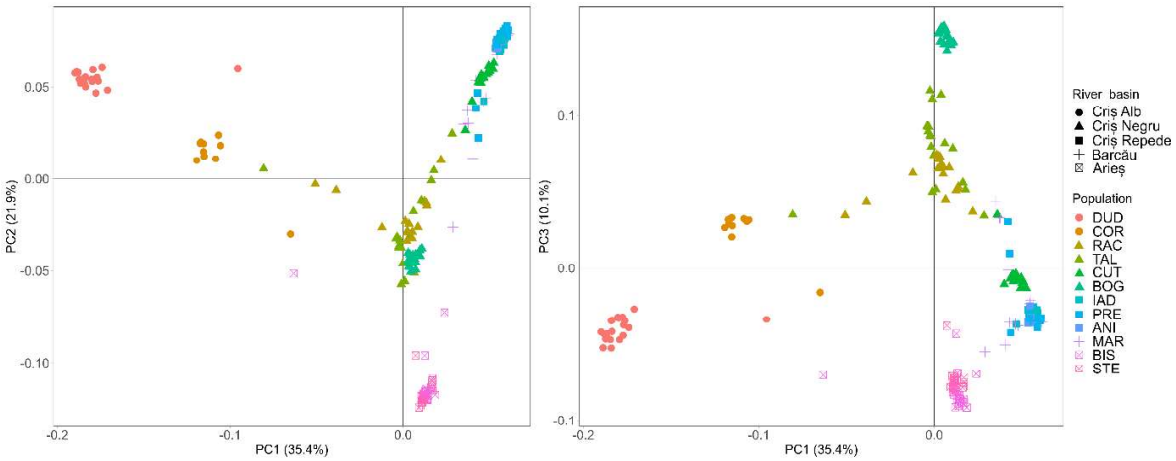


Figure V-4. PCA. Different colours denote different populations, and different symbols the river basins to which populations belong. A – PC1 and PC2, B – PC1 and PC3. Population acronyms: DUD – Dudușoaia, COR – Corbului, TAL – Tâlniciorii, BOG – Boga, RAC – Racu, CUT – Cuților, IAD – Iadei, PRE – Preluca, ANI – Anișelului, MAR – Mare, STE – Starpă, BIS – Bistrii.

The results of genetic structure based on shared co-ancestry matrices are represented in Figure V-5. Generally, individuals shared higher genetic similarity and co-ancestry with individuals from the same river basin. Furthermore, the clustering dendrogram showed clustering of the populations based mainly on the river basin. The populations from the river basin Criș Alb had the highest levels of ancestry within the same river basin compared to other populations. CUT (Criș Negru basin), MAR (Barcău basin), IAD and ANI (Criș Repede basin) populations formed one cluster and shared a more recent ancestry than with other populations. The individuals belonging to the TAL, RAC and BOG populations (all located in Criș Negru basin) grouped together. Population BOG had higher similarity within the population than with other populations. The TAL and RAC populations did not separate clearly but showed similarity between the two populations.

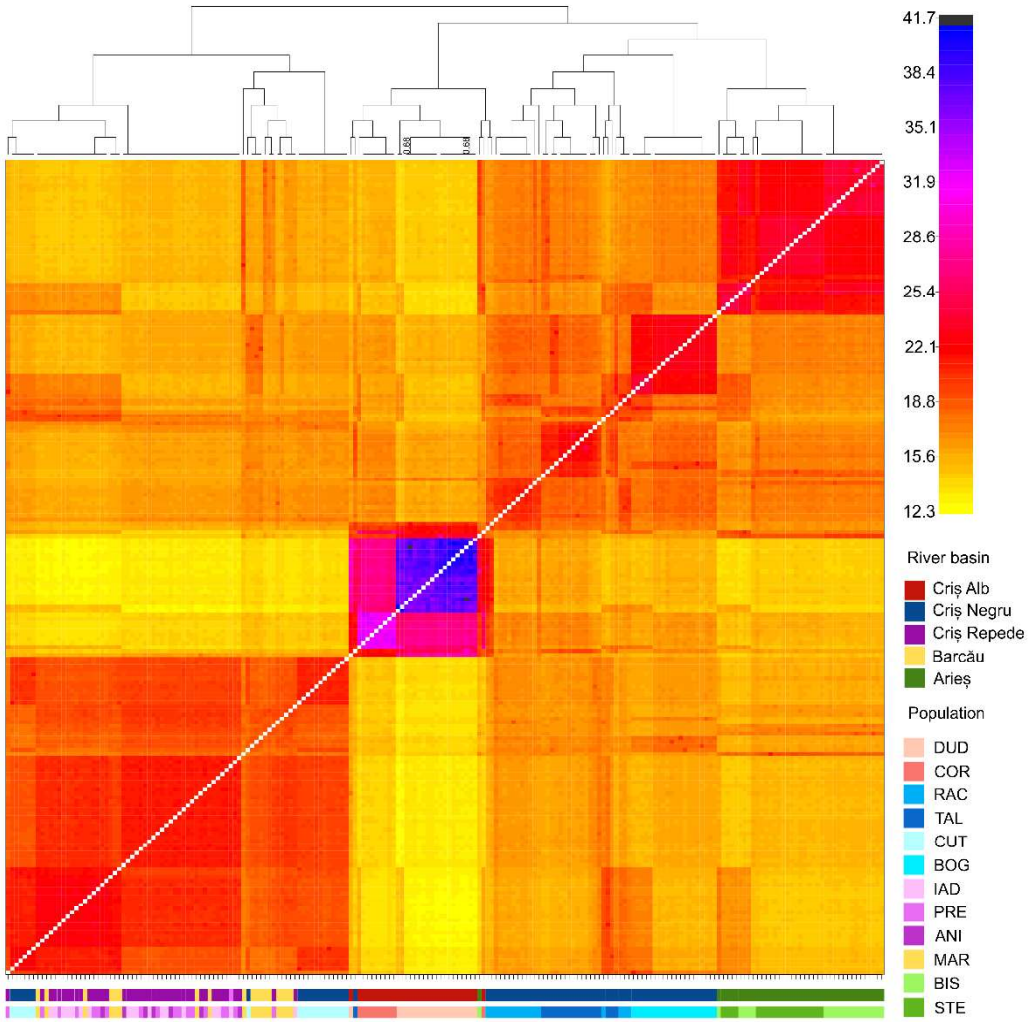


Figure V-5. Co-ancestry matrix of pairwise genetic similarity between the individuals. Darker (blue and black) colours represent high level of genetic similarity and co-ancestry (relatedness), and light colours lower level of co-ancestry. The clustering of individuals is shown in a dendrogram on top of the matrix. Posterior probabilities values are 1 unless indicated on branches. Colour bars indicate river basins and populations. Population acronyms: DUD – Dudușoaia, COR – Corbului, TAL – Tâlniciorii, BOG – Boga, RAC – Racu, CUT – Cuților, IAD – Iadei, PRE – Preluca, ANI – Anişelului, MAR – Mare, STE - Starpă, BIS – Bistrii.

5.4. Discussion

In this study, we applied reduced representation genome sequencing of 235 crayfish individuals to assess the population genomic structure and variation among the populations of the freshwater crayfish *A. bihariensis*, an endemic species to the Apuseni Mountains in Romania with the most restricted range of all freshwater crayfish species in Europe. We show that the populations' genomic structure reflects the distribution of the populations in the river basins. Moreover, the identified low genetic diversity presents a risk for the populations and highlights the need for targeted conservation actions.

5.4.1. Population genetic diversity and structure

Based on 4 875 SNPs from 12 populations covering the entire distribution range of the species, we observed an overall low genetic diversity, with almost no private alleles within populations, low heterozygosity, and low polymorphism. We confirmed our hypothesis that population structuring mainly depends on river basins. Low genetic diversity is usually attributed to small population sizes, genetic inbreeding and/or genetic drift, and bottlenecks (Frankham et al., 2010). In small and isolated populations, there is a higher likelihood of loss of rare alleles and of low migration rates (Meffe & Carroll, 1997). This causes small effective population sizes (N_e) and deficiency of heterozygotes, hence resulting in decreased observed heterozygosity and lower genetic variation (Bassitta et al., 2021). Private alleles were found only in the DUD population, and on the river basin level, in Criș Alb and Criș Negru river basins, possibly because of the geographical isolation of these populations. Private alleles are shared among populations within river basins, but are unique among the river basins. Therefore the estimated number of private alleles is higher for the Criș Alb river basin compared to the DUD population alone. Our results showed small N_e for all populations, below the recommended 100/1000 rule for avoiding inbreeding and maintaining evolutionary potential (Frankham et al., 2014). These low values indicate a higher extinction risk in the long-term, especially for species with low reproductive rates (Pérez-Pereira et al., 2022). Higher values of N_e based on heterozygote excess methods were estimated for DUD, IAD and PRE populations. However, the upper limit of the confidence interval, as well as Ne values for COR, ANI and BOG populations, show indefinite values, indicating the method cannot provide a precise estimation, possibly because of small sample size or missing information. In our analysis, F_{IS} values were around 0, which suggests random-mating in the population (Nichols, 2017). Furthermore, there have been no reports of mortality or severe

natural events, such as drought, floods, or disease outbreaks, which could indicate a bottleneck event. Therefore, we interpret the low values of genetic diversity as the result of a combination of genetic forces, low dispersal capability, as well as habitat fragmentation.

The previous study based on five microsatellite loci, observed heterozygosity (H_O) levels ranged from 0.325 to 0.834 in *A. bihariensis* (Pârvulescu et al., 2020). Lower values of heterozygosity observed in our study (0.164 – 0.311) are expected since SNPs have lower mutation rates than microsatellites and can present only two allelic states in diploid species (Haasl & Payseur, 2011; Zimmerman et al., 2020). Reports on the population genetics of other European freshwater crayfish species have mostly been based on microsatellite markers (Gross et al., 2021; Lovrenčić, Temunović, Bonassin, et al., 2022; Schrimpf et al., 2014, 2017), while genome-wide studies using SNPs are still lacking. However, studies using SNPs on other species from the order Decapoda showed similar or lower heterozygosity levels compared to this study. In the lobster species *Panulirus homarus* and *Panulirus ornatus,* H_O ranged from 0.05 to 0.15, and the average H_O for the crayfish species *Procambarus clarkii* was 0.0047 (Farhadi, Jeffs, et al., 2022; Farhadi, Pichlmueller, et al., 2022; Yi et al., 2018). In our study, low genetic diversity was also seen in the small percentage of polymorphic loci and private alleles in the populations. All summary statistics (heterozygosity, number of private alleles, and percentage of polymorphic loci) suggested limited genetic variation within the analysed populations. Several studies showed that low genetic diversity can lead to faster extinction (Kardos et al., 2021; Lynch et al., 2016; Razgour et al., 2019). In those cases, the population can only persist if it is able to adapt to environmental change, or able to migrate to other areas. Thus, a lower population's ability to adapt to changing environments can increase vulnerability to environmental pressures (Reed & Frankham, 2003). Based on the fixation coefficient (F_{ST}), which is a measure of population differentiation, the highest differentiation was observed between the populations of the river basin Criș Alb and the rest of the populations, with values above the significant level of differentiation (i.e., $F_{ST} = 0.15$ (Frankham, 2003)). We detected the strongest differentiation based on F_{ST} values between the Criș Alb river basin populations and the rest of the populations. Similar results were obtained based on PCA with Criș Alb and Arieș river basin populations, both being most differentiated from the rest of the populations. fastStructure analysis revealed a most likely grouping of populations into eight genetic clusters, with single populations belonging to unique clusters. The Criș Alb river basin populations and the Arieș river basin populations form a separate cluster, indicating longer isolation periods from the other populations. Strong genetic

structure on a narrow geographic range has been observed in other crayfish populations of European *Astacus astacus* (Gross et al., 2021; Lovrenčić, Temunović, Gross, et al., 2022) and Australian *Euastacus bispinosus* (Miller et al., 2014) and *Euastacus armatus* (Whiterod et al., 2017). This observation is expected for species with low vagility and limited dispersal capacity, such as crayfish (Clay et al., 2020).

Based on co-ancestry analysis, which indicates the degree of sharing haplotypes between individuals/populations (Lawson et al., 2012), the populations group mostly according to river basins. The populations of the Criș Repede river basin and the population MAR and CUT show possible presence of gene flow. The same populations also belong to a joint genetic cluster based on our analyses, and have lower F_{ST} values indicating recent gene flow between these populations, which reduces genetic differentiation. Considering their geographic location in a karstic area, the underground connectivity between CUT population and Criș Repede river basin is highly plausible (Matočec et al., 2002; West et al., 2020). Given the proximity (ca. 50 m) of the stream heads of the population MAR (Barcău river basin) to several tributaries of the Criș Repede river basin, plus proximity to the settlement Făgetu (Sălaj County), the MAR population, unique in the Barcău river basin, is likely a result of human-mediated translocation, as it has been already hypothesised based on microsatellites (Pârvulescu et al., 2020). In the latter case, no recent or past karstic substrate could allow underground connections of the MAR population to the Criș Repede populations. Microsatellites were often used more frequently in genetic population studies. However, large SNP datasets have higher resolution power and can detect the population's genetic structure more reliably (Haasl & Payseur, 2011). Microsatellite markers can also overestimate the genetic variability because of the intrinsic high mutation rates, which leads to homoplasy, making it difficult to discern ancestry from independent mutations (Zimmerman et al., 2020). In our case, the SNPs provided a higher resolution of the population structure and showed more precision in clustering analyses than microsatellite data for the same populations, where the individuals were grouped into only one cluster (Pârvulescu et al., 2020).

5.4.2. Conservation

Freshwater crayfish are considered keystone species with important ecological functions in maintaining the structure and functioning of their ecosystems (Reynolds et al., 2013). Thus, the low genetic diversity of the Idle Crayfish and the small effective population sizes are particularly concerning. Reduced genetic diversity in a population can lead to decreased adaptability and decreased resilience to environmental changes, making populations less

capable of effectively performing their ecological functions (Barrett & Schluter, 2008). The decline of Idle Crayfish populations can cause cascading effects through stream ecosystems, with potentially dramatic consequences on their biodiversity (Barnett et al., 2020). Therefore, whenever present, native crayfish should be regarded as umbrella species of conservation focus in protected areas such as the Apuseni mountains, which is regarded as a biodiversity hotspot (Bálint et al., 2011).

Even though *A. bihariensis* is found in multiple protected areas, there are no species-specific conservation programs in place yet, making it vulnerable especially to anthropogenic influence. The results presented here can provide important information to build an appropriate conservation program. Based on the genetic diversity of the populations, considerations of the adaptive potential of the populations are needed to make efficient conservation decisions, as populations with different genetic diversities have different capabilities of responding and minimising the effect of changing environments (Holderegger et al., 2006). Identifying the genetic characteristics of the individuals and populations can inform translocations for endangered species with the goal of restoring a population and increasing its genetic diversity (Weeks et al., 2011). Considering the population structure revealed in this study, and the identified population differentiation according to their river basins, our results suggest that translocations of this species would be possible among populations within the same gene pool. Such actions could be applied between the populations in the river basins Criș Repede and Barcău, as well as within the populations of the river basin Arieș, considering their genetic similarity. However, translocation actions should also take into consideration the genetically unique populations within the river basins, BOG (Criș Negru) and DUD and COR (Criș Alb), which contribute to the overall intraspecific diversity of this species.

Reintroduction and translocation actions have been proven successful for the crayfish species *Astacus astacus* (Linnaeus, 1758) and *Austropotamobius pallipes* (Lereboullet, 1858) (*sensu lato*) whose populations were extinct due to the crayfish plague disease (Schulz et al., 2002). Although useful, the translocations can carry risks of introducing diseases, such as the crayfish plague. Since known carriers of *A. astaci* have already been identified in the Romanian's freshwaters (Ungureanu et al., 2020), conservation actions need to consider the potential threat of the disease introduction in unaffected populations (Manenti et al., 2021). Knowing the risks and problems that translocation carries, reintroduction and translocation actions should remain a last resort in the face of the eventual extinction of some populations.

Until then, appropriate measures to preserve habitats and thoroughly prevent colonisation by invasive species should remain the main efforts in the short and medium term.

The highest conservation priority should be addressed towards the populations carrying unique genetic composition (Criș Alb and Criș Negru populations). Prospectively, as some populations might be more adaptable to potential environmental changes, further studies are needed to identify SNPs associated with more resilient phenotypes. Reference genomes can be highly useful to identify SNPs involved in specific traits or disease resistance (Formenti et al., 2022; Theissinger et al., 2023). However, it is still challenging to generate high-quality reference genomes, especially for non-model invertebrate species with large and repetitive genomes, characteristic of Decapoda (Rutz et al., 2023). Even in species where reference genomes are available, whole genome re-sequencing of a large number of individuals needed for population genomic studies is a financially exhausting and time-consuming approach. ddRADseq is useful for obtaining genomic SNPs without a reference genome, and is appropriate for monitoring as a reproducible and low-priced method.

5.5. Conclusion

In this study, the ddRAD approach allowed the identification of genetically distinct populations which require monitoring and priority in the conservation management of the species. Using genomic approaches for monitoring the genetic diversity of a species allows more comprehensive information than traditional single markers. Our work emphasizes the urgent need to implement a habitat preservation policy for these highly threatened populations. Moreover, we point out the need for a reference genome for this endemic species, to identify the genotypes associated with the most resilient phenotypes. In the case of *A. bihariensis*, there is an urgent need to bring this species to the forefront as a priority sequencing target to ensure the monitoring and conservation of the species.

Funding

This work was funded by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI–UEFISCDI, project number PN-III-P4-IDPCE-2020-1187, within PNCDI III. Lena Bonassin was funded by the Agence Nationale de la Recherche (GEODE: ANR-21-CE02-0028) and Ljudevit Luka Boštjančić was funded by the Deutsche Forschungsgemeinschaft (GEODE: DFG TH 1807/7 − 1).

Acknowledgement

The authors kindly acknowledge the support of Leonie Schardt with laboratory management and Juliane Romahn for the support with server management.

References

Allendorf, F. W., & Lundquist, L. L. (2003). Introduction: Population biology, evolution, and control of Invasive Species. *Conservation Biology*, *17*(1), 24–30. https://doi.org/10.1046/j.1523-1739.2003.02365.x

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*.

Bálint, M., Ujvárosi, L., Theissinger, K., Lehrian, S., Mészáros, N., & Pauls, S. U. (2011). The Carpathians as a Major Diversity Hotspot in Europe. In *Biodiversity Hotspots* (pp. 189–205). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-20992-5_11

Barnett, Z. C., Adams, S. B., Ochs, C. A., & Garrick, R. C. (2020). Crayfish populations genetically fragmented in streams impounded for 36–104 years. *Freshwater Biology*, *65*(4), 768–785. https://doi.org/10.1111/fwb.13466

Barrett, R. D. H., & Schluter, D. (2008). Adaptation from standing genetic variation. In *Trends in Ecology and Evolution* (Vol. 23, Issue 1, pp. 38–44). Elsevier Current Trends. https://doi.org/10.1016/j.tree.2007.09.008

Bassitta, M., Brown, R. P., Pérez-Cembranos, A., Pérez-Mellado, V., Castro, J. A., Picornell, A., & Ramon, C. (2021). Genomic signatures of drift and selection driven by predation and human pressure in an insular lizard. *Scientific Reports*, *11*(1), 1–13. https://doi.org/10.1038/s41598-021-85591-x

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. https://doi.org/10.1111/mec.12354

Çilingir, F. G., Hansen, D., Bunbury, N., Postma, E., Baxter, R., Turnbull, L., Ozgul, A., & Grossen, C. (2022). Low-coverage reduced representation sequencing reveals subtle within-island genetic structure in Aldabra giant tortoises. *Ecology and Evolution*, *12*(3), 1–13. https://doi.org/10.1002/ece3.8739

Clay, M., Brannock, P. M., Barbour, M., Feminella, J. W., Santos, S. R., & Helms, B. S. (2020). Strong Population Structure and Differentiation within and among Burrowing Bog Crayfish Species of Southern Alabama Wetlands. *Wetlands*, *40*(5), 1595–1606. https://doi.org/10.1007/s13157-020-01273-w

Do, C., Waples, R. S., Peel, D., Macbeth, G. M., Tillett, B. J., & Ovenden, J. R. (2014). NeEstimator v2: Re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Molecular Ecology Resources*, *14*(1), 209–214. https://doi.org/10.1111/1755-0998.12157

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Farhadi, A., Jeffs, A. G., & Lavery, S. D. (2022). Genome-wide SNPs in the spiny lobster Panulirus homarus reveal a hybrid origin for its subspecies. *BMC Genomics*, *23*(1), 750. https://doi.org/10.1186/s12864-022-08984-w

Farhadi, A., Pichlmueller, F., Yellapu, B., Lavery, S., & Jeffs, A. (2022). Genome-wide SNPs reveal fine-scale genetic structure in ornate spiny lobster Panulirus ornatus throughout Indo-West Pacific Ocean. *ICES Journal of Marine Science*, *79*(6), 1931–1941. https://doi.org/10.1093/icesjms/fsac130

Formenti, G., Theissinger, K., Fernandes, C., Bista, I., Bombarely, A., Bleidorn, C., Ciofi, C., Crottini, A., Godoy, J. A., Höglund, J., Malukiewicz, J., Mouton, A., Oomen, R. A., Paez, S., Palsbøll, P. J., Pampoulie, C., Ruiz-López, M. J., Svardal, H., Theofanopoulou, C., … Zammit, G. (2022). The era of reference genomes in conservation genomics. *Trends in Ecology and Evolution*, *37*(3), 197–202. https://doi.org/10.1016/j.tree.2021.11.008

Francis, R. M. (2017). pophelper: an R package and web app to analyse and visualize population structure. *Molecular Ecology Resources*, *17*(1), 27–32. https://doi.org/10.1111/1755-0998.12509

Frankham, R. (2003). Genetics and conservation biology. *Comptes Rendus - Biologies*, *326*(SUPPL. 1), 22–29. https://doi.org/10.1016/s1631-0691(03)00023-4

Frankham, R., Ballou, J. D., & Briscoe, D. A. (2010). *Introduction to Conservation Genetics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809002

Frankham, R., Bradshaw, C. J. A., & Brook, B. W. (2014). Genetics in conservation management: Revised recommendations for the 50/500 rules, Red List criteria and population viability analyses. *Biological Conservation*, *170*, 56–63. https://doi.org/https://doi.org/10.1016/j.biocon.2013.12.036

Gross, R., Lovrenčić, L., Jelić, M., Grandjean, F., Đuretanović, S., Simić, V., Burimski, O., Bonassin, L., Groza, M.-I., & Maguire, I. (2021). Genetic diversity and structure of the noble crayfish populations in the Balkan Peninsula revealed by mitochondrial and microsatellite DNA markers. *PeerJ*, *9*(August), e11838. https://doi.org/10.7717/peerj.11838

Groza, M. I., Cupea, D., Lovrenčić, L., & Maguire, I. (2021). First record of the stone crayfish in the Romanian lowlands. *Knowledge & Management of Aquatic Ecosystems*, *2020-Janua*(422), 27. https://doi.org/10.1051/KMAE/2021026

Haasl, R. J., & Payseur, B. A. (2011). Multi-locus inference of population structure: A comparison between single nucleotide polymorphisms and microsatellites. *Heredity*, *106*(1), 158–171. https://doi.org/10.1038/hdy.2010.21

Holderegger, R., Kamm, U., & Gugerli, F. (2006). Adaptive vs. neutral genetic diversity: Implications for landscape genetics. *Landscape Ecology*, *21*(6), 797–807. https://doi.org/10.1007/s10980-005-5245-9

Inkscape. (2020). *Inkscape project*.

Ion, M. C., Ács, A.-R., Laza, A. V., Lorincz, I., Livadariu, D., Lamoly, A. M., Goia, B., Togor, A., Iorgu, E. I., Ștefan, A., Popa, O. P., & Pârvulescu, L. (2024). Conservation status of the idle crayfish Austropotamobius bihariensis Pârvulescu, 2019. *Global Ecology and Conservation*, *50*, e02847. https://doi.org/10.1016/j.gecco.2024.e02847

Jamieson, I. G. (2007). Has the debate over genetics and extinction of island endemics truly been resolved? *Animal Conservation*, *10*(2), 139–144. https://doi.org/10.1111/j.1469-1795.2006.00095.x

Jelić, M., Klobučar, G. I. V., Grandjean, F., Puillandre, N., Franjević, D., Futo, M., Amouret,

J., & Maguire, I. (2016). Insights into the molecular phylogeny and historical biogeography of the white-clawed crayfish (Decapoda, Astacidae). *Molecular Phylogenetics and Evolution*, *103*, 26–40. https://doi.org/10.1016/j.ympev.2016.07.009

Jenkins, T. L., Ellis, C. D., & Stevens, J. R. (2019). SNP discovery in European lobster (Homarus gammarus) using RAD sequencing. *Conservation Genetics Resources*, *11*(3), 253–257. https://doi.org/10.1007/s12686-018-1001-8

Jussila, J., Edsman, L., Maguire, I., Diéguez-Uribeondo, J., & Theissinger, K. (2021). Money Kills Native Ecosystems: European Crayfish as an Example. *Frontiers in Ecology and Evolution*, *9*. https://doi.org/10.3389/fevo.2021.648495

Kardos, M., Armstrong, E. E., Fitzpatrick, S. W., Hauser, S., Hedrick, P. W., Miller, J. M., Tallmon, D. A., & Chris Funk, W. (2021). The crucial role of genome-wide genetic variation in conservation. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(48). https://doi.org/10.1073/pnas.2104642118

Kouba, A., Petrusek, A., & Kozák, P. (2014). Continental-wide distribution of crayfish species in Europe: update and maps. *Knowledge and Management of Aquatic Ecosystems*, *413*, 05. https://doi.org/10.1051/kmae/2014007

Kozák, P., Ďuriš, Z., Petrusek, A., Buřič, M., Horká, I., Kouba, A., Kozubíková-Balcarová, E., & Policar, T. (2015). *Crayfish Biology and Culture*. University of South Bohemia in České Budějovice, Faculty of Fisheries and Protection of Waters.

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of Population Structure using Dense Haplotype Data. *PLoS Genetics*, *8*(1), e1002453. https://doi.org/10.1371/journal.pgen.1002453

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 Evolutionary Biology*, *20*(1), 146. https://doi.org/10.1186/s12862-020-01709-1

Lovrenčić, L., Temunović, M., Bonassin, L., Grandjean, F., Austin, C. M., & Maguire, I. (2022). Climate change threatens unique genetic diversity within the Balkan biodiversity hotspot – The case of the endangered stone crayfish. *Global Ecology and Conservation*, *39*(July). https://doi.org/10.1016/j.gecco.2022.e02301

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. *Scientific Reports*, *12*(1), 2040. https://doi.org/10.1038/s41598-022-06027-8

Lynch, M., Conery, I. J., & Burger, R. (2016). *Mutation Accumulation and the Extinction of Small Populations Author (s): Michael Lynch , John Conery and Reinhard Burger Source : The American Naturalist , Vol . 146 , No . 4 (Oct ., 1995), pp . 489-518 Published by : The University of Chicago Press . 146*(4), 489–518.

Malinsky, M., Trucchi, E., Lawson, D. J., & Falush, D. (2018). RADpainter and fineRADstructure: Population Inference from RADseq Data. *Molecular Biology and Evolution*, *35*(5), 1284–1290. https://doi.org/10.1093/molbev/msy023

Manenti, R., Barzaghi, B., Nessi, A., Cioccarelli, S., Villa, M., & Ficetola, G. F. (2021). Not Only Environmental Conditions but Also Human Awareness Matters: A Successful Post-Crayfish Plague Reintroduction of the White-Clawed Crayfish (Austropotamobius pallipes) in Northern Italy. *Frontiers in Ecology and Evolution*, *9*.

https://doi.org/10.3389/fevo.2021.621613

Matočec, S. G., Bakran-Petricioli, T., Bedek, J., Bukovec, D., Buzjak, S., Franičević, M., Jalžić, B., Kerovec, M., Kletečki, E., Kralj, J., Kružić, P., Kučinić, M., Kuhta, M., Matočec, N., Ozimec, R., Rada, T., Štamol, V., Ternjej, I., & Tvrtković, N. (2002). An overview of the cave and interstitial biota of Croatia. *Natura Croatica*, *11*(SUPP), 1–102.

Meffe, G. K., & Carroll, C. (1997). *Principles of Conservation Biology* (2nd ed.). Sinauer Associates Inc.

Miller, A. D., Sweeney, O. F., Whiterod, N. S., Van Rooyen, A. R., Hammer, M., & Weeks, A. R. (2014). Critically low levels of genetic diversity in fragmented populations of the endangered Glenelg spiny freshwater crayfish Euastacus bispinosus. *Endangered Species Research*, *25*(1), 43–55. https://doi.org/10.3354/esr00609

Nichols, H. J. (2017). The causes and consequences of inbreeding avoidance and tolerance in cooperatively breeding vertebrates. *Journal of Zoology*, *303*(1), 1–14. https://doi.org/10.1111/jzo.12466

Pacioglu, O., Theissinger, K., Alexa, A., Samoilă, C., Sîrbu, O.-I., Schrimpf, A., Zubrod, J. P., Schulz, R., Pîrvu, M., Lele, S.-F., Jones, J. I., & Pârvulescu, L. (2020). Multifaceted implications of the competition between native and invasive crayfish: a glimmer of hope for the native's long-term survival. *Biological Invasions*, *22*(2), 827–842. https://doi.org/10.1007/s10530-019-02136-0

Paris, J. R., Stevens, J. R., & Catchen, J. M. (2017). Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution*, *8*(10), 1360–1373. https://doi.org/10.1111/2041-210X.12775

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

Pârvulescu, L., Iorgu, E. I., Zaharia, C., Ion, M. C., Satmari, A., Krapal, A. M., Popa, O. P., Miok, K., Petrescu, I., & Popa, L. O. (2020). The future of endangered crayfish in light of protected areas and habitat fragmentation. *Scientific Reports*, *10*(1), 1–12. https://doi.org/10.1038/s41598-020-71915-w

Pârvulescu, L., Pacioglu, O., & Hamchevici, C. (2011). The assessment of the habitat and water quality requirements of the stone crayfish (Austropotamobius torrentium) and noble crayfish (Astacus astacus) species in the rivers from the Anina Mountains (SW Romania). *Knowledge and Management of Aquatic Ecosystems*, *401*(401), 03. https://doi.org/10.1051/KMAE/2010036

Pârvulescu, L., Pérez-Moreno, J. L., Panaiotu, C., Drăguț, L., Schrimpf, A., Popovici, I. D., Zaharia, C., Weiperth, A., Gál, B., Schubart, C. D., & Bracken-Grissom, H. (2019). A journey on plate tectonics sheds light on European crayfish phylogeography. *Ecology and Evolution*, *9*(4), 1957–1971. https://doi.org/10.1002/ece3.4888

Pârvulescu, L., Schrimpf, A., Kozubíková, E., Cabanillas Resino, S., Vrålstad, T., Petrusek, A., & Schulz, R. (2012). Invasive crayfish and crayfish plague on the move: first detection of the plague agent Aphanomyces astaci in the Romanian Danube. *Diseases of Aquatic Organisms*, *98*(1), 85–94. https://doi.org/10.3354/DAO02432

Pérez-Pereira, N., Wang, J., Quesada, H., & Caballero, A. (2022). Prediction of the minimum effective size of a population viable in the long term. *Biodiversity and Conservation*,

31(11), 2763–2780. https://doi.org/10.1007/s10531-022-02456-z

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*(5). https://doi.org/10.1371/journal.pone.0037135

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Raj, A., Stephens, M., & Pritchard, J. K. (2014). FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, *197*(2), 573–589. https://doi.org/10.1534/genetics.114.164350

Razgour, O., Forester, B., Taggart, J. B., Bekaert, M., Juste, J., Ibáñez, C., Puechmaille, S. J., Novella-Fernandez, R., Alberdi, A., & Manel, S. (2019). Considering adaptive genetic variation in climate change vulnerability assessment reduces species range loss projections. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(21), 10418–10423. https://doi.org/10.1073/pnas.1820663116

Reed, D. H., & Frankham, R. (2003). Correlation between fitness and genetic diversity. *Conservation Biology*, *17*(1), 230–237. https://doi.org/10.1046/j.1523-1739.2003.01236.x

Reynolds, J., Souty-Grosset, C., & Richardson, A. (2013). Ecological roles of crayfish in freshwater and terrestrial habitats. *Freshwater Crayfish*, *19*(2), 197–218. https://doi.org/10.5869/fc.2013.v19-2.197

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*(8). https://doi.org/10.3390/genes14081627

Satmari, A., Miok, K., Ion, M. C., Zaharia, C., Schrimpf, A., & Pârvulescu, L. (2023). Headwater refuges: Flow protects Austropotamobius crayfish from Faxonius limosus invasion. *NeoBiota*, *89*, 71–94. https://doi.org/10.3897/neobiota.89.110085

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. *Conservation Genetics*, *18*(6), 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 (A stacus astacus) reveals multiple refugia. *Freshwater Biology*, *59*(4), 761–776. https://doi.org/10.1111/fwb.12302

Schulz, R., Stucki, T., & Souty-Grosset, C. (2002). Roundtable Session 4a: Management: Reintroductions and Restocking. *Bulletin Français de La Pêche et de La Pisciculture*, *367*, 917–922. https://doi.org/10.1051/kmae:2002075

Swahn, J.-Ö. (2004). The cultural history of crayfish. *Bulletin Français de La Pêche et de La Pisciculture*, *372–73*, 243–251.

Tarandek, A., Lovrenčić, L., Židak, L., Topić, M., Grbin, D., Gregov, M., Ćurko, J., Hudina,

S., & Maguire, I. (2023). Characteristics of the Stone Crayfish Population along a Disturbance Gradient—A Case Study of the Kustošak Stream, Croatia. *Diversity*, *15*(5), 591. https://doi.org/10.3390/d15050591

Theissinger, K., Edsman, L., Maguire, I., Diéguez-Uribeondo, J., & Jussila, J. (2022). Nothing can go wrong–Introduction of alien crayfish to Europe. *PLOS Water*, *1*(11), e0000062. https://doi.org/10.1371/journal.pwat.0000062

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., … Zammit, G. (2023). How genomics can help biodiversity conservation. *Trends in Genetics*, *39*(7), 545–559. https://doi.org/10.1016/j.tig.2023.01.005

Tickner, D., Opperman, J. J., Abell, R., Acreman, M., Arthington, A. H., Bunn, S. E., Cooke, S. J., Dalton, J., Darwall, W., Edwards, G., Harrison, I., Hughes, K., Jones, T., Leclère, D., Lynch, A. J., Leonard, P., McClain, M. E., Muruven, D., Olden, J. D., … Young, L. (2020). Bending the Curve of Global Freshwater Biodiversity Loss: An Emergency Recovery Plan. *BioScience*, *70*(4), 330–342. https://doi.org/10.1093/biosci/biaa002

Ungureanu, E., Mojžišová, M., Tangerman, M., Ion, M. C., Parvulescu, L., & Petrusek, A. (2020). The spatial distribution of Aphanomyces astaci genotypes across Europe: introducing the first data from Ukraine. *Freshwater Crayfish*, *25*(1), 77–87. https://doi.org/10.5869/fc.2020.v25-1.077

Weeks, A. R., Sgro, C. M., Young, A. G., Frankham, R., Mitchell, N. J., Miller, K. A., Byrne, M., Coates, D. J., Eldridge, M. D. B., Sunnucks, P., Breed, M. F., James, E. A., & Hoffmann, A. A. (2011). Assessing the benefits and risks of translocations in changing environments: a genetic perspective. *Evolutionary Applications*, *4*(6), 709–725. https://doi.org/10.1111/j.1752-4571.2011.00192.x

West, K. M., Richards, Z. T., Harvey, E. S., Susac, R., Grealy, A., & Bunce, M. (2020). Under the karst: detecting hidden subterranean assemblages using eDNA metabarcoding in the caves of Christmas Island, Australia. *Scientific Reports*, *10*(1), 21479. https://doi.org/10.1038/s41598-020-78525-6

Whiterod, N. S., Zukowski, S., Asmus, M., Gilligan, D., & Miller, A. D. (2017). Genetic analyses reveal limited dispersal and recovery potential in the large freshwater crayfish Euastacus armatus from the southern Murray-Darling Basin. *Marine and Freshwater Research*, *68*(2), 213–225. https://doi.org/10.1071/MF16006

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Willi, Y., Kristensen, T. N., Sgro, C. M., Weeks, A. R., Ørsted, M., & Hoffmann, A. A. (2022). Conservation genetics as a management tool: The five best-supported paradigms to assist the management of threatened species. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(1), 1–10. https://doi.org/10.1073/pnas.2105076119

Williams-Subiza, E. A., & Epele, L. B. (2021). Drivers of biodiversity loss in freshwater environments: A bibliometric analysis of the recent literature. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *31*(9), 2469–2480. https://doi.org/10.1002/aqc.3627

Woodruff, D. S. (2001). Populations, Species, and Conservation Genetics. In *Encyclopedia of*

Biodiversity (pp. 811–829). Elsevier. https://doi.org/10.1016/B0-12-226865-2/00355-2

Yi, S., Li, Y., Shi, L., Zhang, L., Li, Q., & Chen, J. (2018). Characterization of Population Genetic Structure of red swamp crayfish, Procambarus clarkii, in China. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-23986-z

Zimmerman, S. J., Aldridge, C. L., & Oyler-Mccance, S. J. (2020). An empirical comparison of population genetic analyses using microsatellite and SNP data for a species of conservation concern. *BMC Genomics*, *21*(1), 1–16. https://doi.org/10.1186/s12864-020-06783-9

.

General discussion

Genetic diversity is recognised as one of the fundamental levels of biodiversity (Genome 10K Community of Scientists, 2009; McMahon et al., 2014; Shaw et al., 2025). The whole-genome assessment of species allows to gain knowledge about both coding and non-coding parts of the genome, which provides information about evolutionary processes, speciation, adaptation and a species' ability to survive changes in its habitat (McMahon et al., 2014). Freshwater crayfish represent key organisms in freshwater ecosystems. However, their population numbers, especially in Europe, have been in decline over the last decades. Globally, freshwater crayfish are a highly diverse taxon and have been mostly studied regarding their physiological and ecological aspects. Genetic studies have primarily focused on single, mitochondrial markers, and/or nuclear microsatellites to understand phylogeny and population diversity (Gross et al., 2021; Lovrenčić et al., 2020; Tarandek et al., 2023), and on transcriptomics studies to investigate the immune response (Boštjančić et al., 2022), while genomic studies have been rare. Studies focusing on genomic characteristics have investigated genome size and chromosome number, but these have been conducted on a small number of species (Boštjančić et al., 2021; Mlinarec et al., 2016; Salvadori et al., 2023). The rarity of genomic studies is most likely related to freshwater crayfish being characterised by large genome sizes and high genome size variation, high chromosome numbers, high heterozygosity, and large number of repeat sequences. These characteristics represent challenges that have hindered whole-genome sequencing studies in freshwater crayfish species.

In this thesis, I aimed to address the challenging genome characterisation of freshwater crayfish and Decapoda at different levels: from repetitive elements, including TEs and satDNA, to SNPs and the whole-genome level. Furthermore, I employed novel methodological approaches to offer new insight into the evolution of freshwater crayfish. In whole genome assemblies of 20 Decapoda and 6 Crustacea species a high amount of TE was identified, and their abundance was correlated with genome size and assembly size. Based on low-coverage genome sequencing of 19 freshwater crayfish species, satDNAs were identified as a major component of the genome and the analysis revealed distinct lineage specific patterns that match the phylogeny of these taxa. Since genomic studies in crayfish rely on optimised protocols, I tested several extraction and sequencing protocols for two freshwater crayfish species with large genomes. Both the choice of DNA extraction protocols and tissue type significantly impacted the success of long read sequencing. A protocol for obtaining

high genomic sequencing yield is proposed. A population genomic study was conducted on the freshwater crayfish *A. bihariensis* using reduced representation sequencing. Low genetic diversity was recorded, and the insights gained based on SNPs provide conservation guidelines for the endangered *A. bihariensis*. Overall, I provide a methodological framework and genomic context that leverage the whole genome, including its non-coding part, that can be exploited for in-depth studies of evolution and conservation status.

6.1. Decapoda genome complexity is driven by repetitive elements

Repetitive DNA sequences are a dynamic component of genomes and are drivers of genome size variation and evolution. TEs are considered key drivers of genome expansion and rearrangement. Their insertions can cause alteration of gene expression and mutations. They influence gene expression by altering regulatory elements of genes, through epigenetic silencing and non-coding RNA transcription (Bourque et al., 2018). Their movement can lead to genomic changes that contribute to evolutionary novelty, but also to dysregulations and diseases. SatDNAs are essential for structural integrity of chromosomes, telomere protection and centromere function (Garrido-Ramos, 2017). SatDNA is involved in regulating gene expression through copy number variation and influence on local chromatin environment. High repeat content usually provides genome plasticity to the species, allowing for rapid adaptation. Furthermore, chromosomal rearrangements caused by RE can lead to reproductive isolation and speciation (Fuller et al., 2019).

A high number of RE was identified across different Decapoda and Crustacea species in Chapter II and Chapter III. The percentage of REs in the genomes and their load (copy number) is correlated with the genomes size (Chapter II) indicating that REs are key in driving genome expansion in these species. In Decapoda species total RE content average was 59.7 %, while in non-Decapoda species it was 46.4 %. The most abundant elements in Decapoda compared to other non-Decapoda are non-LTR/LINE elements (Chapter II), a pattern identified also within freshwater crayfish species (Chapter III). Larger genome sizes in Decapoda species were associated with expansion of specific classes of TE. Moreover, within freshwater crayfish species, satDNA contributes to large parts of the genome, especially within Astacidae species (Chapter III). In these species, with even larger genome sizes (12 – 18 Gb) than in other Decapoda species, the genomic content could be influenced by the amplification of satDNA families. Intraspecific and between closely relates species

genome size variation coupled with high satDNA variation has been observed in grasshopper genomes (Shah et al., 2020). Generally, except as stand-alone repeats, tandem repeats have also been found as part of TEs, especially Class II DNA transposons MITE and Helitron, which help in the spread of tandem repeats through the genome (Šatović & Plohl, 2013; Scalvenzi & Pollet, 2014). However, in the studied Crustacea and Decapoda, DNA transposons are the least represented TE class, with up to 10% of the genome, except in the Dendrobranchiata species where they contribute up to 20% of the genome (Chapter II). This shows that in different species, different elements are more prevalent. For instance, Penelope elements (PLE) are the most abundant in Astacidea species compared to other Decapoda species. However, within the Astacidea the presence of PLE differs in abundance: in the family Astacidae and Parastacidae up to 5% are PLE, while in the Cambaridae and Cambaroididae families, PLE are present only in certain species with up to 2% (Chapter III.).

The RepeatExplorer/TAREAN pipeline is designed for *de novo* repeat analysis using low-coverage whole-genome sequencing reads (Novák et al., 2020). This makes it ideal for non-model organisms that lack a repeat library or a whole genome assembly. The tool works with unassembled reads bypassing the difficulty of assembling highly repetitive regions. Furthermore, TAREAN is designed to specifically identify satellite DNA and tandem repeats, which are often overlooked by other tools (Novák et al., 2017). The downside of the tool is that the process is computationally intensive due to the all-to-all sequence comparison. RepeatModeler, used in Chapter II, creates repeat libraries from assembled genomes (Flynn et al., 2020). Because of the assembly requirement, it is not suitable for species where none is available. The program identifies new repeat sequences that are not present in public databases, though this process can be lengthy for large genomes with a high number of repeats. The species-specific library created by RepeatModeler can be used in RepeatMasker to align and annotate DNA sequences (Tempel, 2012).

The characterisation of the repeatome in Chapter II and Chapter III showed that the RE proportion and content was variable among the higher taxonomic orders, and more similar among closely related species. In both studies, based on satDNA in Chapter III and TEs in Chapter II, the repeat content showed a phylogenetic signal. Both the quality and quantity of shared REs shows higher similarity within groups, i.e. family (Chapter III) or order (Chapter II), than among groups. In both cases, there were exceptions of species that did not cluster according to the phylogeny. This could be explained by differential RE loss or expansion where some REs got removed in the genomes of certain species or there is a proliferation of

specific Res due to loss of repressive mechanisms. Horizontal transfer of TEs (HTT) could also plays a role in the differentiation of the repeatome, making the species affected by HTT more similar than their phylogenetic relatives (H.-H. Zhang et al., 2020). HTT events can be promoted by endogenous viruses (Gilbert & Feschotte, 2018). In Chapter II, integrated viruses were found in the genomes of *M. nipponense* and *H. americanus*, potentially contributing to HTT events.

In species with larger genome sizes, I did not observe an increase in the diversity of TEs (Chapter II). This was expected since other studies have shown that the diversity of TE increases only up to ~500 Mb genomes, while above that threshold there is only amplification of the TEs already present (Elliott & Gregory, 2015). Even in closely related species, the activity of TEs can largely influence the genome size. In the Isopoda species *Armadillidium nasatum* and *A. vulgare*, the genome sizes differ by ~500 Mb (1.2 Gb vs. 1.7 Gb, respectively), which is attributed to the higher transposition activity in the species with the larger genome (Becking et al., 2020). The genome size measurement of *A. bihariensis* and *A. astacus* revealed large genome sizes, 11.58 and 16.89 Gb, respectively (Chapter III). This result was within the order of magnitude expected for the Astacidae family (Boštjančić et al., 2021; Hultgren et al., 2018; Söderhäll et al., 2022). The two species share 95% of identified RE clusters, but their copy number could be different, depending on different transposition activity or amplification of satDNAs.

The species *A. bihariensis* and *A. astacus* belong to the family Astacidae which generally stands out within Decapoda for their large genome sizes. The average genome size is more than two time larger than in the Cambaridae family. Large genomes resulting from accumulation of REs, may be due to a relaxation of selective pressures. REs can provide regulatory elements and modulate gene expression. Furthermore, when there is an accumulation of RE and increase in genome size, satDNA, involved in chromosome integrity, can accumulate in heterochromatin and centromere and help stabilise the now larger chromosomes (Shah et al., 2020). Thus, satDNA can be viewed as a cause but also a consequence of large genomes. On the other hand, large genome sizes may impose costs for the organism, such as slower cell division, increased energy consumption and slower growth and development limiting the overall fitness (Lertzman-Lepofsky et al., 2019). Genome size in Crustaceans has in general been shown to correlate with life cycle and habitat (i.e., terrestrial, marine or freshwater) but does not follow phylogeny (Alfsnes et al., 2017). Freshwater crayfish are also characterised by high chromosome numbers ranging from 2n

=102 to 2n= 276 (Boštjančić et al., 2021), however, there is no correlation observed between genome size and chromosome number (Jeffery, 2015).

Chromosome numbers are altered by polyploidisation, translocation, sequence amplification, fusions, and fissions (Schubert, 2007). Polyploidisation events include the duplication of the entire set of chromosomes because of errors during meiosis or mitosis (Mayrose & Lysak, 2021). Polyploidy is common in plants, but rare in animals. Fusions and fissions of chromosomes alter the chromosome number while preserving the total genome size. These events are considered the primary reason for the variation of chromosome numbers between closely related species (Mackintosh et al., 2023). Amplification of specific sequences can increase the chromosome size and lead to a formation of small chromosomes which increases the chromosome number and the genome size (Li et al., 2017). Translocations involve the exchange of genetic material between chromosomes and can lead to a gain or loss of entire chromosomal segments and alter the chromosome number (Canoy et al., 2022). The chromosomal rearrangements are promoted by satDNA and TE further contributing to evolutionary change (Louzada et al., 2020).

Between the different Crustacea species, the TE sequences showed high divergence values, indicating ancient expansion events (Chapter II). In contrast, satDNA sequences within freshwater crayfish species showed low divergence and recent expansion events (Chapter III). The sequence divergence landscape of TE and satDNA in *P. clarkii* shows a similar pattern, indicating the expansion of RE copies could happen simultaneously. *Cherax destructor* and *C. quadricarinatus* divergence landscapes show low divergence of satDNA sequences (Chapter III), and higher divergences for TE sequences (Chapter II). The low divergence of satDNA sequences in these two *Cherax* species is following the pattern of concerted evolution. This mechanism corrects and homogenises the satDNA sequences within the genome, preventing the accumulation of mutations. Unlike satDNA, TEs do not undergo concerted evolution, but after insertion TEs accumulate mutations independently of other copies. The divergence patterns are a result of different rates and evolution mechanisms of these two types of repetitive DNA.

The insertion of TE in specific genomic regions can affect gene expression (Schrader & Schmitz, 2019). Different TE families have different insertion sites preferences which enables the existence of multiple families in the genome (Wells & Feschotte, 2020). Diverse DNA transposons target 5’ upstream gene regions which allows gene expression modulation for the host and is benefiting the “survival” of the TE (Spradling et al., 2011). Autonomous

TE, which include LINEs and LTR elements, that carry their own promoters have greater likelihood of disrupting the expression of the genes. Therefore, LTR elements are usually not found in gene rich regions, but in regions with low recombination rates, such as peri-centromeric chromatin (Wells & Feschotte, 2020). TEs are often inserted into regions of the genome where they can leave a trace, making them suitable for studying macroevolutionary timescales (Bourque et al., 2018). SatDNA, because of their quick evolution, can be used as species specific markers. However, certain satDNA can be found in multiple species and families where they are conserved for millions of years (dos Santos et al., 2021; Petraccioli et al., 2015). Such sequence was also found in the freshwater crayfish species (Chapter III). The PlSAT3-411 was identified as conserved across all freshwater crayfish families, making it one of the most ancient satDNA discovered so far. The conservation might reflect the functional role of the satDNA sequence in the pericentromeric region.

Genomic changes between closely related species and among populations can be detected with satDNA and TE (Bourgeois & Boissinot, 2019), however, more suitable are SNPs which derive from point mutations and errors during DNA replication. The high number of repeats can also lead to a high number of SNPs. In Chapter V, in total ~1.3 million SNPs were identified for the species *A. bihariensis*. In studies on *Panulirus homarus* and *Penaeus monodon* the total number of identified SNPs was only around 100 k (Farhadi et al., 2022; Vu et al., 2021). The discrepancy is due to *A. bihariensis* having a five to ten fold larger genome size than *P. homarus* and *P. monodon* (11.58 Gb compared to 1.3 Gb and 2.2 Gb, respectively), as in general, larger genomes have more SNPs (Montanari et al., 2023). Most of the SNPs found within a genome do not have a significant impact on the phenotype or on the evolution of the individual. SNPs that occur in genes or regulatory regions can have a direct effect on the gene expression or alter protein functions. For the adaptive evolution of the species, SNPs that are found in gene coding regions are more useful than the ones in repetitive sequences (Sudan et al., 2019). The functional characterisation of SNPs was beyond the scope of the study in Chapter V, however, I expect that a lot of the SNPs come from repetitive DNA sequences and thus do not carry adaptive potential, while still reflecting the population structure.

While SNPs are key to adaptive evolution, the high number of RE, particularly satDNA, also present a unique potential for adaptation to environmental pressures. The freshwater crayfish species studied in Chapter III exhibited a high number of satDNA families, the highest proportion were satellites below 100 bp in length. This has already been observed in other

Decapoda species, such as *P. leptodactylus*, *Litopenaeus vannamei* and several *Macrobrachium* species (Boštjančić et al., 2021; Molina et al., 2020; X. Zhang et al., 2019). In *L. vannamei* simple repeats accounted for 23.93% of the genome, constituting the highest proportion among animal genomes, and were mostly found in intergenic regions and introns of protein coding genes (X. Zhang et al., 2019). In *Macrobrachium* species, simple repeats were localised on chromosome arms, but were absent from centromeric AT-rich regions (Molina et al., 2020). In general, the simple repeats found among introns may regulate gene expression, while repeats found within TEs could contribute to DNA recombination with TEs (X. Zhang et al., 2019). The localisation and high abundance of short satDNA sequences, could create a genetic architecture that allows adaptation to environmental changes or stressors. It has been shown that in plant species, the length of simple repeats can change the gene expression, thus improving the organism's ability to endure temperature variations (Reinar et al., 2021).

Moreover, simple repeat sequence expansion can act as templates to small RNAs which are involved in epigenetic silencing (Sureshkumar et al., 2025). In the shrimp *P. vannamei* it has been shown that small RNAs play a complex role in the immune response during a viral infection, modulating both viral and host gene expression (Luangtrakul et al., 2025). The high amount of simple repeats identified in crayfish in this thesis, could thus play an important role in the crayfish immune response. The high turnover of satDNA leads to rapid divergence between closely related species which can create reproductive barriers (Ferree & Prasad, 2012). SatDNA sequences are often transcribed into non-coding RNAs that are involved in epigenetic regulation and heterochromatin formation, allowing for rapid adaptive changes without the alteration of gene sequences (Fonseca-Carvalho et al., 2024). Furthermore, mutations in the centromeric satDNA can disrupt the compatibility of centromeres and associated proteins leading to speciation (Melters et al., 2013). Ultimately, the RE in Decapoda evolution are key drivers of genome expansion and provide a source of genomic plasticity. However, their specific function and dynamics, specifically the interplay between TE and satDNA and their potential adaptive function, remain important avenues for future research.

6.2. Optimised methodologies are crucial for studying freshwater crayfish genomes

Given the genomic characteristics of Decapoda, i.e., large genomes and high repetitive DNA content, the commonly used genomic methodologies are often not efficient. Especially in whole genome sequencing of such complex genomes, the bioinformatic tools used for genome assembly do not perform as expected leading to incomplete and fragmented assemblies. In addition to the challenges in the bioinformatic processing of the data generated by sequencing methods, difficulties can arise in the sequencing process and protocols prior to sequencing. Many studies working with non-model organisms and invertebrate species have shown the need for optimisation of protocols used for long read sequencing (Howard et al., 2025). These protocols tend to be standardised but often there are species-specific adaptations needed to overcome biological challenges such as tough exoskeleton, high contaminant content or low tissue availability. In *Daphnia spp*., DNA extraction protocols were modified on the step of mechanical disruption of tissue. The chitin released from the carapace was shown to overestimate DNA concentration during measurements, therefore a proteinase K digestion step was performed to digest the tissue without disrupting the carapace (Athanasio et al., 2016). In the shrimp *Penaeus monodon* the CTAB method, usually efficient in removing proteins and neutral polysaccharides (S. C. Tan & Yiap, 2009), resulted in low DNA purity due to ineffective removal of organic contaminants such as acidic polysaccharides. Therefore, a commercial kit was used for HMW DNA extraction (Angthong et al., 2020). In contrast, species specific modifications of protocols are often not needed for vertebrate genomes studies. Vertebrate genomes also tend to have fewer repetitive sequences and lower polyploidy rates. They are therefore comparatively easier to sequence and assemble and are the focus of extensive research leading to many available resources (Rhie et al., 2021). The laboratory protocols are well established for vertebrate tissue and the high demand for sequencing vertebrate genomes has decreased the costs of the service. This made vertebrate genome sequencing much more affordable and accessible as compared to invertebrates. The lack of standardised protocols for non-model invertebrate organisms highlights the critical need for methodological development.

In Chapter IV, I therefore focused on the optimisation of DNA extraction and library preparation protocols for long read sequencing technologies. In Chapter IV a combination of different extraction methods, tissue types and sequencing technologies was tested for long read sequencing of two freshwater crayfish species belonging to the family Astacidae, *A.*

astacus and *A. bihariensis* with large genomes of 11.58 Gb and 16.89 Gb, respectively. For these species, the salting-out DNA extraction protocol coupled with sorbitol wash purification and the PacBio amplification-based library preparation strategy were identified as highly suitable for obtaining large amounts of data needed for the genome assembly of the two species. Moreover, the different DNA extraction protocols showed different performance based on the tissue type used, identifying muscle from abdomen and pereopods as the tissue most appropriate for DNA extraction because of high yield and purity of DNA. The muscle tissue has high cell density (Fukuzawa, 2001), is uniform and free of complex cell types and metabolites found in other tissues, such as the often-used hepatopancreas or haemolymph. Here, a non-kit-based DNA extraction protocol is favoured, while commercial kits showed lower performance (Chapter IV). Commercial kits are designed for a broad range of sample types and are not optimised for tissue with unique challenges. A non-kit-based protocol allows the adjustment of every step and thus outperform commercial kits. In cases were standardisation and automation of protocols or high throughput is needed, commercial kits could be more appropriate because of the higher reproducibility (Wallinger et al., 2017), however, higher per-sample costs that are usually associated with kit-based protocols need to be considered.

In addition to high DNA yield, high DNA quality is crucial for successful sequencing performance (Dahn et al., 2022; Howard et al., 2025). The hepatopancreas, a major organ involved in metabolism contains enzymes, specifically nucleases, that can degrade DNA (McGaw & Curtis, 2024). Haemolymph, a common source of DNA in other organisms, has lower cells concentration and contains proteins that cause coagulation (Gianazza et al., 2021), which interfere with the DNA extraction and result in low DNA yield. In Crustacea tissues the most common contaminants are polyphenols and polysaccharides present in the exoskeleton. Polysaccharides are often co-precipitated with the DNA and can inhibit downstream enzymatic reactions (Fang et al., 1992). Polyphenols are a key component of the crustacean immune system (Hong et al., 2024). When the tissue is damaged, polyphenols are oxidised into quinones that irreversibly bind to the DNA. These DNA-bound metabolites could be then inhibiting the PacBio sequencing polymerase and blocking the passage of the molecule through the Nanopore. Such issues have been noted in the brown algae where the DNA bound quinones inhibited Nanopore sequencing with available pores declining rapidly (Pearman et al., 2024).

In Chapter IV, PacBio sequencing methodology outperformed ONT Nanopore for generating sequencing data because of higher data yield. On average, PacBio amplification-free protocol yielded 3.35 Gb per run, while Nanopore yielded 1.08 Gb per run. ONT Nanopore is based on a single molecule of DNA passing through a protein pore. If there are contaminants present in the sample, or damages to the DNA molecule, the passage of the DNA will be blocked causing the pore to become inactive and the data will not be usable (Pearman et al., 2024). PacBio sequencing is less affected by inhibitors compared to ONT Nanopore. The library preparation in PacBio sequencing has multiple steps effective at cleaning up the sample and creating a library of high integrity (PacBio, 2025). ONT Nanopore libraries are prepared in a smaller number of steps with fewer clean-up steps to maintain the length of the DNA molecule. Therefore, it is also affected more by short fragments potentially present in the sample (Lopez et al., 2019).

In Chapter IV, the amplification-based PacBio approach was selected over the amplification-free approach to generate sequencing data. The PacBio amplification-based approach yielded 27.9 Gb per run. In cases were DNA-bound metabolites are inhibiting the sequencing polymerase, amplification-based library preparation protocols can improve sequencing (Männer et al., 2024). PCR amplification of DNA molecules is carried out using polymerases with high fidelity, processivity and reduce bias (Bein et al., 2025). This creates copies of molecules that are free of contaminants and are further copied. The sequencing polymerase, on the other hand, is synthesising a new DNA strand based on the SMRTbell circular template (Travers et al., 2010). The sequencing polymerase is meant to synthesise a long DNA strand without detaching but it is highly sensitive to DNA-bound molecules which cause the polymerase to stop, leading to a failed or incomplete reads (Korlach et al., 2010; PacBio, 2018).

In the case of small Isopoda species, where tissue is limited due to small body size, the species have large genome size (average 7.59 Gb (Jeffery, 2015)), and the DNA sequences poorly, reaching sufficient coverage is challenging. Therefore, the challenge is addressed by combining amplification free and amplification based PacBio library preparation approaches (Howard et al., 2025). In Chapter IV, amplification-based PacBio library preparation approach showed higher sequencing yield and longer reads than the amplification-free approach. While the general best practice is to avoid PCR amplification steps due to potential bias towards specific genomic regions and PCR errors (Howard et al., 2025; Männer et al., 2024), this bias can be mitigated. For example, using different PCR polymerases that amplify

different genomic regions and then combining the data can improve the quality of genome assemblies (Bein et al., 2025; Männer et al., 2024). Nonetheless, the combination of amplification-based and amplification-free approaches is currently necessary for crayfish genome sequencing. Overall, we successfully generated substantial genomic data, 640.26 Gb for *A. astacus* and 243.15 Gb for *A. bihariensis*, sufficient for *de novo* genome assemblies, corresponding to 38x and 21x coverage, respectively.

Improving genome contiguity is crucial for accurate annotation and downstream genomic analysis such as variant detection, analysis of repetitive regions or comparative genomics studies (Grau et al., 2018). For genes, which are often large and contain multiple exons and introns, a fragmented assembly presents a challenge. These genes can be split across multiple contigs making it difficult to annotate (Kim et al., 2022). Furthermore, regulatory elements can be distant from the gene itself, and in a fragmented assembly located on a different contig. This is a challenge for studies on gene expression and regulation patterns. In RNAseq based gene expression studies, RNA transcripts are usually mapped to a reference genome if available. When the genome is incomplete or fragmented, the reads may be mapped in erroneous locations or not mapped at all, which distorts the gene expression data (Chen et al., 2023).

Similarly, in fragmented genome assemblies, the repeat annotation is often incomplete or missing. The workflow proposed in Chapter II allowed the identification of around 10% more REs compared to standard approaches. This workflow uses a two-step approach in which repeats are first identified prior to assembly and genome annotation and secondly, the obtained library of identified repeats is combined with public repeat databases to improve annotation. This approach allows for better identification of repeats and satDNA, especially in species that are not extensively represented in public databases. The overall improvement in annotation among the studied species ranged from 1 to 20% more identified repeats. This indicates that the repeat annotation could not be significantly improved for certain genomes, likely due to excessive fragmentation or the need for additional manual curation. The *de novo* identification of repeats, in particular satDNA, is extremely important in Decapoda, as satDNA constitutes a large proportion of the genome. *De novo* identification of repeats from low coverage sequencing applied in Chapter III, allowed the characterisation of the repeatome of 19 species without a genome assembly. The use of Repeat Explorer and TAREAN pipeline allowed the identification of satDNAs, which usually remain unidentified by other repeat tools. This demonstrates the value of a whole genome-free approach for

species lacking a complete genome assembly, offering a foundation for future genomic studies.

In addition to gene and repeat annotation, the quality of a genome assembly is also critical for accurate identification of SNPs (Florea et al., 2011). A fragmented assembly produces partial genes or predicts a smaller number of genes than there are present preventing the complete identification of SNPs in gene regions. Furthermore, mis-assembled genomes can have misconstructed structural variants which leads to erroneous SNP calling (Hurgobin & Edwards, 2017). SNPs identified from a whole genome assembly offer comprehensive coverage across the entire genome, including gene regions and non-coding regions. Among the alternatives used for SNP identification without a reference genome is also double-digest restriction-site associated DNA sequencing (ddRADseq). Considering the challenges in long-read genome sequencing and the lack of a reference genome, in Chapter V, a ddRADseq approach was applied to populations of *A. bihariensis*, allowing for the identification of SNPs from a reduced portion of the genome. ddRADseq is useful because it can be used without a reference genome and significantly reduces costs compared to WGS, especially for species with large genomes. Moreover, the quality of DNA can affect SNP identification (Montanari et al., 2023), therefore, to ensure the highest SNP data quality, the DNA extraction protocol developed in Chapter IV was utilised for all DNA extractions in Chapter V.

Overcoming the limitations of studying freshwater crayfish genomes requires the application of novel and/or alternative methodologies. The work detailed across Chapters II - V addresses the technical and biological challenges inherent in studying these complex genomes. By applying specific techniques, such as modified DNA extraction protocols and a novel repeat annotation workflow, my research provides effective alternatives for efficient genomic data generation. Using specialised methods, such as ddRADseq, enables a wide range of downstream analyses, such as population structure and phylogenetic analyses, or phenotype-genotype association studies. These findings underscore the need for continued methodological development to advance the understanding of crayfish and crustacean genomics.

Genomic projects for crayfish and other organisms with large genomes require specialised sequencing technology and substantial amount of data. While long read sequencing is becoming more accessible and the costs are decreasing, the large amount of data needed to achieve sufficient coverage for downstream applications still makes these projects financially prohibitive for many researchers. Furthermore, technical and biological failures are seldom

reported in scientific literature. Even when considerable time and resources have been spent on troubleshooting issues with DNA extraction, library preparation or bioinformatic pipelines, these difficulties are not well documented. Future advances in crustacean and other non-model organisms genomics therefore depend not only on the development of improved methodologies but also on reporting these challenges. With the advancement of genomic initiatives such as the Darwin Tree of Life Project (DToL), the Vertebrate Genome Project (VGP), and the European Reference Genome Atlas (ERGA) initiative, these challenges are being reported more frequently (e.g., Howard et al., 2025), and reviews on key difficulties are now being published (e.g., Reichel et al., 2025). In this regard, the comprehensive and systematic comparison of different protocols provided in this thesis, will greatly facilitate future genomic studies on crayfish genomes.

6.3. Genomic approaches for informing conservation strategies of endangered freshwater crayfish

Genetic diversity is critical to individual and population fitness, enables a species to respond to environmental changes and determines their evolutionary potential for long-term survival (Shaw et al., 2025). Globally, there is a loss of intra-specific genetic diversity over time, and once lost this diversity is difficult to restore, reflecting in permanent loss in species adaptability and ecosystem resilience. Conservation actions that include carefully planned translocation for population reinforcement have been proven to increase genetic diversity within populations, while actions on improving environmental conditions and increase population growth, may reduce the decline of genetic diversity (Shaw et al., 2025). Both genetic and genomic data informs on population structure, effective population size, inbreeding, isolation and fragmentation of the populations (Höglund, 2009; McMahon et al., 2014). However, the use of a high number of markers across the genome can provide more accurate estimates than genetic datasets based on few markers (Supple & Shapiro, 2018). In the case of the idle crayfish *A. bihariensis*, genetic data based on five microsatellite loci revealed strong genetic structuring between populations (Pârvulescu et al., 2020), however, the genomic SNP dataset showed additional structuring within populations (Chapter V). The latter can help identifying unique genetic patterns, population history and gene flow. Furthermore, the low heterozygosity identified in the populations indicates the reduced ability for adaptation, and therefore an increased vulnerability of the populations.

The traditionally used genetic markers in population genetic studies are microsatellites which are short repetitive sequences with high mutation rate. Because of their variability they can provide enough information for genetically distinguishing populations and individuals (Putman & Carbone, 2014). However, their limited number usually used in population genetic studies provides a low-resolution view on the genome. SNPs mutate at a lower rate than microsatellites but are more abundant throughout the genome. A high abundance of analysed SNPs provides a much higher resolution and can detect more subtle differences between populations. The more comprehensive genetic evaluation of populations provided by SNPs is crucial for conservation efforts (Zimmerman et al., 2020). The SNP data obtained in Chapter V helped in the conservation status assessment of the *A. bihariensis*, which was classified as endangered (Pârvulescu et al., 2025). This study identified populations with unique genetic diversity that were defined by (Ács et al., 2025) to meet the criteria to be

considered ark sites, where the individuals are protected from threats. Moreover, it was shown that the gene flow between populations is highly restricted and structured by river basins (Chapter V). This indicates there is reduction in genetic exchange which poses a significant threat for the species genetic health.

Austropotamobius bihariensis is an endemic species, with small and isolated populations (Pârvulescu et al., 2020). These populations are characterised by deficiency of heterozygotes, lower genetic variation and have a higher likelihood of loss of rare alleles and low migration rates. These factors influence small effective population sizes (N_e). N_e is defined as the number of individuals that participates in producing the next generation and in an idealised population would maintain the same level of genetic diversity as the real population (Frankham, 2019). N_e is identified as a key metric for monitoring within population genetic diversity and provides information about evolutionary processes by quantifying the random effects of genetic drift (Mastretta-Yanes et al., 2024; Waples, 2025). N_e identified in *A. bihariensis* populations was below 100 (Chapter V) and similarly low values were identified in the endangered Nashville crayfish *Faxonius shoupi* (Hurt et al., 2022). It has been found that N_e is negatively correlated with genome size and TE expansion in fruit flies, isopods and killifishes indicating the role of genetic drift in determining recent differences in genome size (Cui et al., 2019; Lefébure et al., 2017; Mérel et al., 2025). However, this effect is not seen at larger taxonomic scales (Marino et al., 2025). Lineages with low N_e can accumulate TEs which have slight deleterious insertions (Lynch, 2007). Consequently, because of strong drift effects in populations with small N_e (Charlesworth et al., 2003), these insertions have a higher chance to fixate as neutral alleles and cause larger genome size (Lynch, 2007; Marino et al., 2025). High TE content and large genome sizes observed across Decapoda species (Chapter II and Chapter III) could relate to small N_e, especially considering the large variation in genome size among closely related species. Still research about N_e in Decapoda is missing to test the correlation between N_e, TE, and genome size.

While the theoretical understanding of how a low N_e can impact genetic health and potentially lead to larger genome sizes is advancing, translating this knowledge into conservation actions requires high resolution data. Whole genome SNPs are often used in conservation studies. The information gained from SNPs can be used in delineating conservation units and identify gene flow among populations (Supple & Shapiro, 2018). For instance, in Chapter V, genetic similarity was identified among populations of *A. bihariensis* belonging to different rivers. This indicates the possibility of human mediated translocation.

In such case, the conservation actions applied would be different than to other populations where there is no human influence. Conservation actions focused on genetically distinct populations usually aim to protect the unique genetic diversity by maintaining natural barriers between the populations, while at the same time preventing inbreeding by maintaining large population size (Hoban et al., 2023). In cases of translocated populations, conservation efforts focus on reducing the risk of outbreeding depression, where mixing genetically distinct populations can lead to offspring with lower fitness (Reid et al., 2025). The translocation brings further risks of transmitting diseases between populations, making preventing disease spread a foremost management action (Warne & Chaber, 2023).

SNPs used in genome-wide association studies (GWAS) can link genotypes to specific phenotypes and are commonly applied in disease-resistance studies (Zimmerman et al., 2020). European freshwater crayfish species are threatened by the crayfish plague disease, however, some populations were shown to be resistant to the disease (Jussila, Francesconi, et al., 2021; Maguire et al., 2016). Future studies could investigate the genetic variants underlying the resistance which could be informative for conservation. Even with the advantages of whole-genome sequencing, SNPs are widespread tools in genetic research. Freshwater crayfish would benefit greatly from more extensive genomic research to better understand their biology, their genome architecture and evolution and apply this knowledge to conservation efforts.

General conclusions

In this thesis I characterised the extraordinary genomes of freshwater crayfish and other Decapoda across complementary levels of organisation in order to clarify how genome characteristics intersect with evolutionary potential and conservation needs. Across whole genome assemblies and low coverage datasets in freshwater crayfish and Decapoda, REs emerged as principal drivers of genome size and structure revealing a phylogenetic signal. Comparative analyses showed Class I non-LTR/LINE elements as particularly abundant, while DNA transposons were less represented. Within freshwater crayfish satDNA was particularly abundant in the Astacidae family, characterised by extremely large genomes. Repeatome clustering mirrors the taxonomical relatedness of species, while a few outliers are present due to differential amplification, loss or horizontal transfer. Divergence landscapes show older TE bursts and recent satellite expansions. Within crayfish, the recent satDNA burst is in line with concerted evolution, and larger genomes containing more copies of common sequence families, rather than more families. TEs and satDNA, therefore, contribute to genome size expansion. At the population level, ddRADseq of *A. bihariensis* populations revealed low diversity, strong spatial sstructuring,and small effective population sizes, providing insights for conservation and management actions. Small N_e increases the fixation probability of deleterious insertion and can thereby facilitate repeat accumulation, while low genetic diversity indicates reduced adaptive capacity. Therefore, the RE dynamics determine the adaptive potential and conservation risk in crayfish.

Methodological advances were essential for observing genomic patterns with a two-step *de novo* repeat identification pipeline improving repeat detection and annotation compared to standard approaches. Furthermore, an optimised DNA extraction protocol enabled me to generate sequencing data from tissue prone to contaminants. The DNA extraction protocol was used prior to both long-read sequencing and ddRAD library preparation. With these methods I provide a workflow for future comparative and conservation genomic studies in crayfish. With optimised protocols, generating high contiguity assemblies is more accessible. Coupling whole genome data with chromosome conformation capture methods, like Hi-C, will help to resolve more complex genomic regions, especially REs. Method development is crucial for the advancement of successful genomic studies, and further optimisations should be done for long read sequencing library preparation without PCR amplification. Transparent

reporting of wet-lab and bioinformatic limitations are key for accelerating standardisation across studies. Even though the assembly of certain Decapoda genomes remains challenging, the constant development of new technologies and bioinformatic tools makes genomic data more accessible and complete. As more complete reference genomes are produced, genomic characterisation should be complemented by functional studies that examine the roles of satDNA and TEs, as well as small RNAs transcribed from repeats. Genome-wide SNPs should be complemented with structural variant analysis and functional characterisation to inform management actions, including phenotypic resilience and disease resistance. The large genome size and RE content should be further explored in relation to the large chromosome number and chromosome number variation between the species. Ultimately, a comprehensive understanding of Decapoda genomics will require multiple approaches, integrating genomic, structural, and functional analyses complemented with ecological data to provide insights crucial for their conservation and management.

Conclusion générale

Dans cette thèse, j'ai caractérisé les génomes exceptionnels des écrevisses d'eau douce et d'autres décapodes à différents niveaux d'organisation complémentaires afin de comprendre comment les caractéristiques génomiques interagissent avec le potentiel évolutif et les besoins de conservation. Grâce à l'analyse d'assemblages de génomes complets et de jeux de données à faible couverture chez les écrevisses d'eau douce et les décapodes, les REs sont apparus comme les principaux déterminants de la taille et de l'organisation du génome, révélant un signal phylogénétique. Les analyses comparatives ont montré que les éléments de classe I non LTR/LINE étaient particulièrement abondants, tandis que les transposons ADN étaient moins représentés. Chez les écrevisses d'eau douce, l'ADN satellite était particulièrement abondant dans la famille des Astacidae, caractérisée par des génomes extrêmement grands. Le clutering des éléments répétés reflète la proximité taxonomique des espèces, bien que quelques valeurs aberrantes soient présentes en raison d'une amplification différentielle, d'une perte ou d'un transfert horizontal. Les profils de divergence des REs montrent des d'anciens pics d'expansion de TEs et des expansions plus récentes d'ADN satellite. Chez les écrevisses, l'expansion récente de l'ADN satellite est conforme à une évolution concertée, et les génomes plus grands contiennent davantage de copies de familles de séquences communes, plutôt que davantage de familles. Les TEs et l'ADN satellite contribuent donc à l'expansion de la taille du génome. Au niveau des populations, le séquençage ddRAD des populations d'*A. bihariensis* a révélé une faible diversité, une forte structuration spatiale et une petite taille effective des populations, fournissant ainsi des informations utiles pour les mesures de conservation et de gestion. Une petite N_e augmente la probabilité de fixation d'insertions délétères et peut ainsi faciliter l'accumulation d'élements répétés, tandis qu'une faible diversité génétique indique une capacité d'adaptation réduite. Par conséquent, la dynamique des REs conditionne le potentiel adaptatif et le niveau de risque pour la conservation chez les écrevisses.

Les avancées méthodologiques ont été essentielles pour observer les motifs génomiques grâce à un pipeline d'identification *de novo* en deux étapes des éléments répétés, qui a amélioré la détection et l'annotation des éléments répétés par rapport aux approches standard. De plus, un protocole d'extraction d'ADN optimisé m'a permis de générer des données de séquençage à partir de tissus particulièrement riches en contaminants. Le protocole

d'extraction d'ADN a été utilisé à la fois pour les extractions d'ADN destinées au séquençage à lecture longue et pour la préparation de la librairie ddRAD. Grâce à ces méthodes, je propose un pipeline pour de futures études comparatives et de génomique de conservation chez les écrevisses. Avec des protocoles optimisés, la génération d'assemblages à haute contiguïté est plus accessible. Le couplage de données de génome complet avec des méthodes de capture de la conformation chromosomique, comme l'approche Hi-C, aidera à résoudre les régions génomiques complexes, en particulier les REs. Le développement de méthodes est essentiel à la réussite des études génomiques, et des optimisations supplémentaires devraient être apportées à la préparation de librairies de séquençage à lecture longue sans amplification par PCR. Une communication transparente sur les limites expérimentales et bio-informatiques est essentielle pour accélérer la standardisation entre les études. Même si l'assemblage de certains génomes de décapodes reste difficile, le développement continu de nouvelles technologies et d'outils bio-informatiques rend les données génomiques plus accessibles et plus complètes. À mesure que des génomes de référence plus complets sont générés, la caractérisation génomique devrait être complétée par des études fonctionnelles pour déterminer les rôles de l'ADNsat et des TEs, ainsi que des petits ARN transcrits à partir d'éléments répétés. L'étude des SNPs à l'échelle du génome devrait être complétée par une analyse des variants structuraux et une caractérisation fonctionnelle afin d'orienter les mesures de gestion, notamment en matière de résilience phénotypique et de résistance aux maladies. La grande taille du génome et la teneur en REs devraient être explorées plus en détail en relation avec le nombre élevé de chromosomes et la variation du nombre de chromosomes entre les espèces. En définitive, une compréhension approfondie de la génomique des décapodes nécessitera la mise en œuvre d'approches multiples, intégrant des analyses génomiques, structurelles et fonctionnelles enrichies par des données écologiques, afin de fournir des informations cruciales pour la conservation et la gestion de ces espèces.

References

Aasegg Araya, R., Reinar, W. B., Tørresen, O. K., Goubert, C., Daughton, T. J., Hoff, S. N. K., Baalsrud, H. T., Brieuc, M. S. O., Komisarczuk, A. Z., Jentoft, S., Cerca, J., & Jakobsen, K. S. (2025). Chromosomal Inversions Mediated by Tandem Insertions of Transposable Elements. *Genome Biology and Evolution*, *17*(8). https://doi.org/10.1093/gbe/evaf131

Abdelrahman, H., ElHady, M., Alcivar-Warren, A., Allen, S., Al-Tobasei, R., Bao, L., Beck, B., Blackburn, H., Bosworth, B., Buchanan, J., Chappell, J., Daniels, W., Dong, S., Dunham, R., Durland, E., Elaswad, A., Gomez-Chiarri, M., Gosh, K., Guo, X., … Zhou, T. (2017). Aquaculture genomics, genetics and breeding in the United States: current status, challenges, and priorities for future research. *BMC Genomics*, *18*(1), 191. https://doi.org/10.1186/s12864-017-3557-1

Abdul-Muneer, P. M. (2014). Application of Microsatellite Markers in Conservation Genetics and Fisheries Management: Recent Advances in Population Structure Analysis and Conservation Strategies. *Genetics Research International*, *2014*, 1–11. https://doi.org/10.1155/2014/691759

Ács, A. R., Ion, M. C., Miok, K., Laza, A. V., Pitic, A., Robnik-Šikonja, M., & Pârvulescu, L. (2025). Threats Assessment of the Endemic Idle Crayfish (Austropotamobius bihariensis Pârvulescu, 2019): Lessons From Long-Term Monitoring. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *35*(1). https://doi.org/10.1002/aqc.70033

Albertson, L. K., & Daniels, M. D. (2018). Crayfish ecosystem engineering effects on riverbed disturbance and topography are mediated by size and behavior. *Freshwater Science*, *37*(4), 836–844. https://doi.org/10.1086/700884

Alderman, D. J. (1996). Geographical spread of bacterial and fungal diseases of crustaceans. *Rev. Sci. Tech. Off. Int. Epiz*, *15*(2).

Alfsnes, K., Leinaas, H. P., & Hessen, D. O. (2017). Genome size in arthropods; different roles of phylogeny, habitat and life history in insects and crustaceans. *Ecology and Evolution*, *7*(15), 5939–5947. https://doi.org/10.1002/ece3.3163

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*, 1–18. https://doi.org/10.7717/peerj.10340

Athanasio, C. G., Chipman, J. K., Viant, M. R., & Mirbahai, L. (2016). Optimisation of DNA extraction from the crustacean Daphnia. *PeerJ*, *2016*(5). https://doi.org/10.7717/peerj.2004

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. *Frontiers in Genetics*, *12*(January), 1–8. https://doi.org/10.3389/fgene.2021.695763

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE*, *3*(10), e3376. https://doi.org/10.1371/journal.pone.0003376

Balachandra, S., Sarkar, S., & Amodeo, A. A. (2022). The Nuclear-to-Cytoplasmic Ratio: Coupling DNA Content to Cell Size, Cell Cycle, and Biosynthetic Capacity. *Annual Review of Genetics*, *56*(1), 165–185. https://doi.org/10.1146/annurev-genet-080320-030537

Becking, T., Gilbert, C., & Cordaux, R. (2020). Impact of transposable elements on genome size variation between two closely related crustacean species. *Analytical Biochemistry*, *600*(May), 113770. https://doi.org/10.1016/j.ab.2020.113770

Bein, B., Chrysostomakis, I., Arantes, L. S., Brown, T., Gerheim, C., Schell, T., Schneider, C., Leushkin, E., Chen, Z., Sigwart, J., Gonzalez, V., Wong, N. L. W. S., Santos, F. R., Blom, M. P. K., Mayer, F., Mazzoni, C. J., Böhne, A., Winkler, S., Greve, C., & Hiller, M. (2025). Long-read sequencing and genome assembly of natural history collection samples and challenging specimens. *Genome Biology*, *26*(1), 25. https://doi.org/10.1186/s13059-025-03487-9

Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., & Dekker, J. (2012). Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods*, *58*(3), 268–276. https://doi.org/10.1016/j.ymeth.2012.05.001

Berger, C., Štambuk, A., Maguire, I., Weiss, S., & Füreder, L. (2018). Integrating genetics and morphometrics in species conservation—A case study on the stone crayfish, Austropotamobius torrentium. *Limnologica*, *69*(July), 28–38. https://doi.org/10.1016/j.limno.2017.11.002

Bista, I., & Lino, A. (2025). *Long-read sequencing for biodiversity analyses - a comprehensive guide*. https://doi.org/10.32942/X2JP8H

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. *Environmental Monitoring and Assessment*, *194*(11), 795. https://doi.org/10.1007/s10661-022-10361-9

Boeke, J. D., Garfinkel, D. J., Styles, C. A., & Fink, G. R. (1985). Ty elements transpose through an RNA intermediate. *Cell*, *40*(3), 491–500. https://doi.org/10.1016/0092-8674(85)90197-7

Boštjančić, L. L., Bonassin, L., Anušić, L., Lovrenčić, L., Besendorfer, V., Maguire, I., Grandjean, F., Austin, C. M., Greve, C., Hamadou, A. Ben, & Mlinarec, J. (2021). The Pontastacus leptodactylus (Astacidae) Repeatome Provides Insight Into Genome Evolution and Reveals Remarkable Diversity of Satellite DNA. *Frontiers in Genetics*, *11*. 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). Host-pathogen coevolution drives innate immune response to Aphanomyces astaci infection in freshwater crayfish: transcriptomic evidence. *BMC Genomics*, *23*(1), 600. https://doi.org/10.1186/s12864-022-08571-z

Bourgeois, Y., & Boissinot, S. (2019). On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. *Genes*, *10*(6), 419. https://doi.org/10.3390/genes10060419

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 06 Biological Sciences 0604 Genetics. *Genome Biology*, *19*(1), 1–12. https://doi.org/10.1186/s13059-018-1577-z

Bracken-Grissom, H. D., Ahyong, S. T., Wilkinson, R. D., Feldmann, R. M., Schweitzer, C. E., Breinholt, J. W., Bendall, M., Palero, F., Chan, T.-Y., Felder, D. L., Robles, R., Chu, K.-H., Tsang, L.-M., Kim, D., Martin, J. W., & Crandall, K. A. (2014). The Emergence of Lobsters: Phylogenetic Relationships, Morphological Evolution and Divergence Time Comparisons of an Ancient Group (Decapoda: Achelata, Astacidea, Glypheidea, Polychelida). *Systematic Biology*, *63*(4), 457–479. https://doi.org/10.1093/sysbio/syu008

Canoy, R. J., Shmakova, A., Karpukhina, A., Shepelev, M., Germini, D., & Vassetzky, Y. (2022). Factors That Affect the Formation of Chromosomal Translocations in Cells. *Cancers*, *14*(20), 5110. https://doi.org/10.3390/cancers14205110

Carvalho, F., Pascoal, C., Cássio, F., Teixeira, A., & Sousa, R. (2022). Combined per-capita and abundance effects of an invasive species on native invertebrate diversity and a key ecosystem process. *Freshwater Biology*, *67*(5), 828–841. https://doi.org/10.1111/fwb.13884

Charlesworth, B., Charlesworth, D., & Barton, N. H. (2003). The Effects of Genetic and Geographic Structure on Neutral Variation. *Annual Review of Ecology, Evolution, and Systematics*, *34*(1), 99–125. https://doi.org/10.1146/annurev.ecolsys.34.011802.132359

Chen, J.-W., Shrestha, L., Green, G., Leier, A., & Marquez-Lago, T. T. (2023). The hitchhikers' guide to RNA sequencing and functional analysis. *Briefings in Bioinformatics*, *24*(1). https://doi.org/10.1093/bib/bbac529

Claros, M. G., Bautista, R., Guerrero-Fernández, D., Benzerki, H., Seoane, P., & Fernández-Pozo, N. (2012). Why Assembling Plant Genome Sequences Is So Challenging. *Biology*, *1*(2), 439–459. https://doi.org/10.3390/biology1020439

Colonna Romano, N., & Fanti, L. (2022). Transposable Elements: Major Players in Shaping Genomic and Evolutionary Patterns. *Cells*, *11*(6), 1048. https://doi.org/10.3390/cells11061048

Coluccia, E., Cannas, R., Cau, A., Deiana, A. M., & Salvadori, S. (2004). B chromosomes in Crustacea Decapoda. *Cytogenetic and Genome Research*, *106*(2–4), 215–221. https://doi.org/10.1159/000079290

Council Directive 92/43/EEC. (2007). EU Habitats Directive Annex II: animal and plant species of community interest whose conservation requires the designation of special areas of conservation. In EC (Ed.), *Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora*. European council.

Crandall, K. A., & Buhay, J. E. (2007). Global diversity of crayfish (Astacidae, Cambaridae, and Parastacidae—Decapoda) in freshwater. In *Freshwater Animal Diversity Assessment* (pp. 295–301). Springer Netherlands. 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. *Journal of Crustacean Biology*, *37*(5), 615–653. https://doi.org/10.1093/jcbiol/rux070

Crandall, K. A., Templeton, A. R., & Neigel, J. (1999). The zoogeography and centers of origin of the crayfish subgenus Procericambarus (Decapoda: Cambaridae). *Evolution*, *53*(1), 123–134. https://doi.org/10.1111/j.1558-5646.1999.tb05338.x

Cui, R., Medeiros, T., Willemsen, D., Iasi, L. N. M., Collier, G. E., Graef, M., Reichard, M., & Valenzano, D. R. (2019). Relaxed Selection Limits Lifespan by Increasing Mutation Load. *Cell*, *178*(2), 385-399.e20. https://doi.org/10.1016/j.cell.2019.06.004

Cumberlidge, N., Hobbs, H. H., & Lodge, D. M. (2015). Class Malacostraca, Order Decapoda. In *Thorp and Covich's Freshwater Invertebrates* (pp. 797–847). Elsevier. https://doi.org/10.1016/B978-0-12-385026-3.00032-2

Dahn, H. A., Mountcastle, J., Balacco, J., Winkler, S., Bista, I., Schmitt, A. D., Pettersson, O. V., Formenti, G., Oliver, K., Smith, M., Tan, W., Kraus, A., Mac, S., Komoroske, L. M., Lama, T., Crawford, A. J., Murphy, R. W., Brown, S., Scott, A. F., … Fedrigo, O. (2022). Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing. *GigaScience*, *11*. https://doi.org/10.1093/gigascience/giac068

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). *Journal of Crustacean Biology*, *43*(3). https://doi.org/10.1093/jcbiol/ruad042

dos Santos, R. Z., Calegari, R. M., Silva, D. M. Z. de A., Ruiz-Ruano, F. J., Melo, S., Oliveira, C., Foresti, F., Uliano-Silva, M., Foresti, F. P., & Utsunomia, R. (2021). A long-term conserved satellite DNA that remains unexpanded in several genomes of Characiformes fish is actively transcribed. *Genome Biology and Evolution*. https://doi.org/10.1093/gbe/evab002

Dufresne, F., & Jeffery, N. (2011). A guided tour of large genome size in animals: what we know and where we are heading. *Chromosome Research*, *19*(7), 925–938. https://doi.org/10.1007/s10577-011-9248-x

Elliott, T. A., & Gregory, T. R. (2015). Do larger genomes contain more diverse transposable elements? *BMC Evolutionary Biology*, *15*(1), 69. https://doi.org/10.1186/s12862-015-0339-8

Espinosa, E., Bautista, R., Larrosa, R., & Plata, O. (2024). Advancements in long-read genome sequencing technologies and algorithms. In *Genomics* (Vol. 116, Issue 3, p. 110842). Academic Press. https://doi.org/10.1016/j.ygeno.2024.110842

Fang, G., Hammar, S., & Grumet, R. (1992). A quick and inexpensive method for removing polysaccharides from plant genomic DNA. *BioTechniques*, *13*(1), 52–54, 56. http://www.ncbi.nlm.nih.gov/pubmed/1503775

Farhadi, A., Pichlmueller, F., Yellapu, B., Lavery, S., & Jeffs, A. (2022). Genome-wide SNPs reveal fine-scale genetic structure in ornate spiny lobster Panulirus ornatus throughout Indo-West Pacific Ocean. *ICES Journal of Marine Science*, *79*(6), 1931–1941.

https://doi.org/10.1093/icesjms/fsac130

Feng, J. B., & Li, J. L. (2008). Twelve polymorphic microsatellites in Oriental river prawn, Macrobrachium nipponense. *Molecular Ecology Resources*, *8*(5), 986–988. https://doi.org/10.1111/j.1755-0998.2008.02129.x

Ferree, P. M., & Prasad, S. (2012). How Can Satellite DNA Divergence Cause Reproductive Isolation? Let Us Count the Chromosomal Ways. *Genetics Research International*, *2012*, 1–11. https://doi.org/10.1155/2012/430136

Florea, L., Souvorov, A., Kalbfleisch, T. S., & Salzberg, S. L. (2011). Genome Assembly Has a Major Impact on Gene Content: A Comparison of Annotation in Two Bos Taurus Assemblies. *PLoS ONE*, *6*(6), e21400. https://doi.org/10.1371/journal.pone.0021400

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. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117

Flynn, J. M., & Yamashita, Y. M. (2024). The implications of satellite DNA instability on cellular function and evolution. *Seminars in Cell & Developmental Biology*, *156*, 152–159. https://doi.org/10.1016/j.semcdb.2023.10.005

Fonseca-Carvalho, M., Veríssimo, G., Lopes, M., Ferreira, D., Louzada, S., & Chaves, R. (2024). Answering the Cell Stress Call: Satellite Non-Coding Transcription as a Response Mechanism. *Biomolecules*, *14*(1), 124. https://doi.org/10.3390/biom14010124

Formenti, G., Theissinger, K., Fernandes, C., Bista, I., Bombarely, A., Bleidorn, C., Ciofi, C., Crottini, A., Godoy, J. A., Höglund, J., Malukiewicz, J., Mouton, A., Oomen, R. A., Paez, S., Palsbøll, P. J., Pampoulie, C., Ruiz-López, M. J., Svardal, H., Theofanopoulou, C., … Zammit, G. (2022). The era of reference genomes in conservation genomics. *Trends in Ecology & Evolution*, *37*(3), 197–202. https://doi.org/10.1016/j.tree.2021.11.008

Frankham, R. (2019). Conservation Genetics. In *Encyclopedia of Ecology* (pp. 382–390). Elsevier. https://doi.org/10.1016/B978-0-12-409548-9.10559-7

Fry, K., & Salser, W. (1977). Nucleotide sequences of HS-α satellite DNA from kangaroo rat dipodomys ordii and characterization of similar sequences in other rodents. *Cell*, *12*(4), 1069–1084. https://doi.org/10.1016/0092-8674(77)90170-2

Fukuzawa, A. (2001). Invertebrate connectin spans as much as 3.5 microm in the giant sarcomeres of crayfish claw muscle. *The EMBO Journal*, *20*(17), 4826–4835. https://doi.org/10.1093/emboj/20.17.4826

Fuller, Z. L., Koury, S. A., Phadnis, N., & Schaeffer, S. W. (2019). How chromosomal rearrangements shape adaptation and speciation: Case studies in Drosophila pseudoobscura and its sibling species Drosophila persimilis. *Molecular Ecology*, *28*(6), 1283–1301. https://doi.org/10.1111/mec.14923

Füreder, L., & Reynolds, J. D. (2003). Is Austropotamobius pallipes a good bioindicator? *BFPP - Bulletin Francais de La Peche et de La Protection Des Milieux Aquatiques*, *370-371 SPEC. ISS.*, 157–163. https://doi.org/10.1051/kmae:2003011

Galbraith, D. W., Harkins, K. R., Maddox, J. M., Ayres, N. M., Sharma, D. P., & Firoozabady, E. (1983). Rapid Flow Cytometric Analysis of the Cell Cycle in Intact

Plant Tissues. *Science*, *220*(4601), 1049–1051. https://doi.org/10.1126/science.220.4601.1049

Garrido-Ramos, M. A. (2017). Satellite DNA: An evolving topic. *Genes*, *8*(9). https://doi.org/10.3390/genes8090230

Genome 10K Community of Scientists. (2009). Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species. *Journal of Heredity*, *100*(6), 659–674. https://doi.org/10.1093/jhered/esp086

Gherardi, F., Souty-Grosset, C., Vogt, G., Diéguez-Uribeondo, J., & Crandall, K. A. (2010). Infraorder Astacidea Latreille, 1802 p.p.: The freshwater crayfish. In *Crustacea* (Vol. 1, pp. 269–423). Koninklijke Brill NV.

Gianazza, E., Eberini, I., Palazzolo, L., & Miller, I. (2021). Hemolymph proteins: An overview across marine arthropods and molluscs. *Journal of Proteomics*, *245*, 104294. https://doi.org/10.1016/j.jprot.2021.104294

Gilbert, C., & Feschotte, C. (2018). Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. *Current Opinion in Genetics & Development*, *49*, 15–24. https://doi.org/10.1016/j.gde.2018.02.007

Grau, J. H., Hackl, T., Koepfli, K.-P., & Hofreiter, M. (2018). Improving draft genome contiguity with reference-derived in silico mate-pair libraries. *GigaScience*, *7*(5). https://doi.org/10.1093/gigascience/giy029

Greenleaf, W. J., & Sidow, A. (2014). The future of sequencing: convergence of intelligent design and market Darwinism. *Genome Biology*, *15*(3), 303. https://doi.org/10.1186/gb4168

Gregory, T. R. (2002). Genome size and developmental complexity. *Genetica*, *115*(1), 131–146. https://doi.org/10.1023/A:1016032400147

Gregory, T. R. (2025). *Animal Genome Size Database*. http://www.genomesize.com

Gross, R., Lovrenčić, L., Jelić, M., Grandjean, F., Đuretanović, S., Simić, V., Burimski, O., Bonassin, L., Groza, M.-I., & Maguire, I. (2021). Genetic diversity and structure of the noble crayfish populations in the Balkan Peninsula revealed by mitochondrial and microsatellite DNA markers. *PeerJ*, *9*(August), e11838. https://doi.org/10.7717/peerj.11838

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. *Nature Ecology & Evolution*, *2*(3), 567–573. https://doi.org/10.1038/s41559-018-0467-9

Harrow, J., Nagy, A., Reymond, A., Alioto, T., Patthy, L., Antonarakis, S. E., & Guigó, R. (2009). Identifying protein-coding genes in genomic sequences. *Genome Biology*, *10*(1), 201. https://doi.org/10.1186/gb-2009-10-1-201

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Heras, S., Planella, L., Caldarazzo, I., Vera, M., García-Marín, J.-L., & Roldán, M. I. (2016). Development and characterization of novel microsatellite markers by Next Generation Sequencing for the blue and red shrimp Aristeus antennatus. *PeerJ*, *4*, e2200. https://doi.org/10.7717/peerj.2200

Herrmann, A., Grabow, K., & Martens, A. (2022). The invasive crayfish Faxonius immunis causes the collapse of macroinvertebrate communities in Central European ponds. *Aquatic Ecology*, *56*(3), 741–750. https://doi.org/10.1007/s10452-021-09935-5

Hess, J. F., Kohl, T. A., Kotrová, M., Rönsch, K., Paprotka, T., Mohr, V., Hutzenlaub, T., Brüggemann, M., Zengerle, R., Niemann, S., & Paust, N. (2020). Library preparation for next generation sequencing: A review of automation strategies. *Biotechnology Advances*, *41*, 107537. https://doi.org/10.1016/j.biotechadv.2020.107537

Hessen, D. O., & Persson, J. (2009). Genome size as a determinant of growth and life-history traits in crustaceans. *Biological Journal of the Linnean Society*, *98*(2), 393–399. https://doi.org/10.1111/j.1095-8312.2009.01285.x

Hoban, S., Bruford, M. W., da Silva, J. M., Funk, W. C., Frankham, R., Gill, M. J., Grueber, C. E., Heuertz, M., Hunter, M. E., Kershaw, F., Lacy, R. C., Lees, C., Lopes-Fernandes, M., MacDonald, A. J., Mastretta-Yanes, A., McGowan, P. J. K., Meek, M. H., Mergeay, J., Millette, K. L., … Laikre, L. (2023). Genetic diversity goals and targets have improved, but remain insufficient for clear implementation of the post-2020 global biodiversity framework. *Conservation Genetics*, *24*(2), 181–191. https://doi.org/10.1007/s10592-022-01492-0

Hogg, C. J., Ottewell, K., Latch, P., Rossetto, M., Biggs, J., Gilbert, A., Richmond, S., & Belov, K. (2022). Threatened Species Initiative: Empowering conservation action using genomic resources. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(4), 1–8. https://doi.org/10.1073/pnas.2115643118

Höglund, J. (2009). *Evolutionary conservation genetics*. Oxford University Press.

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. *Knowledge and Management of Aquatic Ecosystems*, *2009*(394–395), 394–395. https://doi.org/10.1051/kmae/2009025

Hong, Q., Chen, Y.-L., Lin, D., Yang, R.-Q., Cao, K.-Y., Zhang, L.-J., Liu, Y.-M., Sun, L.-C., & Cao, M.-J. (2024). Expression of polyphenol oxidase of Litopenaeus vannamei and its characterization. *Food Chemistry*, *432*, 137258. https://doi.org/10.1016/j.foodchem.2023.137258

Hossain, M. A., Lahoz-Monfort, J. J., Burgman, M. A., Böhm, M., Kujala, H., & Bland, L. M. (2018). Assessing the vulnerability of freshwater crayfish to climate change. *Diversity and Distributions*, *24*(12), 1830–1843. https://doi.org/10.1111/ddi.12831

Houben, A., Banaei-Moghaddam, A. M., Klemme, S., & Timmis, J. N. (2014). Evolution and biology of supernumerary B chromosomes. *Cellular and Molecular Life Sciences*, *71*(3), 467–478. https://doi.org/10.1007/s00018-013-1437-7

Howard, C., Denton, A., Jackson, B., Bates, A., Jay, J., Yatsenko, H., Raman, P. S., Thomas, A., Oatley, G., do Amaral, R. V., Göktan, Z. E., Gómez, J. P. N., Lucey, I. C., Sinclair, E., Quail, M. A., Blaxter, M., Howe, K., & Lawniczak, M. K. N. (2025). *On the path to reference genomes for all biodiversity: lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species*. https://doi.org/10.1101/2025.04.11.648334

Huang, S., Ding, J., Deng, D., Tang, W., Sun, H., Liu, D., Zhang, L., Niu, X., Zhang, X., Meng, M., Yu, J., Liu, J., Han, Y., Shi, W., Zhang, D., Cao, S., Wei, Z., Cui, Y., Xia, Y.,

… Liu, Y. (2013). Draft genome of the kiwifruit Actinidia chinensis. *Nature Communications*, *4*(1), 2640. https://doi.org/10.1038/ncomms3640

Hudina, S., Hock, K., & Žganec, K. (2014). The role of aggression in range expansion and biological invasions. *Current Zoology*, *60*(3), 401–409. https://doi.org/10.1093/czoolo/60.3.401

Hultgren, K. M., Jeffery, N. W., Moran, A., & Gregory, T. R. (2018). Latitudinal variation in genome size in crustaceans. *Biological Journal of the Linnean Society*, *123*(2), 348–359. https://doi.org/10.1093/biolinnean/blx153

Hurgobin, B., & Edwards, D. (2017). SNP Discovery Using a Pangenome: Has the Single Reference Approach Become Obsolete? *Biology*, *6*(1), 21. https://doi.org/10.3390/biology6010021

Hurt, C., Hildreth, P., & Williams, C. (2022). A genomic perspective on the conservation status of the endangered Nashville crayfish (Faxonius shoupi). *Conservation Genetics*, *23*(3), 589–604. https://doi.org/10.1007/s10592-022-01438-6

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. *Frontiers in Ecology and Evolution*, *10*(August), 1–15. https://doi.org/10.3389/fevo.2022.930888

Ion, M. C., Ács, A.-R., Laza, A. V., Lorincz, I., Livadariu, D., Lamoly, A. M., Goia, B., Togor, A., Iorgu, E. I., Ştefan, A., Popa, O. P., & Pârvulescu, L. (2024). Conservation status of the idle crayfish Austropotamobius bihariensis Pârvulescu, 2019. *Global Ecology and Conservation*, *50*, e02847. https://doi.org/10.1016/j.gecco.2024.e02847

IUCN. (2001). IUCN Red List categories and criteria. *IUCN Species Survival Commission*.

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 239. https://doi.org/10.1186/s13059-016-1103-0

Jeffery, N. W. (2015). *Genome size diversity and evolution in Crustacea*. University of Guelph.

Jiang, N. (2013). Overview of Repeat Annotation and De Novo Repeat Identification. In T. Peterson (Ed.), *Plant Transposable Elements* (Vol. 1057, pp. 275–287). Humana Press. https://doi.org/10.1007/978-1-62703-568-2

Johnson, W. E., Onorato, D. P., Roelke, M. E., Land, E. D., Cunningham, M., Belden, R. C., McBride, R., Jansen, D., Lotz, M., Shindle, D., Howard, J., Wildt, D. E., Penfold, L. M., Hostetler, J. A., Oli, M. K., & O'Brien, S. J. (2010). Genetic Restoration of the Florida Panther. *Science*, *329*(5999), 1641–1645. https://doi.org/10.1126/science.1192891

Jussila, J., Edsman, L., Maguire, I., Diéguez-Uribeondo, J., & Theissinger, K. (2021). Money Kills Native Ecosystems: European Crayfish as an Example. *Frontiers in Ecology and Evolution*, *9*. https://doi.org/10.3389/fevo.2021.648495

Jussila, J., Francesconi, C., Theissinger, K., Kokko, H., & Makkonen, J. (2021). Is Aphanomyces astaci Losing its Stamina: A Latent Crayfish Plague Disease Agent from Lake Venesjärvi, Finland. *Freshwater Crayfish*, *26*(2), 139–144. https://doi.org/10.5869/fc.2021.v26-2.139

Kalendar, R., Ivanov, K. I., Samuilova, O., Kairov, U., & Zamyatnin, A. A. (2023). Isolation of High-Molecular-Weight DNA for Long-Read Sequencing Using a High-Salt Gel Electroelution Trap. *Analytical Chemistry*, *95*(48), 17818–17825. https://doi.org/10.1021/acs.analchem.3c03894

Kardos, M., Armstrong, E. E., Fitzpatrick, S. W., Hauser, S., Hedrick, P. W., Miller, J. M., Tallmon, D. A., & Chris Funk, W. (2021). The crucial role of genome-wide genetic variation in conservation. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(48). https://doi.org/10.1073/pnas.2104642118

Kawai, T., & Crandall, K. A. (2016). Global Diversity and Conservation of Freshwater Crayfish (Crustacea: Decapoda: Astacoidea). In *A Global Overview of the Conservation of Freshwater Decapod Crustaceans* (pp. 65–114). Springer International Publishing. https://doi.org/10.1007/978-3-319-42527-6_3

Kim, J., Lee, C., Ko, B. J., Yoo, D. A., Won, S., Phillippy, A. M., Fedrigo, O., Zhang, G., Howe, K., Wood, J., Durbin, R., Formenti, G., Brown, S., Cantin, L., Mello, C. V., Cho, S., Rhie, A., Kim, H., & Jarvis, E. D. (2022). False gene and chromosome losses in genome assemblies caused by GC content variation and repeats. *Genome Biology*, *23*(1), 204. https://doi.org/10.1186/s13059-022-02765-0

Korlach, J., Bjornson, K. P., Chaudhuri, B. P., Cicero, R. L., Flusberg, B. A., Gray, J. J., Holden, D., Saxena, R., Wegener, J., & Turner, S. W. (2010). *Real-Time DNA Sequencing from Single Polymerase Molecules* (pp. 431–455). https://doi.org/10.1016/S0076-6879(10)72001-2

Kouba, A., Petrusek, A., & Kozák, P. (2014). Continental-wide distribution of crayfish species in Europe: update and maps. *Knowledge and Management of Aquatic Ecosystems*, *413*, 05. https://doi.org/10.1051/kmae/2014007

Kyriakidou, M., Tai, H. H., Anglin, N. L., Ellis, D., & Strömvik, M. V. (2018). Current Strategies of Polyploid Plant Genome Sequence Assembly. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.01660

Laffitte, M., Baudry, T., Guilmet, M., Andrieu, T., Poulet, N., Duperray, T., Carine, D., Collas, M., Moumen, B., Sudres, M., & Grandjean, F. (2023). A new invader in freshwater ecosystems in France: the rusty crayfish Faxonius rusticus. *BioInvasions Records*, *12*(2), 457–468. https://doi.org/10.3391/bir.2023.12.2.10

Laggis, A., Baxevanis, A. D., Charalampidou, A., Maniatsi, S., Triantafyllidis, A., & Abatzopoulos, T. J. (2017). Microevolution of the noble crayfish (Astacus astacus) in the Southern Balkan Peninsula. *BMC Evolutionary Biology*, *17*(1), 1–19. https://doi.org/10.1186/s12862-017-0971-6

Lanciano, S., & Cristofari, G. (2020). Measuring and interpreting transposable element expression. *Nature Reviews Genetics*, *21*(12), 721–736. https://doi.org/10.1038/s41576-020-0251-y

Lavalli, K., & Spanier, E. (2016). Predator Adaptations of Decapods. *Lifestyles and Feeding Biology. Edited By Martin Thiel and Les Watling. © 2015 Oxford University Press. Published 2015 by Oxford University Press*, *January 2015*, 190–228.

Lécher, P., Defaye, D., & Noel, P. (1995). Chromosomes and nuclear DNA of crustacea. *Invertebrate Reproduction and Development*, *27*(2), 85–114. https://doi.org/10.1080/07924259.1995.9672440

Lee, M., Kim, Y., Nam, D., & Cho, K. (2023). Impacts of the accumulated extinction of endangered species on stream food webs. *Global Ecology and Conservation*, *48*, e02747. https://doi.org/10.1016/j.gecco.2023.e02747

Lefébure, T., Morvan, C., Malard, F., François, C., Konecny-Dupré, L., Guéguen, L., Weiss-Gayet, M., Seguin-Orlando, A., Ermini, L., Sarkissian, C. Der, Charrier, N. P., Eme, D., Mermillod-Blondin, F., Duret, L., Vieira, C., Orlando, L., & Douady, C. J. (2017). Less effective selection leads to larger genomes. *Genome Research*, *27*(6), 1016–1028. https://doi.org/10.1101/gr.212589.116

Lertzman-Lepofsky, G., Mooers, A. Ø., & Greenberg, D. A. (2019). Ecological constraints associated with genome size across salamander lineages. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1911), 20191780. https://doi.org/10.1098/rspb.2019.1780

Li, S.-F., Su, T., Cheng, G.-Q., Wang, B.-X., Li, X., Deng, C.-L., & Gao, W.-J. (2017). Chromosome Evolution in Connection with Repetitive Sequences and Epigenetics in Plants. *Genes*, *8*(10), 290. https://doi.org/10.3390/genes8100290

Liao, M., Xu, M., Hu, R., Xu, Z., Bonvillain, C., Li, Y., Li, X., Luo, X., Wang, J., Wang, J., Zhao, S., & Gu, Z. (2024). The chromosome-level genome assembly of the red swamp crayfish Procambarus clarkii. *Scientific Data*, *11*(1), 1–8. https://doi.org/10.1038/s41597-024-03718-x

Liu, J., Sun, Y., Chen, Q., Wang, M., Li, Q., Zhou, W., & Cheng, Y. (2023). Genetic Diversity Analysis of the Red Swamp Crayfish Procambarus clarkii in Three Cultured Populations Based on Microsatellite Markers. *Animals*, *13*(11), 1881. https://doi.org/10.3390/ani13111881

Lopez, R., Chen, Y.-J., Dumas Ang, S., Yekhanin, S., Makarychev, K., Racz, M. Z., Seelig, G., Strauss, K., & Ceze, L. (2019). DNA assembly for nanopore data storage readout. *Nature Communications*, *10*(1), 2933. https://doi.org/10.1038/s41467-019-10978-4

Louzada, S., Lopes, M., Ferreira, D., Adega, F., Escudeiro, A., Gama-carvalho, M., & Chaves, R. (2020). Architecture and Plasticity — An Evolutionary and Clinical A ffair. *Genes*.

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 Evolutionary Biology*, *20*(1), 146. https://doi.org/10.1186/s12862-020-01709-1

Lovrenčić, L., Temunović, M., Bonassin, L., Grandjean, F., Austin, C. M., & Maguire, I. (2022). Climate change threatens unique genetic diversity within the Balkan biodiversity hotspot – The case of the endangered stone crayfish. *Global Ecology and Conservation*, *39*(July). https://doi.org/10.1016/j.gecco.2022.e02301

Luangtrakul, W., Wongdontri, C., Jaree, P., Boonchuen, P., Somboonviwat, K., Sarnow, P., & Somboonwiwat, K. (2025). Unveiling the impact of shrimp piRNAs on WSSV infection and immune modulation. *Fish & Shellfish Immunology*, *158*, 110124. https://doi.org/10.1016/j.fsi.2025.110124

Lynch, M. (2007). *The origins of genome architecture*. Sinauer associates.

Mackintosh, A., Vila, R., Laetsch, D. R., Hayward, A., Martin, S. H., & Lohse, K. (2023).

Chromosome Fissions and Fusions Act as Barriers to Gene Flow between Brenthis Fritillary Butterflies. *Molecular Biology and Evolution*, *40*(3). https://doi.org/10.1093/molbev/msad043

Maguire, I., Jelić, M., Klobučar, G., Delpy, M., Delaunay, C., & Grandjean, F. (2016). Prevalence of the pathogen Aphanomyces astaci in freshwater crayfish populations in Croatia. *Diseases of Aquatic Organisms*, *118*(1), 45–53. https://doi.org/10.3354/dao02955

Malicki, M., Spaller, T., Winckler, T., & Hammann, C. (2020). DIRS retrotransposons amplify via linear, single-stranded cDNA intermediates. *Nucleic Acids Research*, *48*(8), 4230–4243. https://doi.org/10.1093/nar/gkaa160

Mann, L., Balasch, K., Schmidt, N., & Heitkam, T. (2024). High-fidelity (repeat) consensus sequences from short reads using combined read clustering and assembly. *BMC Genomics*, *25*(1), 109. https://doi.org/10.1186/s12864-023-09948-4

Männer, L., Schell, T., Spies, J., Galià-Camps, C., Baranski, D., Ben Hamadou, A., Gerheim, C., Neveling, K., Helfrich, E. J. N., & Greve, C. (2024). Chromosome-level genome assembly of the sacoglossan sea slug Elysia timida (Risso, 1818). *BMC Genomics*, *25*(1), 941. https://doi.org/10.1186/s12864-024-10829-7

Marino, A., Debaecker, G., Fiston-Lavier, A.-S., Haudry, A., & Nabholz, B. (2025). Effective population size does not explain long-term variation in genome size and transposable element content in animals. *ELife*, *13*. https://doi.org/10.7554/eLife.100574.3

Marn, N., Hudina, S., Haberle, I., Dobrović, A., & Klanjšček, T. (2022). Physiological performance of native and invasive crayfish species in a changing environment: insights from Dynamic Energy Budget models. *Conservation Physiology*, *10*(1). https://doi.org/10.1093/conphys/coac031

Martinsen, L., Venanzetti, F., Johnsen, A., Sbordoni, V., & Bachmann, L. (2009). Molecular evolution of the pDo500 satellite DNA family in Dolichopoda cave crickets (Rhaphidophoridae). *BMC Evolutionary Biology*, *9*(1), 301. https://doi.org/10.1186/1471-2148-9-301

Mason, A. S. (2015). *SSR Genotyping* (pp. 77–89). https://doi.org/10.1007/978-1-4939-1966-6_6

Mastretta-Yanes, A., da Silva, J. M., Grueber, C. E., Castillo-Reina, L., Köppä, V., Forester, B. R., Funk, W. C., Heuertz, M., Ishihama, F., Jordan, R., Mergeay, J., Paz-Vinas, I., Rincon-Parra, V. J., Rodriguez-Morales, M. A., Arredondo-Amezcua, L., Brahy, G., DeSaix, M., Durkee, L., Hamilton, A., … Hoban, S. (2024). Multinational evaluation of genetic diversity indicators for the Kunming-Montreal Global Biodiversity Framework. *Ecology Letters*, *27*(7). https://doi.org/10.1111/ele.14461

Mayrose, I., & Lysak, M. A. (2021). The Evolution of Chromosome Numbers: Mechanistic Models and Experimental Approaches. *Genome Biology and Evolution*, *13*(2). https://doi.org/10.1093/gbe/evaa220

McCord, R. P., Kaplan, N., & Giorgetti, L. (2020). Chromosome Conformation Capture and Beyond: Toward an Integrative View of Chromosome Structure and Function. *Molecular Cell*, *77*(4), 688–708. https://doi.org/10.1016/j.molcel.2019.12.021

McGaw, I. J., & Curtis, D. L. (2024). Feeding and digestive processes. In *Ecophysiology of*

the European Green Crab (Carcinus Maenas) and Related Species (pp. 81–101). Elsevier. https://doi.org/10.1016/B978-0-323-99694-5.00012-X

McLaughlin, C. M., Hinshaw, C., Sandoval-Arango, S., Zavala-Paez, M., & Hamilton, J. A. (2025). Redlisting genetics: towards inclusion of genetic data in IUCN Red List assessments. *Conservation Genetics*, *26*(2), 213–223. https://doi.org/10.1007/s10592-024-01671-1

McMahon, B. J., Teeling, E. C., & Höglund, J. (2014). How and why should we implement genomics into conservation? *Evolutionary Applications*, *7*(9), 999–1007. https://doi.org/10.1111/eva.12193

Melters, D. P., Bradnam, K. R., Young, H. A., Telis, N., May, M. R., Ruby, J. G., Sebra, R., Peluso, P., Eid, J., Rank, D., Garcia, J. F., DeRisi, J. L., Smith, T., Tobias, C., Ross-Ibarra, J., Korf, I., & Chan, S. W. L. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology*, *14*(1), R10. https://doi.org/10.1186/gb-2013-14-1-r10

Mérel, V., Tricou, T., Burlet, N., & Haudry, A. (2025). Relaxed Purifying Selection is Associated with an Accumulation of Transposable Elements in Flies. *Molecular Biology and Evolution*, *42*(6). https://doi.org/10.1093/molbev/msaf111

Mlinarec, J., Porupski, I., Maguire, I., & Klobučar, G. (2016). Comparative karyotype investigations in the white-clawed crayfish Austropotamobius pallipes (Lereboullet, 1858) species complex and stone crayfish A. torrentium (Schrank, 1803) (Decapoda: Astacidae). *Journal of Crustacean Biology*, *36*(1), 87–93. https://doi.org/10.1163/1937240X-00002390

Molina, W. F., Costa, G. W. W. F., Cunha, I. M. C., Bertollo, L. A. C., Ezaz, T., Liehr, T., & Cioffi, M. B. (2020). Molecular Cytogenetic Analysis in Freshwater Prawns of the Genus Macrobrachium (Crustacea: Decapoda: Palaemonidae). *International Journal of Molecular Sciences*, *21*(7), 2599. https://doi.org/10.3390/ijms21072599

Montanari, S., Deng, C., Koot, E., Bassil, N. V, Zurn, J. D., Morrison-Whittle, P., Worthington, M. L., Aryal, R., Ashrafi, H., Pradelles, J., Wellenreuther, M., & Chagné, D. (2023). A multiplexed plant–animal SNP array for selective breeding and species conservation applications. *G3: Genes, Genomes, Genetics*, *13*(10). https://doi.org/10.1093/g3journal/jkad170

Munoz-Lopez, M., & Garcia-Perez, J. (2010). DNA Transposons: Nature and Applications in Genomics. *Current Genomics*, *11*(2), 115–128. https://doi.org/10.2174/138920210790886871

NCBI Genomes. (n.d.). *Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information*. https://www.ncbi.nlm.nih.gov/home/genomes/

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 Research*, *45*(12), e111–e111. https://doi.org/10.1093/nar/gkx257

Novák, P., Neumann, P., & Macas, J. (2020). Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nature Protocols*, *15*(11), 3745–3776. https://doi.org/10.1038/s41596-020-0400-y

Olsson, K. (2008). *Dynamics of omnivorous crayfish in freshwater ecosystems*. Lund University.

Otto, F. (1992). Preparation and staining of cells for high-resolution DNA analysis. In A. Radbruch (Ed.), *Flow cytometry and cell sorting* (pp. 101–104). Springer-Verlag.

PacBio. (2018). *Low Yield Troubleshooting Guide*. *101-627-900–01*.

PacBio. (2021). *SMRT® sequencing — Delivering highly accurate long reads to drive discovery in life science*. *102-193–63*.

PacBio. (2025). *Preparing whole genome and metagenome libraries using SMRTbell® prep kit 3.0 Procedure & checklist*. *102-166-600 REV07*.

Paez, S., Kraus, R. H. S., Shapiro, B., Gilbert, M. T. P., Jarvis, E. D., & Vertebrate Genomes Project Conservation Group. (2022). Reference genomes for conservation. *Science*, *377*(6604), 364–366.

Pârvulescu, L., Ion, M., & Ács, A. R. (2025). Austropotamobius bihariensis. In *IUCN Red List of Threatened Species* (Issue e.T260451001A260451021). https://doi.org/10.2305/IUCN.UK.2025-1.RLTS.T260451001A260451021.en

Pârvulescu, L., Iorgu, E. I., Zaharia, C., Ion, M. C., Satmari, A., Krapal, A. M., Popa, O. P., Miok, K., Petrescu, I., & Popa, L. O. (2020). The future of endangered crayfish in light of protected areas and habitat fragmentation. *Scientific Reports*, *10*(1), 1–12. https://doi.org/10.1038/s41598-020-71915-w

Pearman, W. S., Arranz, V., Carvajal, J. I., Whibley, A., Liau, Y., Johnson, K., Gray, R., Treece, J. M., Gemmell, N. J., Liggins, L., Fraser, C. I., Jensen, E. L., & Green, N. J. (2024). A cry for kelp: Evidence for polyphenolic inhibition of Oxford Nanopore sequencing of brown algae. *Journal of Phycology*, *60*(6), 1601–1610. https://doi.org/10.1111/jpy.13513

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 Evolutionary Biology*, *19*(1), 1–15. https://doi.org/10.1186/s12862-018-1324-9

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*(5). https://doi.org/10.1371/journal.pone.0037135

Petraccioli, A., Odierna, G., Capriglione, T., Barucca, M., Forconi, M., Olmo, E., & Biscotti, M. A. (2015). A novel satellite DNA isolated in Pecten jacobaeus shows high sequence similarity among molluscs. *Molecular Genetics and Genomics*, *290*(5), 1717–1725. https://doi.org/10.1007/s00438-015-1036-4

Pezer, Ž., Brajković, J., Feliciello, I., & Ugarković, Đ. (2012). Satellite DNA-Mediated Effects on Genome Regulation. In M. A. Garrido-Ramos (Ed.), *Genome Dynamics* (Volume 7, pp. 153–169). Karger.

Plohl, M., Meštrović, N., & Mravinac, B. (2012). Satellite DNA Evolution. *Genome Dyn* 7, 126–52. https://doi.org/10.1159/000337122

Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I., & Bird, C. E. (2014).

Demystifying the RAD fad. *Molecular Ecology*, *23*(24), 5937–5942. https://doi.org/10.1111/mec.12965

Putman, A. I., & Carbone, I. (2014). Challenges in analysis and interpretation of microsatellite data for population genetic studies. *Ecology and Evolution*, *4*(22), 4399–4428. https://doi.org/10.1002/ece3.1305

Raza, S., Shoaib, M. W., & Mubeen, H. (2016). Molecular Markers: an Introduction and Applications. *International Journal of Scientific and Research Publications*, *6*(3), 221–223.

Reichel, K., Pohjoismäki, J., Astrin, J., Böhne, A., Bortoluzzi, C., Campo, J. del, Ciofi, C., Di-Nizo, C. B., Divakar, P. K., Greve, C., Hampl, V., Hilgers, L., Laine, V. N., Leonard, J. A., Lozano-Fernandez, J., Bilela, L. L., Mazzoni, C., McCartney, A., Melo-Ferreira, J., … Guttry, C. De. (2025). Addressing key challenges in sample handling for high-quality reference genome generation https://doi.org/10.22541/au.174889351.10048878/v1

Reid, B. N., Hofmeier, J., Crockett, H., Fitzpatrick, R., Waters, R., & Fitzpatrick, S. W. (2025). Balancing Inbreeding and Outbreeding Risks to Inform Translocations Throughout the Range of an Imperiled Darter. *Evolutionary Applications*, *18*(3). https://doi.org/10.1111/eva.70088

Reinar, W. B., Lalun, V. O., Reitan, T., Jakobsen, K. S., & Butenko, M. A. (2021). Length variation in short tandem repeats affects gene expression in natural populations of Arabidopsis thaliana. *The Plant Cell*, *33*(7), 2221–2234. https://doi.org/10.1093/plcell/koab107

Reynolds, J., Souty-Grosset, C., & Richardson, A. (2013). Ecological roles of crayfish in freshwater and terrestrial habitats. *Freshwater Crayfish*, *19*(2), 197–218. https://doi.org/10.5869/fc.2013.v19-2.197

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), 737–746. https://doi.org/10.1038/s41586-021-03451-0

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*, *13*(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Richard, G.-F., Kerrest, A., & Dujon, B. (2008). Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes. *Microbiology and Molecular Biology Reviews*, *72*(4), 686–727. https://doi.org/10.1128/mmbr.00011-08

Richman, N. I., Böhm, M., Adams, S. B., Alvarez, F., Bergey, E. A., Bunn, J. J. S., Burnham, Q., Cordeiro, J., Coughran, J., Crandall, K. A., Dawkins, K. L., DiStefano, R. J., Doran, N. E., Edsman, L., Eversole, A. G., Füreder, L., Furse, J. M., Gherardi, F., Hamr, P., … Collen, B. (2015). Multiple drivers of decline in the global status of freshwater crayfish (Decapoda: Astacidea). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1662), 20140060. https://doi.org/10.1098/rstb.2014.0060

Robles, F., de la Herrán, R., Ludwig, A., Ruiz Rejón, C., Ruiz Rejón, M., & Garrido-Ramos, M. A. (2004). Evolution of ancient satellite DNAs in sturgeon genomes. *Gene*, *338*(1),

133–142. https://doi.org/10.1016/j.gene.2004.06.001

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. *Scientific Reports*, *6*(June), 1–14. https://doi.org/10.1038/srep28333

Salem, A. (2024). *A Review of Transposable Elements in Crustacea*. Williams Honors College.

Salvadori, S., Deidda, F., Carugati, L., Melis, R., Costa, E., Sibiriu, M., & Coluccia, E. (2023). Chromosomal mapping of ribosomal clusters and telomeric sequences (TTAGGn in nine species of lobsters (Crustacea, Decapoda) . *The European Zoological Journal*, *90*(1), 443–453. https://doi.org/10.1080/24750263.2023.2217188

Sambrook, J., & Russell, D. W. (2006). Purification of Nucleic Acids by Extraction with Phenol:Chloroform. *Cold Spring Harbor Protocols*, *2006*(1), pdb.prot4455. https://doi.org/10.1101/pdb.prot4455

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., Slocombe, P. M., & Smith, M. (1977). Nucleotide sequence of bacteriophage φX174 DNA. *Nature*, *265*(5596), 687–695. https://doi.org/10.1038/265687a0

Šatović-Vukšić, E., & Plohl, M. (2023). Satellite DNAs—From Localized to Highly Dispersed Genome Components. In *Genes* (Vol. 14, Issue 3). https://doi.org/10.3390/genes14030742

Šatović, E., & Plohl, M. (2013). Tandem repeat-containing MITEs in the clam Donax trunculus. *Genome Biology and Evolution*, *5*(12), 2549–2559. https://doi.org/10.1093/gbe/evt202

Sayer, C. A., Fernando, E., Jimenez, R. R., Macfarlane, N. B. W., Rapacciuolo, G., Böhm, M., Brooks, T. M., Contreras-MacBeath, T., Cox, N. A., Harrison, I., Hoffmann, M., Jenkins, R., Smith, K. G., Vié, J. C., Abbott, J. C., Allen, D. J., Allen, G. R., Barrios, V., Boudot, J. P., … Darwall, W. R. T. (2025). One-quarter of freshwater fauna threatened with extinction. *Nature*, *638*(8049), 138–145. https://doi.org/10.1038/s41586-024-08375-z

Scalvenzi, T., & Pollet, N. (2014). Insights on genome size evolution from a miniature inverted repeat transposon driving a satellite DNA. *Molecular Phylogenetics and Evolution*, *81*, 1–9. https://doi.org/10.1016/j.ympev.2014.08.014

Scheben, A., Batley, J., & Edwards, D. (2017). Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant Biotechnology Journal*, *15*(2), 149–161. https://doi.org/10.1111/pbi.12645

Schell, T., Greve, C., & Podsiadlowski, L. (2025). Establishing genome sequencing and assembly for non-model and emerging model organisms: a brief guide. *Frontiers in Zoology*, *22*(1), 1–27. https://doi.org/10.1186/s12983-025-00561-7

Schrader, L., & Schmitz, J. (2019). The impact of transposable elements in adaptive evolution. *Molecular Ecology*, *28*(6), 1537–1549. https://doi.org/10.1111/mec.14794

Schrimpf, A., Theissinger, K., Dahlem, J., Maguire, I., Pârvulescu, L., Schulz, H. K., & Schulz, R. (2014). Phylogeography of noble crayfish (A stacus astacus) reveals multiple refugia. *Freshwater Biology*, *59*(4), 761–776. https://doi.org/10.1111/fwb.12302

Schubert, I. (2007). Chromosome evolution. *Current Opinion in Plant Biology*, *10*(2), 109–115. https://doi.org/10.1016/j.pbi.2007.01.001

Shah, A., Hoffman, J. I., & Schielzeth, H. (2020). Comparative Analysis of Genomic Repeat Content in Gomphocerine Grasshoppers Reveals Expansion of Satellite DNA and Helitrons in Species with Unusually Large Genomes. *Genome Biology and Evolution*, *12*(7), 1180–1193. https://doi.org/10.1093/gbe/evaa119

Shao, C., Sun, S., Liu, K., Wang, J., Li, S., Liu, Q., Deagle, B. E., Seim, I., Biscontin, A., Wang, Q., Liu, X., Kawaguchi, S., Liu, Y., Jarman, S., Wang, Y., Wang, H. Y., Huang, G., Hu, J., Feng, B., … Fan, G. (2023). The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights. *Cell*, *186*(6), 1279-1294.e19. https://doi.org/10.1016/j.cell.2023.02.005

Shaw, R. E., Farquharson, K. A., Bruford, M. W., Coates, D. J., Elliott, C. P., Mergeay, J., Ottewell, K. M., Segelbacher, G., Hoban, S., Hvilsom, C., Pérez-Espona, S., Ruņģis, D., Aravanopoulos, F., Bertola, L. D., Cotrim, H., Cox, K., Cubric-Curik, V., Ekblom, R., Godoy, J. A., … Grueber, C. E. (2025). Global meta-analysis shows action is needed to halt genetic diversity loss. *Nature*, *638*(8051), 704–710. https://doi.org/10.1038/s41586-024-08458-x

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Skinner, D. M., & Beattie, W. G. (1974). Characterization of a pair of isopycnic twin crustacean satellite deoxyribonucleic acids, one of which lacks one base in each strand. *Biochemistry*, *13*(19), 3922–3929. https://doi.org/10.1021/bi00716a017

Söderhäll, I., Fasterius, E., Ekblom, C., & Söderhäll, K. (2022). Characterization of hemocytes and hematopoietic cells of a freshwater crayfish based on single-cell transcriptome analysis. *IScience*, *25*(8). https://doi.org/10.1016/j.isci.2022.104850

Sollars, E. S. A., Harper, A. L., Kelly, L. J., Sambles, C. M., Ramirez-Gonzalez, R. H., Swarbreck, D., Kaithakottil, G., Cooper, E. D., Uauy, C., Havlickova, L., Worswick, G., Studholme, D. J., Zohren, J., Salmon, D. L., Clavijo, B. J., Li, Y., He, Z., Fellgett, A., McKinney, L. V., … Buggs, R. J. A. (2017). Genome sequence and genetic diversity of European ash trees. *Nature*, *541*(7636), 212–216. https://doi.org/10.1038/nature20786

Soto, I., Ahmed, D. A., Beidas, A., Oficialdegui, F. J., Tricarico, E., Angeler, D. G., Amatulli, G., Briski, E., Datry, T., Dohet, A., Domisch, S., England, J., Feio, M. J., Forcellini, M., Johnson, R. K., Jones, J. I., Larrañaga, A., L'Hoste, L., Murphy, J. F., … Haubrock, P. J. (2023). Long-term trends in crayfish invasions across European rivers. *Science of The Total Environment*, *867*, 161537. https://doi.org/10.1016/j.scitotenv.2023.161537

Spradling, A. C., Bellen, H. J., & Hoskins, R. A. (2011). Drosophila P elements preferentially transpose to replication origins. *Proceedings of the National Academy of Sciences*, *108*(38), 15948–15953. https://doi.org/10.1073/pnas.1112960108

Storer, J., Hubley, R., Rosen, J., & Smit, A. (2022). Methodologies for the De novo Discovery of Transposable Element Families. *Genes*, *13*(4), 709. https://doi.org/10.3390/genes13040709

Sudan, J., Singh, R., Sharma, S., Salgotra, R. K., Sharma, V., Singh, G., Sharma, I., Sharma, S., Gupta, S. K., & Zargar, S. M. (2019). ddRAD sequencing-based identification of inter-genepool SNPs and association analysis in Brassica juncea. *BMC Plant Biology*, *19*(1), 594. https://doi.org/10.1186/s12870-019-2188-x

Supple, M. A., & Shapiro, B. (2018). Conservation of biodiversity in the genomics era. *Genome Biology*, *19*(1), 131. https://doi.org/10.1186/s13059-018-1520-3

Sureshkumar, S., Chhabra, A., Guo, Y., & Balasubramanian, S. (2025). Simple sequence repeats and their expansions: role in plant development, environmental response and adaptation. *New Phytologist*, *247*(2), 504–517. https://doi.org/10.1111/nph.70173

Tan, G., Opitz, L., Schlapbach, R., & Rehrauer, H. (2019). Long fragments achieve lower base quality in Illumina paired-end sequencing. *Scientific Reports*, *9*(1), 2856. https://doi.org/10.1038/s41598-019-39076-7

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. *Frontiers in Genetics*, *11*(March). https://doi.org/10.3389/fgene.2020.00201

Tan, S. C., & Yiap, B. C. (2009). DNA, RNA, and Protein Extraction: The Past and The Present. *BioMed Research International*, *2009*(1). https://doi.org/10.1155/2009/574398

Tarandek, A., Lovrenčić, L., Židak, L., Topić, M., Grbin, D., Gregov, M., Ćurko, J., Hudina, S., & Maguire, I. (2023). Characteristics of the Stone Crayfish Population along a Disturbance Gradient—A Case Study of the Kustošak Stream, Croatia. *Diversity*, *15*(5), 591. https://doi.org/10.3390/d15050591

Tempel, S. (2012). Using and Understanding RepeatMasker. In Y. Bigot (Ed.), *Mobile Genetic Elements*. Humana Press. https://doi.org/10.1007/978-1-61779-603-6_2

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., … Zammit, G. (2023). How genomics can help biodiversity conservation. *Trends in Genetics*, *39*(7), 545–559. https://doi.org/10.1016/j.tig.2023.01.005

Toon, A., Pérez-Losada, M., Schweitzer, C. E., Feldmann, R. M., Carlson, M., & Crandall, K. A. (2010). Gondwanan radiation of the Southern Hemisphere crayfishes (Decapoda: Parastacidae): evidence from fossils and molecules. *Journal of Biogeography*, *37*(12), 2275–2290. https://doi.org/10.1111/j.1365-2699.2010.02374.x

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 Research*, *47*(21), 10994–11006. https://doi.org/10.1093/nar/gkz841

Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S., & Turner, S. W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*, *38*(15), e159–e159. https://doi.org/10.1093/nar/gkq543

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, *13*(1), 36–46.

https://doi.org/10.1038/nrg3117

Trigodet, F., Lolans, K., Fogarty, E., Shaiber, A., Morrison, H. G., Barreiro, L., Jabri, B., & Eren, A. M. (2022). High molecular weight DNA extraction strategies for long-read sequencing of complex metagenomes. *Molecular Ecology Resources*, *22*(5), 1786–1802. https://doi.org/10.1111/1755-0998.13588

Veitia, R. A., & Bottani, S. (2009). Whole Genome Duplications and a 'Function' for Junk DNA? Facts and Hypotheses. *PLoS ONE*, *4*(12), e8201. https://doi.org/10.1371/journal.pone.0008201

Victoriano, P. F., & D'Elía, G. (2021). Evolving in islands of mud: old and structured hidden diversity in an endemic freshwater crayfish from the Chilean hotspot. *Scientific Reports*, *11*(1), 8573. https://doi.org/10.1038/s41598-021-88019-8

Vieira, M. L. C., Santini, L., Diniz, A. L., & Munhoz, C. de F. (2016). Microsatellite markers: What they mean and why they are so useful. *Genetics and Molecular Biology*, *39*(3), 312–328. https://doi.org/10.1590/1678-4685-GMB-2016-0027

Vu, N. T. T., Zenger, K. R., Silva, C. N. S., Guppy, J. L., & Jerry, D. R. (2021). Population Structure, Genetic Connectivity, and Signatures of Local Adaptation of the Giant Black Tiger Shrimp (Penaeus monodon) throughout the Indo-Pacific Region. *Genome Biology and Evolution*, *13*(10). https://doi.org/10.1093/gbe/evab214

Wallinger, C., Staudacher, K., Sint, D., Thalinger, B., Oehm, J., Juen, A., & Traugott, M. (2017). Evaluation of an automated protocol for efficient and reliable <scp>DNA</scp> extraction of dietary samples. *Ecology and Evolution*, *7*(16), 6382–6389. https://doi.org/10.1002/ece3.3197

Wang, J., Itgen, M. W., Wang, H., Gong, Y., Jiang, J., Li, J., Sun, C., Sessions, S. K., & Mueller, R. L. (2021). Gigantic Genomes Provide Empirical Tests of Transposable Element Dynamics Models. *Genomics, Proteomics & Bioinformatics*, *19*(1), 123–139. https://doi.org/10.1016/j.gpb.2020.11.005

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, *39*(11), 1348–1365. https://doi.org/10.1038/s41587-021-01108-x

Waples, R. S. (2025). The Idiot's Guide to Effective Population Size. *Molecular Ecology*. https://doi.org/10.1111/mec.17670

Warne, R. K., & Chaber, A.-L. (2023). Assessing Disease Risks in Wildlife Translocation Projects: A Comprehensive Review of Disease Incidents. *Animals*, *13*(21), 3379. https://doi.org/10.3390/ani13213379

Wells, J. N., & Feschotte, C. (2020). A Field Guide to Eukaryotic Transposable Elements. *Annual Review of Genetics*, *54*, 539–561. https://doi.org/10.1146/annurev-genet-040620-022145

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., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Wenne, R. (2023). Single Nucleotide Polymorphism Markers with Applications in Conservation and Exploitation of Aquatic Natural Populations. *Animals*, *13*(6), 1–25. https://doi.org/10.3390/ani13061089

Williams-Subiza, E. A., & Epele, L. B. (2021). Drivers of biodiversity loss in freshwater environments: A bibliometric analysis of the recent literature. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *31*(9), 2469–2480. https://doi.org/10.1002/aqc.3627

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. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1901), 20190079. https://doi.org/10.1098/rspb.2019.0079

Wu, Y., Pegoraro, A. F., Weitz, D. A., Janmey, P., & Sun, S. X. (2022). The correlation between cell and nucleus size is explained by an eukaryotic cell growth model. *PLOS Computational Biology*, *18*(2), e1009400. https://doi.org/10.1371/journal.pcbi.1009400

Xu, Y., Tang, Y., Feng, W., Yang, Y., & Cui, Z. (2023). Comparative Analysis of Transposable Elements Reveals the Diversity of Transposable Elements in Decapoda and Their Effects on Genomic Evolution. *Marine Biotechnology*, *25*(6), 1136–1146. https://doi.org/10.1007/s10126-023-10265-w

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*(5), 3274–3284. https://doi.org/10.1016/j.ygeno.2021.07.017

Yirgu, M., Kebede, M., Feyissa, T., Lakew, B., Woldeyohannes, A. B., & Fikere, M. (2023). Single nucleotide polymorphism (SNP) markers for genetic diversity and population structure study in Ethiopian barley (Hordeum vulgare L.) germplasm. *BMC Genomic Data*, *24*(1), 1–13. https://doi.org/10.1186/s12863-023-01109-6

Yuan, J., Yu, Y., Zhang, X., Li, S., Xiang, J., & Li, F. (2023). Recent advances in crustacean genomics and their potential application in aquaculture. *Reviews in Aquaculture*, *July 2022*, 1501–1521. https://doi.org/10.1111/raq.12791

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

Zeng, X., Zhao, W., Kanika, N. H., Dong, Y., Hou, X., Chen, X., Wang, J., & Wang, C. (2025). Comparative Analysis of Transposable Element Evolution in Crustaceans. *Genome Biology and Evolution*, *17*(7). https://doi.org/10.1093/gbe/evaf115

Zhang, H.-H., Peccoud, J., Xu, M.-R.-X., Zhang, X.-G., & Gilbert, C. (2020). Horizontal transfer and evolution of transposable elements in vertebrates. *Nature Communications*, *11*(1), 1362. https://doi.org/10.1038/s41467-020-15149-4

Zhang, X., Yuan, J., Sun, Y., Li, S., Gao, Y., Yu, Y., Liu, C., Wang, Q., Lv, X., Zhang, X., Ma, K. Y., Wang, X., Lin, W., Wang, L., Zhu, X., Zhang, C., Zhang, J., Jin, S., Yu, K., … Xiang, J. (2019). Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. *Nature Communications*, *10*(1), 356. https://doi.org/10.1038/s41467-018-08197-4

Zimmerman, S. J., Aldridge, C. L., & Oyler-Mccance, S. J. (2020). An empirical comparison of population genetic analyses using microsatellite and SNP data for a species of conservation concern. *BMC Genomics*, *21*(1), 1–16. https://doi.org/10.1186/s12864-020-06783-9

Appendix

Status and author contributions of publications included in the thesis

Chapter II

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

Conceptualization, C.R., L.B., C.F., L.L.B., K.T. and O.L.; *methodology*, C.R., D.M., Lj.L.B., L.B., C.F. and O.L.; *software*, C.R. and A.K.; *visualisation*, C.R.; *writing—original draft preparation*, C.R., K.T. and O.L.; *writing—review and editing*, C.R., C.F., L.B., Lj.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*.

In this chapter, I contributed to the conceptualisation and methodology of the study, interpretation of the results, as well as the writing of the manuscript. I was involved in the development of the workflow, software choice and annotation protocol.

Chapter III

Bonassin, L., Boštjančić, L.L., Rutz, C., Francesconi, C., Schardt, L., Baranski, D., Greve, C., Pârvulescu, L., Mlinarec, J., Besendorfer, V., Maguire, I., Theissinger, K., Lecompte, O. The extraordinary satellitome diversity of freshwater crayfish: a driver of genome evolution. Under review in *BMC Mobile DNA* doi:10.21203/rs.3.rs-7499918/v1

Conceptualisation L.B., L.L.B., I.M., J.M., V.B, K.T., O.L.; *Data curation* L.B.; *Formal analysis* L.B.; *Investigation* L.B., L.L.B., D.B., C.G.; *Methodology* L.B., L.L.B., J.M., V.B.; *Software* L.B., C.R., C.F.; Visualisation L.B., L.L.B., *Resources* L.S., L.P., K.T., O.L.; *Writing – original draft* L.B.; *Validation, Writing – rewieving and editing* L.B., L.L.B., C.R., C.F., L.S., D.B., C.G., L.P., J.M., V.B., I.M., K.T, O.L; *Supervision* L.P., J.M., V.B., I.M., K.T., O.L.; *Project administration* L.L.B, L.P., K.T., O.L.; Funding acquisition L.L.B., L.P., J.M., V.B., I.M., K.T., O.L. *All authors read and approved the final version of the manuscript.*

In this chapter I lead the investigation and the study. This included study conceptualisation and development and the implementation of the research methodology. I performed DNA extraction for Illumina sequencing. I curated and analysed the sequencing data. I framed and established the analytical and visualisation approach of the study and wrote the original draft of the manuscript.

Chapter IV

Bonassin, L., Rutz, C., Boštjančić, L.L., Francesconi, C., Schardt, L., Greve, C., Ben Hamadou, A., Baranski, D., Gerheim, C., Feldmeyer, B., Ploch, S., Kress., A., Wollenweber, T.E., Pârvulescu, L., Lecompte, O., Theissinger, K. (in preparation). From DNA extraction to long read sequencing: workflow challenges of giant genomes of two non-model decapod species

Conceptualisation L.B., O.L., K.T.; Sample acquisition L.P.; *Formal analysis* L.B.; *Investigation* L.B., L.S., A.B.H., C. Ge; *Methodology* L.B., L.L.B., C.F., L.S., C.Gr, A.B.H, D.B., C.Ge, B.F, S.P., T.E.W, O.L., K.T.; *Software* L.B., C.R., A.K.; *Visualisation* L.B.; *Resources* L.S., C.Gr, S.P., T.E.W., O.L., K.T.; *Validation, Writing – original draft* L.B.; *Writing – rewieving and editing* L.B, C.R., L.L.B., C.F., L.S., C.G., A.B.H., C.H, B.F., S.P., A.K., T.E.W., L.P., O.L., K.T.; *Supervision* O.L, K.T; *Project administration* O.L., K.T.; Funding acquisition L.P., O.L., K.T.

In this chapter I lead the investigation and the study. This included study conceptualisation and development and the implementation of the research methodology I performed DNA extractions and library preparations. I customised and optimised the protocols. I framed and established the analytical and visualisation approach of the study and wrote the original draft of the manuscript.

Chapter V

Bonassin, L., Pârvulescu, L., Boštjančić, L.L., Francesconi, C., Paetsch, J., Rutz, C., Lecompte, O., Theissinger, K. (2024) Genomic insights into the conservation status of the Idle Crayfish *Austropotamobius bihariensis* Pârvulescu, 2019: low genetic diversity in the endemic crayfish species of the Apuseni Mountains. BMC *Ecol Evo* 24, 78. https://doi.org/10.1186/s12862-024-02268-5

Conceptualisation L.P, K.T., L.B; *Sample acquisition* L.P.; *Methodology* L.B., L.P., Lj.L.B, C.F., C.R., O.L., K.T.; *Investigation* L.B., J.P.; *Data curation* L.B.; *Formal analysis* L.B.;

Software L.B; *Visualisation* L.B., L.P.; *Funding acquisition* L.P., O.L., K.T.; *Project administration* L.P., O.L., K.T.; *Supervision* L.P., O.L., K.T.; *Validation* L.B., L.P., Lj.L.B, C.F., J.P., C.R., O.L., K.T.; *Writing-original draft* L.B. *All authors read and approved the final version of the manuscript*.

In this chapter I performed DNA extractions and methodological approach. I was involved in the study conceptualisation and framed all analytical and visualisation approaches of the study and wrote the original draft of the manuscript.

Supplementary material

Supplementary material is provided below for each chapter. Due to size or extension of the files incompatible with a printed version, the remaining Supplementary material will be made available only digitally.

Supplementary material - Chapter II

The following supporting information can be downloaded at:

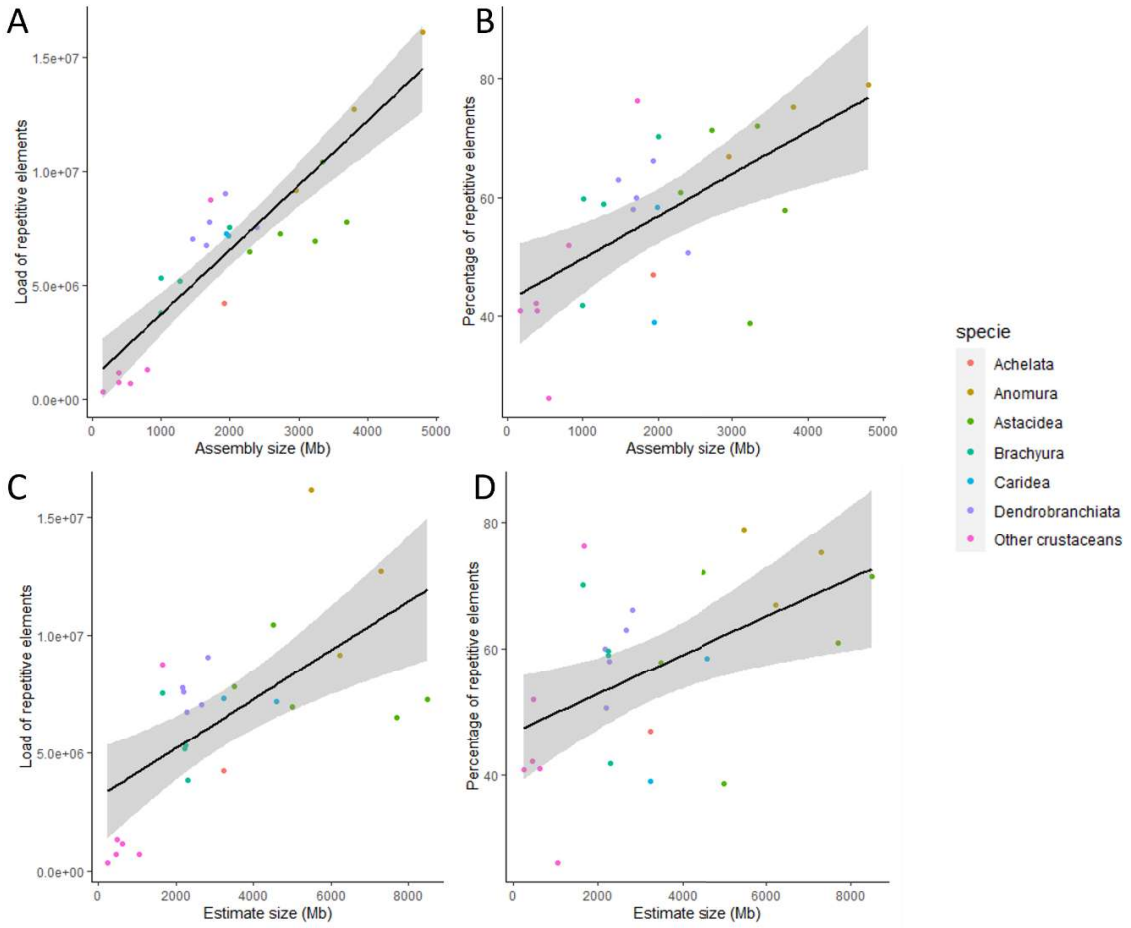
https://www.mdpi.com/article/10.3390/genes14081627/s1

Supplementary file II-1. crustaceans_RE_library.fa;

Supplementary table II-1. Assembly metrics. Contig and scaffold N50, number of scaffold and type of reads produced. L: long reads; S: short reads; O: optical mapping; M: mate pair reads.

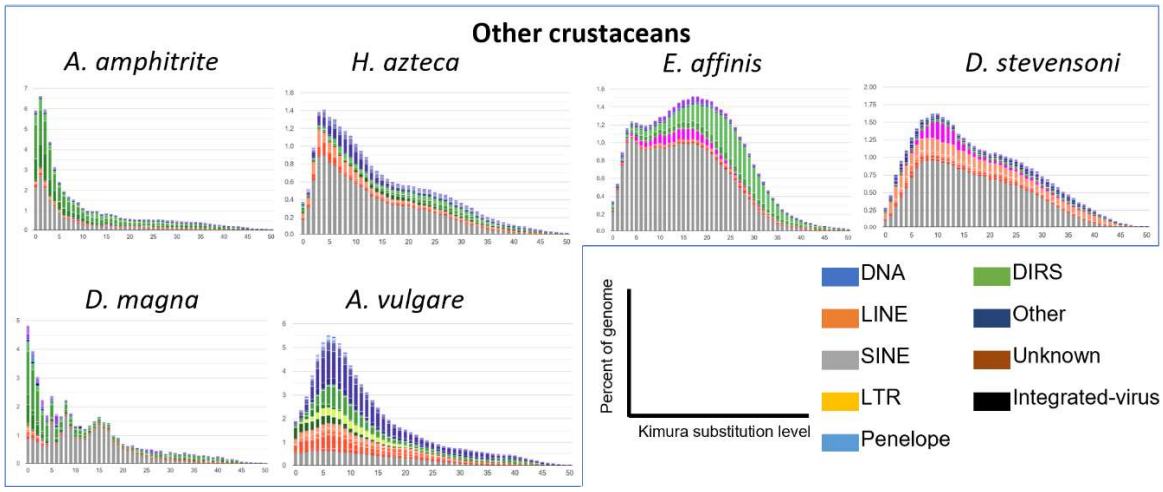
Genomes	contig N50	Number of scaffolds	Scaffold N50	Type of reads
Penaeus chinensis	470.2 kb	1060	36.9 Mb	L
Penaeus indicus	463.4 kb	11166	34.4 Mb	L+S
Penaeus japonicus	132.8 kb	18210	234.9kb	L+S
Penaeus monodon	45.2 kb	26875	44.9 Mb	L+S
Penaeus vannamei	86.9 kb	4682	605.6 kb	L+S
Caridina multidentata	819 bp	2750712	819 pb	S
Macrobrachium nipponense	267.3 kb	24	83 Mb	L
Panulirus oranatus	5.4 kb	403881	8.1 kb	S
Procambarus virginalis	12.2kb	169498	144.4 kb	L
Procambarus clarkii	217.7 kb	24238	17 Mb	L
Cherax destructor	80.9 kb	98662	87.2 kb	L+S
Cherax quadricarinatus	3.3 kb	508682	33.2 kb	L+S
Homarus americanus	133.3 kb	47245	759.6 kb	L+S
Paralithodes camtschaticus	5.8 kb	859811	7 kb	S
Paralithodes platypus	147.8 kb	6958	51.2 Mb	L
Birgus latro	5.3 kb	767134	6.3 kb	S

Chionoecetes opilio	149.6 kb	26514	208.1 kb	L+S
Eriocheir sinensis	3.2 Mb	4311	17.6 Mb	S+O
Portunus trituberculatus	4.1 Mb	523	21.8 Mb	L+S
Callinectes sapidus	9.3 kb	3967	18.8 Mb	L+S+O
Amphibalanus amphitrite	536.8 kb	/	/	L+S
Armadillidium vulgare	38.4 kb	43541	51.1 kb	L+S
Daphnia magna	1.5 Mb	308	12.5 Mb	L+S
Darwinula stevensoni	38.5 kb	62117	56.4 kb	M+S
Eurytemora affinis	67.7 kb	6171	252.3 kb	S
Hyalella azteca	112.9 kb	17395	213.8 kb	S



Supplementary figure II-1. Correlation between genome size and REs. Correlation plot between assembly or estimate genome size and load or percentage of REs. Order and suborder are grouped by colours. A. Correlation between assembly size and the load of REs. Spearman rank correlation test: ρ=0.83, p-value=1.939E-6. B. Correlation between assembly size and the percentage of REs. Spearman rank correlation test: ρ=0.54, p-value=0.00502. C. Correlation between estimate genome size and the load of REs. Spearman rank correlation

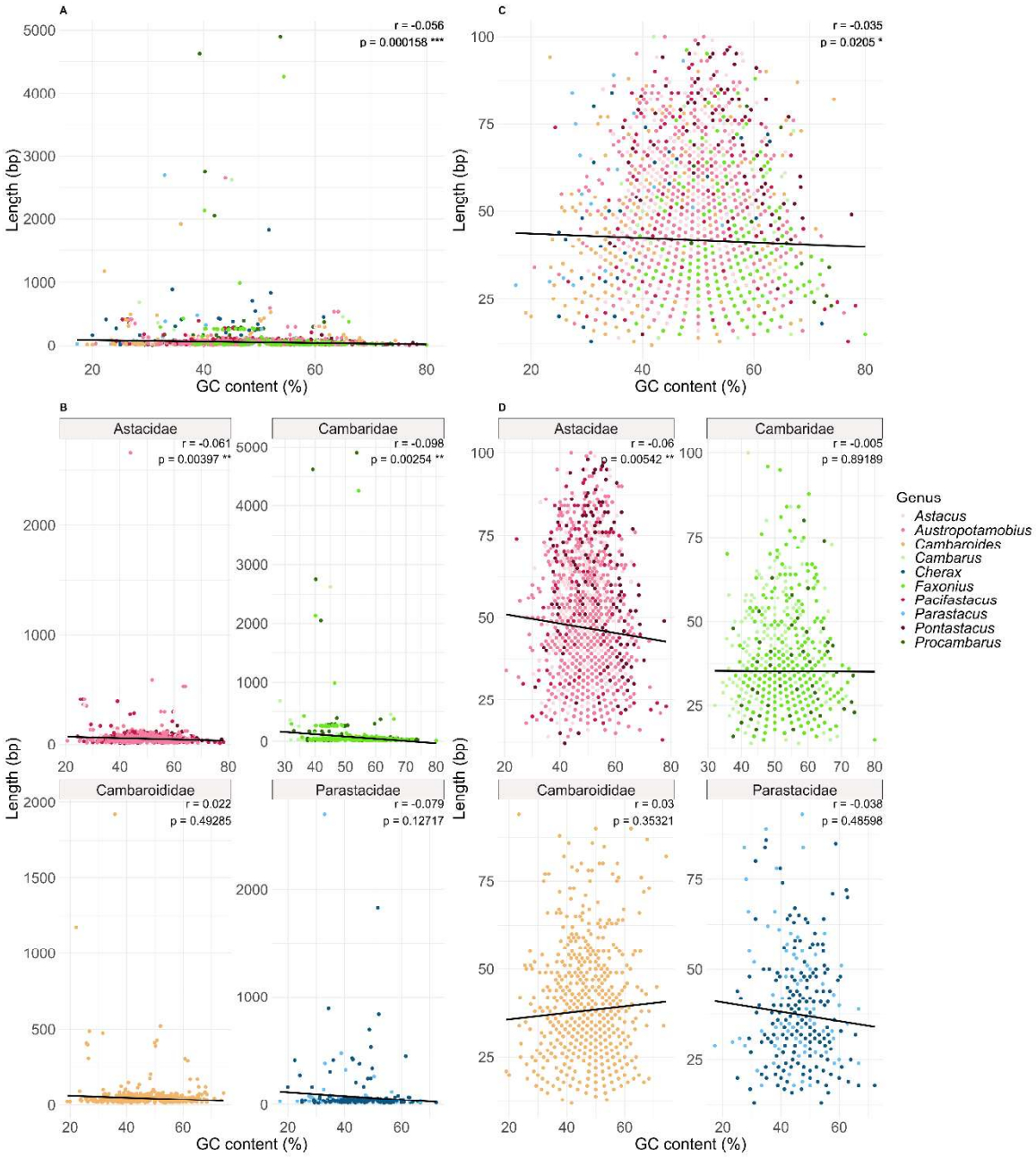
test: ρ=0.57, p-value=0.00208. D. Correlation between estimate genome size and the percentage of REs. Spearman rank correlation test: ρ=0.4, p-value=0.02745.



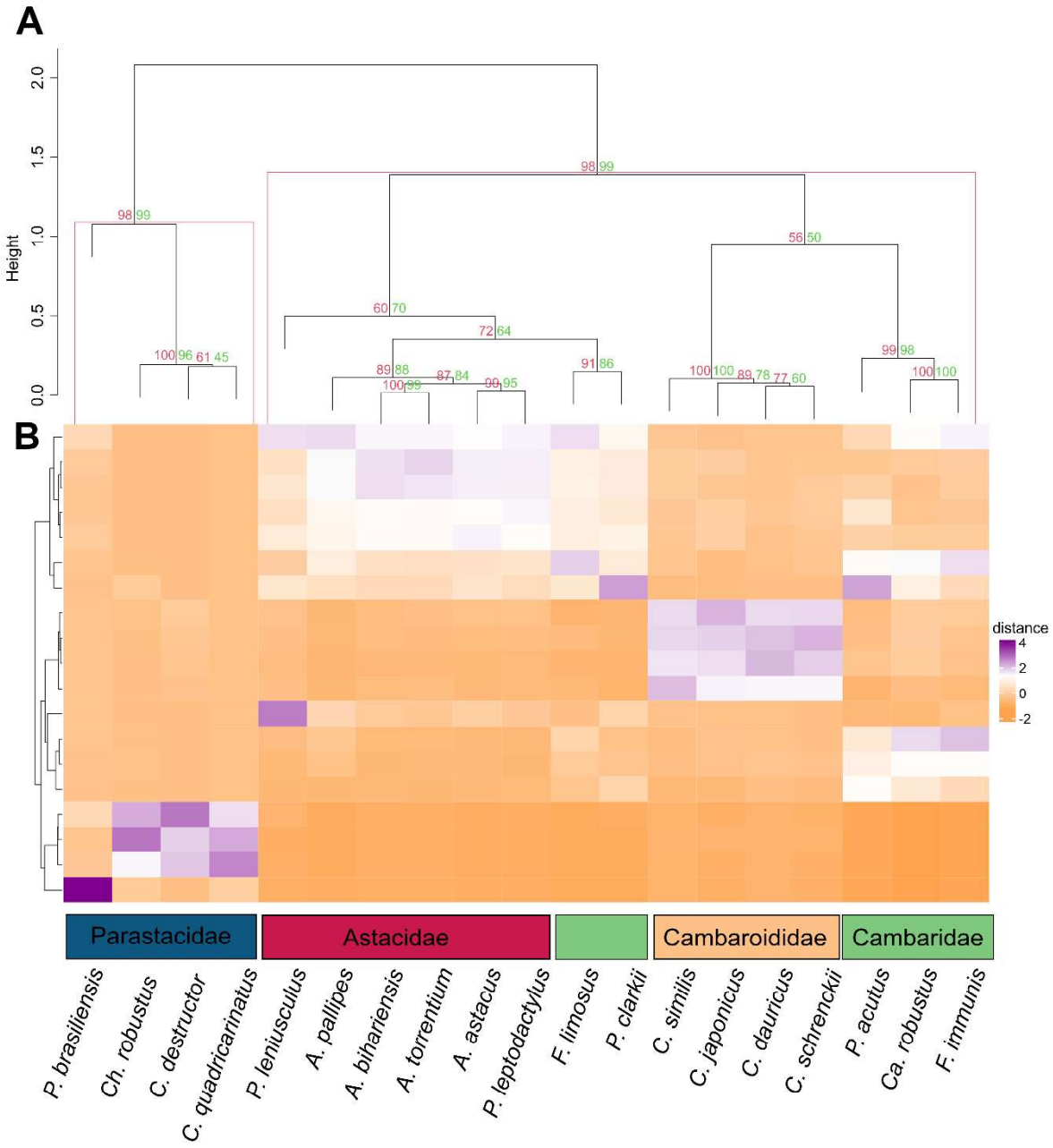
Supplementary figure II-2. Sequence divergence distribution of TEs. TE accumulation history based on Kimura 2P distance. Sequence divergence on the x-axis. On the y-axis, the percentage of the genome represented by each TE type, the scale is different for each genome depending on the percentage occupied. TE type indicated by the colour chart.

Crustaceans non-Decapoda species studied present a large fraction of unknown elements in their sequence divergence distribution. Unknown elements can largely bias the analysis. Indeed, the identification of these unknown can change interpretation. *A. amphitrite* and *D. magna* present active TE with an expansion of LTR and unknown elements. In *D. magna*, we can also observe a peak of unknown elements at 15% of divergence. We can notice in *H. azteca* a peak at 5% of divergence of unknown elements. For *E. affinis* genome there is almost no distinction between the two peaks, around 4% of divergence and 10% to 30%. LTR can be the predominant elements of the oldest event, but an identification of unknown elements can change the interpretation. In *D. stevensoni* there is a peak between 5% to 10% of divergence of DNA transposons and unknown. *A. vulgare* present a high peak at a Kimura distance of 5% to 10% with an augmentation of the coverage of DNA transposons, LINE and LTR elements. At really low Kimura divergence, we can observe an increasing coverage of Penelope elements.

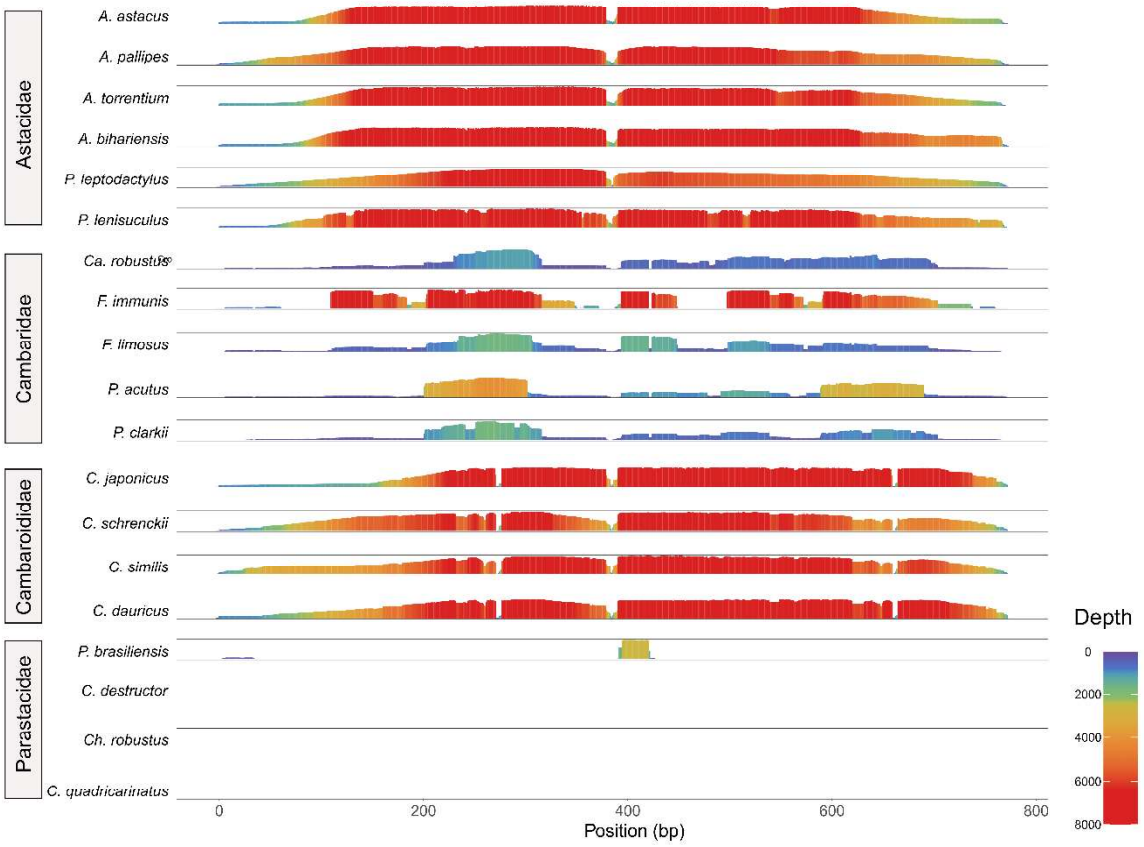
Supplementary material - Chapter III



Supplementary figure III-1. Correlation of GC content (%) and repeat unit length (bp) for (A) overall satellites (B) overall satellites per family, (C) minisatellites and (D) minisatellite per family. Colours indicate different genera. Correlation was tested using Spearman rank correlation test with significance level $\alpha=0.05$. Significance levels are indicated as follows: $p < 0.05$ *, $p < 0.01$ **, and $p < 0.001$ ***.



Supplementary figure III-2. (A) Cluster dendrogram and (B) heatmap showing hierarchical clustering of satDNA sequences in all 19 species based on observed/expected number of edges between species in RepeatExplorer2 analysis. In (A) red numbers on nodes indicate Approximately Unbiased (AU) p-value, while green numbers on nodes indicate Bootstrap Probability (BP) values. Clusters with AU larger than 95% are highlighted by rectangles. In (B) colours indicate distance values.



Supplementary figure III-3. Colour enhanced profile of PlSAT3-411 satellite DNA family against each species. Different colours indicate coverage. The height of each bar indicates the coverage of base variant in the reads

Supplementary table III-1. Overview of the analysed species, including their genus, family, accession number of the reads, number of sequencing reads obtained in this study, mitochondrial genome accession number and genome size (Gb)
https://figshare.com/s/d0629e60ce817faf53b4

Supplementary table III-2. Flow cytometry genome size measurement of haemolymph from *Astacus astacus* and *Austropotamobius bihariensis* obtained by PI fluorescence dye excitation with three chopping buffers.

Species	Chopping buffer	Genome size (Mb)	Average genome size (Mb)
A. astacus	Galbraith et al. (Galbraith et al., 1983)	17495	16891
A. astacus	Galbraith et al. (Galbraith et al., 1983)	19472	
A. astacus	Galbraith et al. (Galbraith et al., 1983)	18894	
A. astacus	Otto et al. (Otto, 1992)	15268	
A. astacus	Phosphate buffer saline	14944	
A. astacus	Phosphate buffer saline	15273	
A. bihariensis	Galbraith et al. (Galbraith et al., 1983)	12240	11583
A. bihariensis	Galbraith et al. (Galbraith et al., 1983)	12170	
A. bihariensis	Phosphate buffer saline	10340	

Supplementary table III-3. Summary of clusters identified in individual RepeatExplorer runs for each crayfish species. For each cluster are indicated the unique cluster name (CL_unique), the supercluster classification, the cluster size, automatic annotation from RepeatExplorer, TAREAN annotation, the final manually curated annotation and genome proportion (%). For satDNA sequences the sequence, length (bp) and GC content (%) are indicated.
https://figshare.com/s/a698b2280377f714bff3

Supplementary table III-4. Summary of clusters identified in the comparative RepeatExplorer run. For each cluster are indicated RepeatExplorer classification, TAREAN classification, total number of reads in a cluster and number of reads in a cluster belonging to a particular species, and correspondence to satDNA in individual clustering.
https://figshare.com/s/1809671357f57ebda826

Supplementary table III-5. Summary of Kruskal-Wallis tests comparing GC content and satDNA repeat length across genera and families. The table reports the tested variable, grouping factor, test statistic (Chi-squared), degrees of freedom (Df), and corresponding p-value for each comparison.

Variable	Grouping	Chi-squared	Df	p-value
Length	Genus	340.9753	9	5.157e-68
Length	Family	325.8408	3	2.537e-70
GC	Genus	336.2363	9	5.252e-67

GC	Family	262.8178	3	1.105e-56

Supplementary table III-6. Spearman rank correlation test between GC content and satDNA repeat length across freshwater crayfish families. The table reports the correlation value (Spearman's rho) and corresponding p-value for each family.

	Family	Correlation	p-value
Satellites	Astacidae	-0.061	0.00397
Satellites	Cambaridae	-0.098	0.00254
Satellites	Cambaroididae	0.022	0.49285
Satellites	Parastacidae	-0.079	0.12717
Minisatellites	Astacidae	-0.060	0.00542
Minisatellites	Cambaridae	-0.005	0.89189
Minisatellites	Cambaroididae	0.030	0.35321
Minisatellites	Parastacidae	-0.038	0.48598

Supplementary table III-7. Results of pairwise Wilcoxon rank-sum tests comparing GC content and satDNA repeat length across genera and families. Adjusted p-values were calculated using the Bonferroni correction to control for multiple comparisons. Significance levels are indicated as follows: $p < 0.05$ *, $p < 0.01$ **, and $p < 0.001$ ***. The table includes the tested variable, grouping factor, compared groups, adjusted p-values, and significance annotations.
https://figshare.com/s/d7fbab812b51db4ea9ba

Supplementary file III-1. Cluster_similarity.sh

https://figshare.com/s/d4fbffa9e20233416b6b

Supplementary file III-2. Cluster_similarity.R

https://figshare.com/s/e852b2baddb01352cc33

Supplementary material - Chapter IV

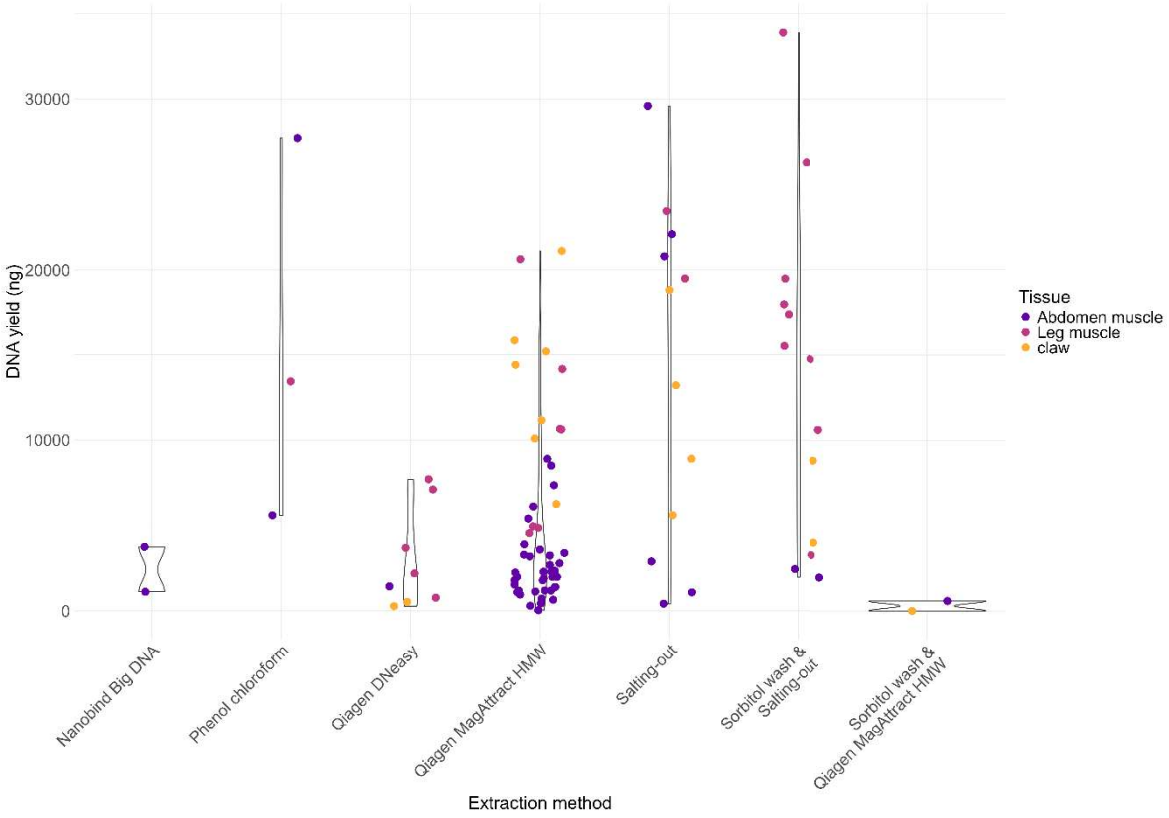
Supplementary table IV-1. Summary of sample preparation and sequencing metrics. Overview of sample, species, tissue type and DNA extraction protocol, DNA quality metrics (DNA yield, absorption ratios A260/280 and A260/230, and fragment length), library preparation information, polymerase, sequencing platform and sequencing run performance (productivity metrics P0, P1 and P2 and HiFi yield).
https://figshare.com/s/cf6d8d5f7667cc2f2f7c

Supplementary table IV-2. Analysis of Variance of Aligned Rank Transformed Data (ART ANOVA) for Tissue and Extraction method effects on DNA yield. Degrees of freedom (Df), residual degrees of freedom (Df.res), F-statistic (F value), and the corresponding p-value (Pr (>F)). Significance codes are indicated as followed: for $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*', not significant 'ns'

	Df	Df. res	F value	Pr (>F)	
Extraction	2	67	12.012	3.481e-05	***
Tissue	2	67	12.887	1.839e-05	***
Extraction:Tissue	4	67	3.232	0.0173	*

Supplementary table IV-3. Post-hoc comparisons for main effects of Extraction method and Tissue type on DNA yield. For each comparison, the estimated difference (estimate), standard error (SE), degrees of freedom (df), t-ratio, and adjusted p-value (p.value) are given. P value adjustment using Tukey method. Significance codes are indicated as followed: for $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*', not significant 'ns'

		estimate	SE	df	t.ratio	p.value	
Qiagen MagAttract HMW	Salting-out	-25.86	7.06	67	-3.664	0.0014	**
Qiagen MagAttract HMW	Sorbitol wash & Salting-out	-30.61	7.59	67	-4.030	0.0004	***
Salting-out	Sorbitol wash & Salting-out	-4.75	9.06	67	-0.524	0.859	ns
Abdomen muscle	Leg muscle	-31.44	6.56	67	-4.796	<.0001	***
Abdomen muscle	Claw	-25.23	6.86	67	-3.676	0.0014	**
Leg muscle	Claw	6.22	6.99	67	0.889	0.649	ns



Supplementary figure IV-1. Total DNA yield (ng) for all extraction methods. Colour indicates tissue types used for DNA extraction.

Supplementary table IV-4. Analysis of Variance of Aligned Rank Transformed Data for Tissue and Extraction method effects on A260/280. Degrees of freedom (Df), residual degrees of freedom (Df.res), F-statistic (F value), and the corresponding p-value (Pr (>F)). Significance codes are indicated as followed: for $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*', not significant 'ns'

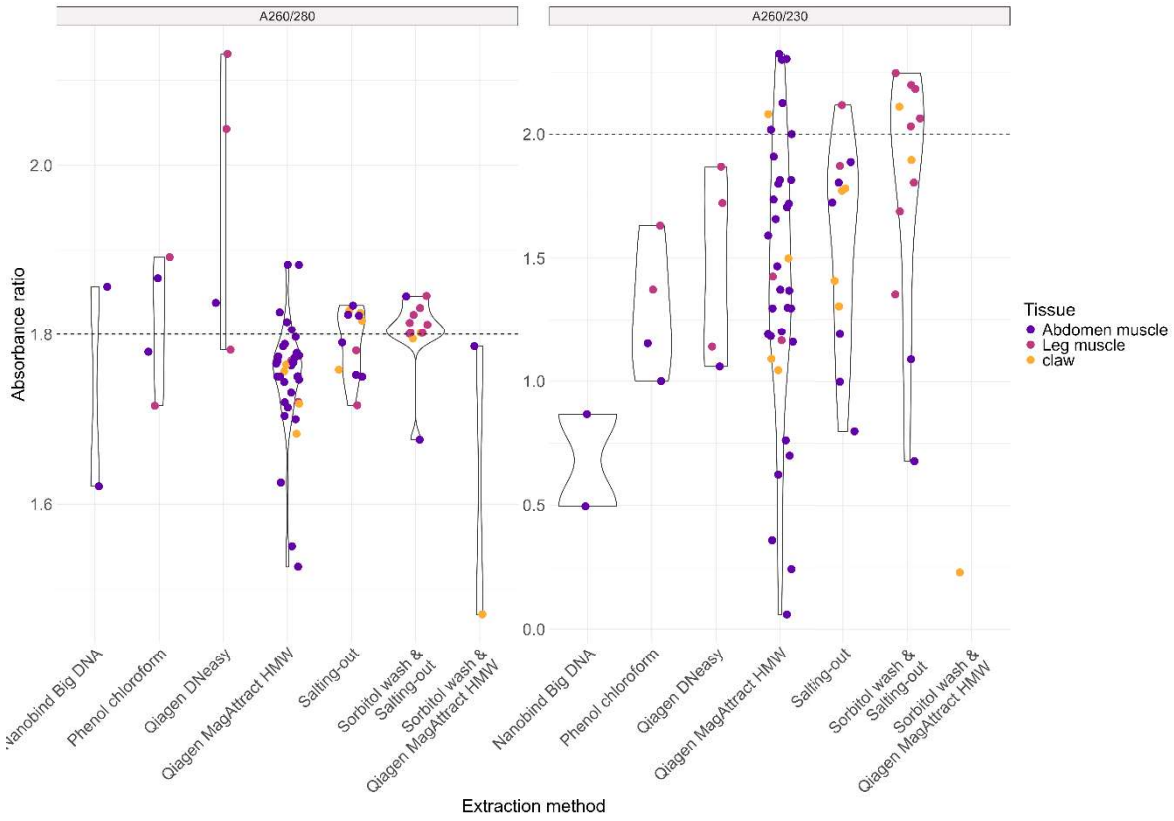
	Df	Df. res	F value	Pr (>F)	
Extraction	2	53	4.031	0.024	*
Tissue	2	53	0.404	0.669	ns
Extraction:Tissue	4	53	0.607	0.659	ns

Supplementary table IV-5. Tukey's Honestly Significant Difference (HSD) post-hoc test for multiple comparisons of means for the A260/280 ratios. For each comparison, the table includes the mean difference (diff), the lower (lwr) and upper (upr) bounds of the 95% confidence interval, and the adjusted p-value (P adj). Significance codes are indicated as followed: for $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*', not significant 'ns'

		diff	lwr	upr	P adj	
Salting-out	Qiagen MagAttract HMW	0.0435	-0.0795	0.166	0.672	ns
Sorbitol wash & Salting-out	Qiagen MagAttract HMW	0.140	0.021	0.259	0.017	*
Sorbitol wash & Salting-out	Salting-out	0.096	-0.0513	0.245	0.265	ns

Supplementary table IV-6. Analysis of Variance of Aligned Rank Transformed Data for Tissue and Extraction method effects on A260/230. Degrees of freedom (Df), residual degrees of freedom (Df.res), F-statistic (F value), and the corresponding p-value (Pr (>F)). Significance codes are indicated as followed: for $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*', not significant 'ns'

	Df	Df. res	F value	Pr (>F)	
Extraction	2	53	2.290	0.111	ns
Tissue	2	53	1.548	0.222	ns
Extraction:Tissue	4	53	1.586	0.192	ns



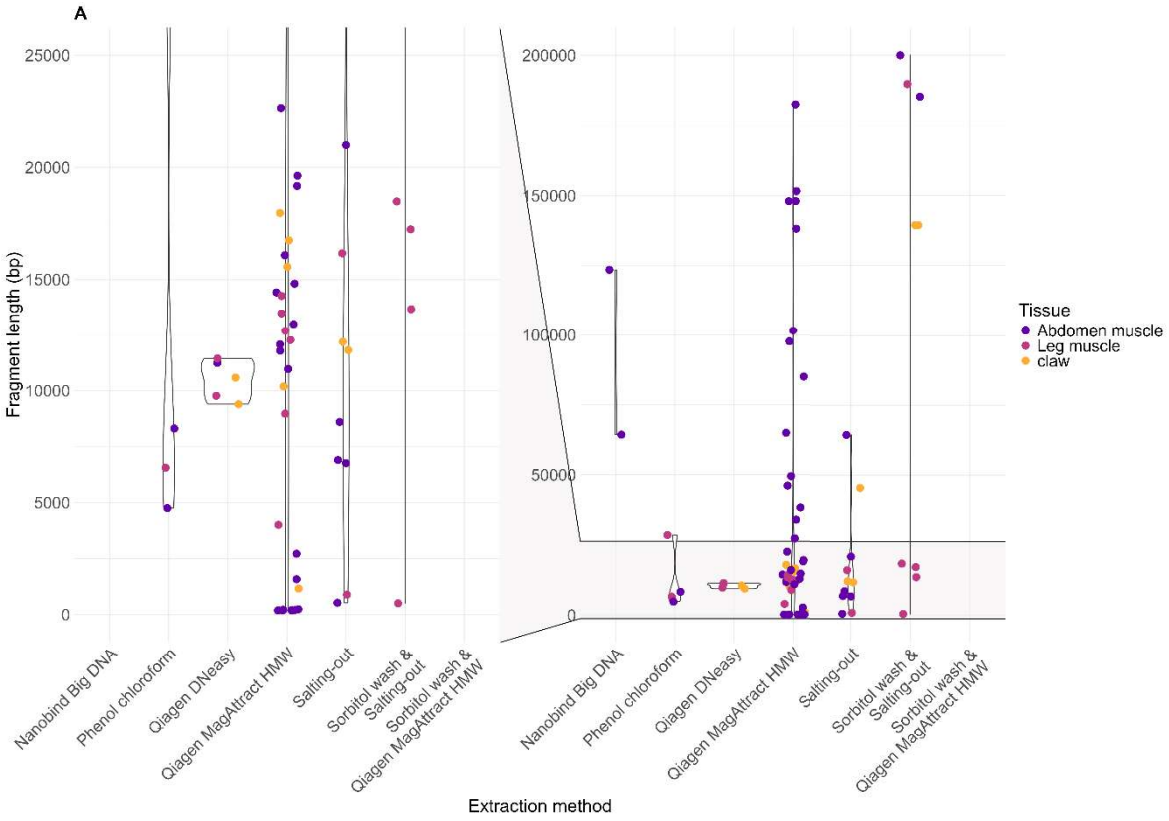
Supplementary figure IV-2. Absorbance ratio A260/280 (left) and A260/230 (right) for all extraction methods. Colour indicates tissue types used for DNA extraction. Dashed lines indicate optimal absorbance ratio values.

Supplementary table IV-7. Analysis of Variance of Aligned Rank Transformed Data for tissue and extraction method effects on DNA fragment length. Degrees of freedom (Df), residual degrees of freedom (Df.res), F-statistic (F value), and the corresponding p-value (Pr (>F)). Significance codes are indicated as followed: for $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*', not significant 'ns'.

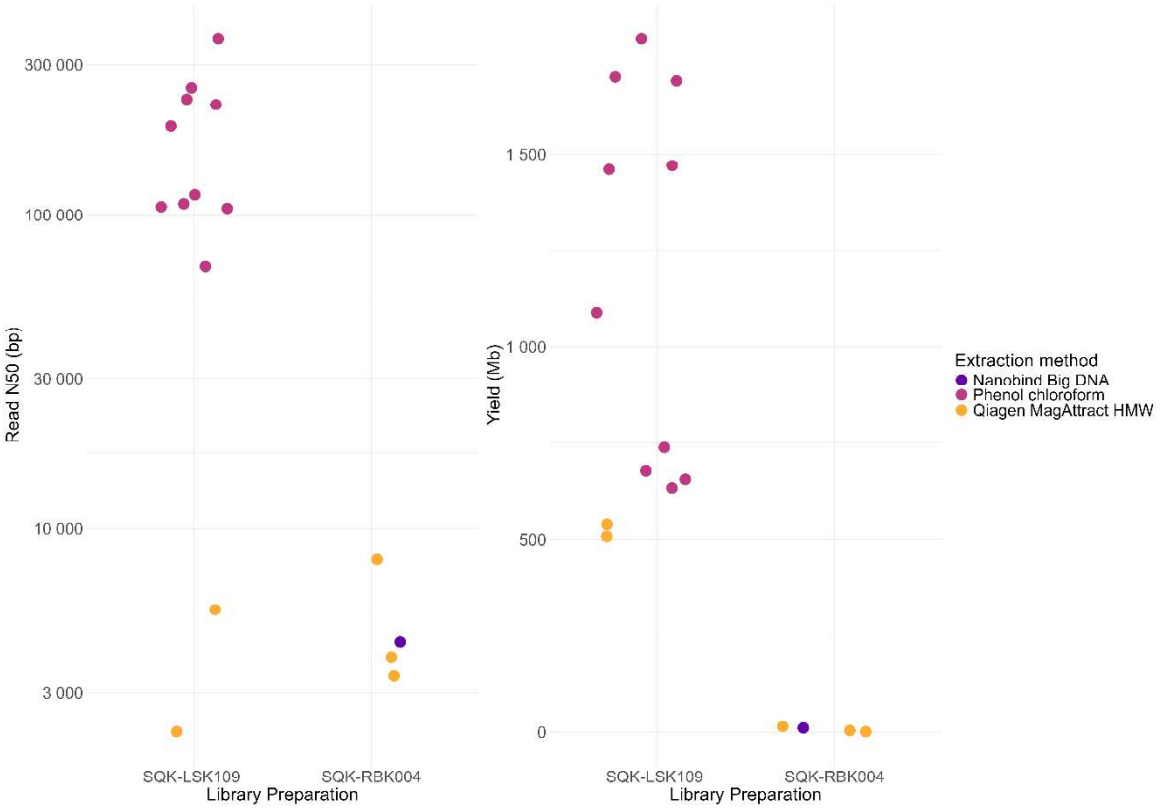
	Df	Df. res	F value	Pr (>F)	
Extraction	2	58	6.165	0.0038	**
Tissue	2	58	3.371	0.041	*
Extraction:Tissue	4	58	3.619	0.0106	*

Supplementary table IV-8. Post-hoc comparisons for main effects of Extraction method and Tissue type on DNA fragment length. For each comparison, the estimated difference (estimate), standard error (SE), degrees of freedom (df), t-ratio, and adjusted p-value (p.value) are given. P value adjustment using Tukey method. Significance codes are indicated as followed: for $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*', not significant 'ns'

		estimate	SE	df	t.ratio	p.value	
Qiagen MagAttract HMW	Salting-out	7.8	7.01	58	1.113	0.5103	ns
Qiagen MagAttract HMW	Sorbitol wash & Salting-out	-21.9	7.50	58	-2.918	0.0137	*
Salting-out	Sorbitol wash & Salting-out	-29.7	8.80	58	-3.370	0.0038	**
Abdomen muscle	Leg muscle	16.82	7.85	58	2.142	0.0903	ns
Abdomen muscle	claw	-3.51	8.26	58	-0.425	0.9054	ns
Leg muscle	claw	20.33	8.67	58	2.346	0.0573	ns



Supplementary figure IV-3. DNA fragment length (bp) at which the highest concentration of DNA extract was observed for all extraction methods. A) Zoomed in y-axis 0-25000 bp and B) y axis 0-200000 bp. Colour indicates tissue types used for DNA extraction.



Supplementary figure IV-4. Comparison of Nanopore sequencing metrics for different DNA extraction methods and library preparation kits. The left panel shows Read N50 (bp), and the right panel shows total sequencing yield (Mb). Colors represent different DNA extraction methods. Library preparation kits include SQK-LSK109 and SQK-RBK004.

Supplementary table IV-9. Wilcoxon sign rank test for HiFi read length and t-test results for HiFi yield between amplification free and amplification-based library preparation methods. For each comparison, the test statistic and p-value are given. Significance codes are indicated as followed: for $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*', not significant 'ns'

			statistic	p-value	significance
HiFi read length	Low input	Ultra-low input	53	0.0408	*
HiFi yield	Low input	Ultra-low input	-18.137	6.66e-21	***

Supplementary table IV-10. Kruskal-Wallis rank sum test for HiFi read length and HiFi yield across sample replicates. Kruskal-Wallis chi-squared statistic and the corresponding p-value are given. Significance codes are indicated as followed: for $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*', not significant 'ns'

sample		Statistic	P value	significance
AA1_70	Read length	0	1	ns
AA1_86	Read length	1	1	ns
AA1_91B	Read length	7	0.429	ns
AB2_1	Read length	3	0.3916	ns
AB2_28	Read length	4	0.406	ns
AB2_29	Read length	2	0.3679	ns
AA1_70	yield	0	1	ns
AA1_86	yield	0	1	ns
AA1_91B	yield	19	0.4568	ns
AB2_1	yield	3	0.3916	ns
AB2_28	yield	4	0.406	ns
AB2_29	yield	2	0.3679	ns

Supplementary table IV-11. t-tests for P0 and P1 Values across library preparation Methods. For each test, the t-statistic and the corresponding p-value are given. Significance codes are indicated as followed: for $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*', not significant 'ns'

			statistic	P value	significance
P0	Low input	Ultra-low input	7.223	1.14e-07	***
P1	Low input	Ultra-low input	-7.904	2.22e-08	***

Supplementary material - Chapter V

The following supplementary material can be downloaded at:

https://doi.org/10.1186/s12862-024-02268-5.

Supplementary table V-1. Sampling site locations, code and sampling size of *A. bihariensis* populations.

River basin	Population	Code	Sample size	GPS N	GPS E
Arieș	Bistra	BIS	20	46,4059	23,0541
Arieș	Starpă	STE	20	46,4728	22,9175
Barcău	Mare	MAR	20	47,1242	22,6216
Criș Alb	Corbului	COR	10	46,4244	22,3438
Criș Alb	Dudușoaia	DUD	20	46,4334	22,2324
Criș Alb	Rănușa	RAN	16	46,4391	22,2672
Criș Negru	Boga	BOG	20	46,6107	22,661
Criș Negru	Cuților	CUT	20	46,8311	22,3977
Criș Negru	Racu	RAC	20	46,6631	22,5255
Criș Negru	Tâlniciorii	TAL	20	46,4182	22,4672
Criș Repede	Anișelului	ANI	10	46,7883	22,8872
Criș Repede	Iadului	IAD	20	46,7447	22,5597
Criș Repede	Preluca	PRE	19	46,7257	22,8813

Supplementary table V-2 River basin, population, number of demultiplexed reads, assembled loci and variant sites after filtering for each individual

Curriculum vitae

ORCID

0000-0003-3987-5907

Languages

Croatian (native proficiency)
Italian (native proficiency)
English (full professional proficiency)
German (elementary proficiency)

Skills

Long-read sequencing
Animal cytogenetics
Population genomics
R, bash programming languages (basic)

2022. – present

PhD student • cotutelle
ICube Laboratory - The Engineering science, computer science and imaging laboratory, University of Strasbourg (France)

RPTU Kaiserslautern – Landau (Germany)
Thesis: Genome characterisation of European freshwater crayfish

Education

2018. – 2021.
University graduate programme in Molecular Biology • Faculty of Science, University of Zagreb (Croatia)

2015. – 2018.
University undergraduate programme in Molecular Biology • Faculty of Science, University of Zagreb (Croatia)

Research Experience

2021.
Research assistant and Erasmus+ internship • Senckenberg Biodiversity and Climate Research Centre, LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Frankfurt am Main (Germany).

2017. – 2021.
Intern • Laboratory for molecular analyses, Division of Zoology, Department of Biology, Faculty of Science, University of Zagreb (Croatia).

Awards

2024. – TBG Young Scientist Grant for scientific course participation

2023. - Award for poster presentation, CrayfIT Regional European IAA Meeting, Pavia (Italy)

2022. - Award for outstanding scientific work Faculty of Science, University of Zagreb (Croatia)

2019. - Rector's award for individual scientific work (one/two authors) in natural sciences „Genetic diversity of the stone crayfish", University of Zagreb (Croatia)

2019. – 2021. Leader of crayfish section, Biology student association – Udruga studenata biologije BIUS

Publications

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

Boštjančić, L.L., Bonassin, L., Anušić, L., Lovrenčić, L., Besendorfer, V., Maguire, I., Grandjean, F., Austin, C.M., Greve, C., Hamadou, A.B., Mlinarec, J. 2021. The *Pontastacus leptodactylus* (Astacidae) Repeatome Provides Insight Into Genome Evolution and Reveals Remarkable Diversity of Satellite DNA. Front. Genet. 11:611745. https://doi.org/10.3389/fgene.2020.611745

Gross, R., Lovrenčić, L., Jelić, M., Grandjean, F., Đuretanović, S., Simić, V., Burimski, O., Bonassin, L., Groza, M., Maguire, I. 2021. Genetic diversity and structure of the noble crayfish populations in the Balkan Peninsula revealed by mitochondrial and microsatellite DNA markers. PeerJ 9:e11838. https://doi.org/10.7717/peerj.11838

Lovrenčić, L., Temunović, M., Bonassin, L., Grandjean, F., Austin, C.M., Maguire, I. 2022. Climate change threatens unique genetic diversity within the Balkan biodiversity hotspot – The case of the endangered stone crayfish. Global Ecology and Conservation 39:e02301. https://doi.org/10.1016/j.gecco.2022.e02301

Dobrović A, Geček S, Klanjšček T, Haberle I, Dragičević P, Pavić D, Petelinec A, Boštjančić L.L., Bonassin L, Theissinger K, Hudina S., 2022, Recurring infection by crayfish plague pathogen only marginally affects survival and growth of marbled crayfish. NeoBiota. 77, 155-177, https://doi.org/10.3897/neobiota.77.87474

Bonassin L, Tarandek A. Monitoring populacija potočnog raka na području Žumberka i Samoborskog gorja (eng. Monitoring of stone crayfish populations in the Žumberak and Samoborsko gorje areas). In: Klarin A, Vizec P, Dupanović A, Požarić F, editors. A Collection of scientific studies of "Žumberak 2020." and "Žumberak 2021." Zagreb: Biology student association – BIUS; 2023. p. 139–51.

Boštjančić, L.L., Francesconi, C., Bonassin, L., Hudina, S., Gračan, R., Maguire, I., Rutz, C., Beck, A., Dobrović, A., Lecompte, O., Theissinger, K., 2023. Temporal dynamics of the immune response in Astacus astacus (Linnaeus, 1758) challenged with Aphanomyces astaci Schikora, 1906. Fish Shellfish Immunol. 143, 109185. https://doi.org/10.1016/j.fsi.2023.109185

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

Bonassin, L., Pârvulescu, L., Boštjančić, L.L., Francesconi, C., Paetsh, J., Rutz, C., Lecompte, O., Theissinger, K. 2024. Genomic insights into the conservation status of *Austropotamobius bihariensis*: low genetic diversity in the endemic crayfish species of the Apuseni Mountains. BMC Ecology and Evolution 24, 78. https://doi.org/10.1186/s12862-024-02268-5

Francesconi, C., Boštjančić, L.L., Bonassin, L., Schardt, L., Rutz, C., Makkonen, J., Schwenk, K., Lecompte, O., Theissinger, K. 2024 High variation of virulence in *Aphanomyces astaci* strains lacks association with pathogenic traits and mtDNA

haplogroups. Journal of Invertebrate Pathology, 206 https://doi.org/10.1016/j.jip.2024.108174

Boštjančić, L.L., Dragičević, P., Bonassin, L., Francesconi, C., Tarandek, A., Rutz, C., Lecompte, O., Theissinger, K. 2024. Expression of C/EBP and Kr-h1 transcription factors under immune stimulation in the noble crayfish. Gene 929, 148813 https://doi.org/10.1016/j.gene.2024.148813

Tarandek, A., Boštjančić, L.L., Francesconi, C., Bonassin, L., Schardt, L., Jussila, J., Kokko, H., Schwenk, K., Hudina, S., Lecompte, O., Theissinger, K. 2025. Characterisation of the noble crayfish immune response to oomycete-derived immunostimulants. Fish and Shellfish Immunology 166, 110666 https://doi.org/10.1016/j.fsi.2025.110666

Paetsch, J., Romahn, J., Theissinger, K., Baranski, D., Bonassin, L., Schardt, L Koschorreck, J., Krehenwinkel, H., Bálint, M. 2025. Two decades of compositional restructuring of soil biodiversity in Germany despite stable α- and β-diversity indices. https://doi.org/10.1101/2025.07.21.665925.

Oral communications

Bonassin, L., Boštjančić, L.L., Anušić, L., Lovrenčić, L., Besendorfer, V., Maguire, I., Grandjean, F., Austin, C.M., Greve, C., Ben Hamadou, A., Lecompte, O., Theissinger, K., Mlinarec, J. Cytogenomic investigation of repetitive elements in the family Astacidae. 2022, 23rd Symposium of the International Association of Astacology, Czech Republic.

Bonassin, L., Rutz, C., Schardt, L., Greve, C., Feldmeyer, B., Kress, A., Pârvulescu, L., Lecompte, O., Theissinger, K. Evaluation of high molecular weight DNA extraction strategies for long read sequencing in two non-model Astacidae (Crustacea: Decapoda) species. 2022, DeRGA International Symposium on "Reference Genomes for Biodiversity", Germany.

Bonassin, L., Paetsch, J., Rutz, C., Schardt, L., Greve, C., Lecompte, O., Pârvulescu, L., Theissinger, K. Genomic approach for the conservation of an endemic crayfish species. 2023, 4th Symposium on Freshwater Biology, Croatia.

Bonassin, L., Paetsch, J., Lecompte, O., Pârvulescu, L., Theissinger, K. Conservation genomics of endemic crayfish species populations using reduced representation sequencing. 2023, Regional European International Association of Astacology CrayfIT, Italy.

Poster presentations

Bonassin, L., Rutz, C., Schardt, L., Greve, C., Feldmeyer, B., Kress, A., Pârvulescu, L., Lecompte, O., Theissinger, K. Evaluation of high molecular weight DNA extraction strategies for long read sequencing in non-model Astacidae species. 2022, Long-Read Sequencing Uppsala 2022, Italy

Bonassin, L., Rutz, C., Schardt, L., Boštjančić, L.L., Francesconi, C., Greve, C., Kress, A., Pârvulescu, L., Lecompte, O., Theissinger, K. Reference genomes for non-model invertebrates: Giant genome sequencing of the endangered Austropotamobius bihariensis. 2023, Regional European International Association of Astacology CrayfIT, Italy.